Université du Québec
INRS-Eau

ESTIMATION NON PARAMÉTRIQUE DES QUANTILES DE CRUE PAR LA MÉTHODE DES NOYAUX

Par Dany Faucher
Baccalauréat en statistique

Mémoire présenté
pour l'obtention
du grade de Maître ès science (M.Sc.)

Jury d'évaluation

Examinateur externe	Abdous Belkacem Département de mathématiques et d'informatique Université du Québec à Trois-Rivières
Examinateur interne	Taha Ouarda INRS-Eau
Directeur de recherche	Peter F. Rasmussen INRS-Eau
Codirecteur de recherche	Bernard Bobée INRS-Eau

Avril 1999

REMERCIEMENTS

Pour la réalisation de ce mémoire, j'ai bénéficié de l'aide de plusieurs personnes que je désire remercier. Tout d'abord, mon directeur de recherche, Professeur Peter F. Rasmussen a été un atout indispensable par son soutien scientifique et financier. Sa disponibilité et son dévouement remarquables font de lui un directeur exemplaire. Ensuite, je tiens à remercier mon co-directeur, le Professeur Bernard Bobée, titulaire de la Chaire en hydrologie statistique, pour ses commentaires et son soutien financier. Les activités de la Chaire m'ont permis de rencontrer le Professeur Upmanu Lall de l'Utah State University que je tiens particulièrement à remercier pour ses importantes recommandations ainsi que Monsieur Jacques Bernier, pour ses commentaires sur la méthodologie. Je tiens aussi à remercier le Professeur Christian Léger de l'Université de Montréal pour ses conseils sur les développements effectués sur sa méthode. Finalement, je souhaite remercier ma famille pour leurs encouragements et leur compréhension.

« No amount of statistical refinement can overcome the disadvantage of not knowing the frequency distribution involved. »

Dooge (1986)

RÉSUMÉ

L'étude des probabilités d'occurrence des crues extrêmes s'effectue généralement à l'aide de méthodes paramétriques. Ces méthodes possèdent des inconvénients qui peuvent faire en sorte d'accroître l'incertitude liée à l'estimation des probabilités. Pour cette raison, les méthodes non paramétriques sont d'un intérêt grandissant. Ce travail a porté plus précisément sur la méthode des noyaux (*kernel method*), méthode non paramétrique qui gagne de plus en plus de popularité en hydrologie. À la lumière des résultats obtenus dans cette étude, la méthode des noyaux constitue une alternative intéressante aux méthodes paramétriques traditionnelles pour l'estimation de quantile de crue. Plusieurs méthodes pour l'estimation du paramètre de lissage ont été étudiée. Parmi celles-ci, deux méthodes sont apparues comme étant plus efficaces que les autres. Par ailleurs, il est généralement stipulé dans la littérature que le choix du type de noyau à considérer n'a que très peu d'importance sur la qualité de l'estimation. Toutefois, comme on doit généralement extrapoler sur la distribution pour estimer les quantiles de crue, on peut croire que les caractéristiques des extrémités de la fonction noyau peut avoir une certaine importance sur la qualité d'estimation. Diverses fonctions ont été étudiées en tant que noyau dans la méthode. Il apparaît clair que même dans un contexte d'extrapolation, l'influence du type de noyau est bien moindre que le choix du paramètre de lissage. Par contre, on a déterminé qu'une certaine gamme de noyaux procuraient une estimation relativement équivalente, tandis que l'utilisation de certains noyaux pouvait nuire considérablement à la qualité de l'estimation. Ce travail constitue une analyse approfondie des propriétés et de l'efficacité de la méthode des noyaux. Plusieurs autres aspects de cette méthode pourraient être explorés dans une étude ultérieure.

TABLE DES MATIERES

LISTE DES FIGURES

LISTE DES TABLEAUX

1. INTRODUCTION

1.1 Problématique

La nature est remplie de phénomènes qui se produisent de façon aléatoire. On qualifie ces processus incertains d'aléas naturels. Les précipitations de pluie ou de neige, le vent et le débit sont des exemples d'aléas. Ces variables aléatoires sont quantifiables et peuvent parfois atteindre un niveau tel qu'elles sont alors qualifiées d'événements extrêmes. Dans le cas des débits par exemple, le printemps est généralement le moment de l'année où le niveau d'eau dans les rivières et ruisseaux est à son plus haut, à cause de la fonte des neiges et des précipitations. On associe alors à cet événement le terme « crue » qui désigne un débit nettement supérieur à ce qui est observé pour le reste de l'année. Suite à des précipitations abondantes, il arrive aussi d'avoir une crue automnale ou une crue au milieu de l'été. Une crue peut-être qualifiée d'événement extrême dans la mesure où elle est relativement supérieure aux crues habituelles. Il est important de noter qu'un débit extrême peut aussi bien être un débit qui est particulièrement faible, on parle alors d'étiage. Des précipitations intenses peuvent être qualifiées d'orage tandis que des vents violents peuvent provoquer un ouragan ou une tornade.

On s'intéresse beaucoup aux événements extrêmes parce qu'ils ont des conséquences économiques importantes. Une rivière qui déborde peut causer des dommages physiques importants aux installations de la ville qu'elle inonde. De fortes précipitations peuvent causer un glissement de terrain qui peut détruire une route ou des habitations. Les ravages provoqués par une tornade, sont généralement considérables et ils engendrent des coûts importants pour la reconstruction. Mais les dommages ne sont malheureusement pas toujours uniquement d'ordre économique. Certains événements extrêmes peuvent aussi causer des pertes de vies humaines ou des dommages non quantifiables à l'environnement.

Lors de la construction de certaines installations, on dimensionne les ouvrages de façon à ce qu'ils soient sécuritaires. La sécurité des gens est un paramètre important à prendre en compte dans la construction d'un barrage par exemple, ce qui vient augmenter les coûts de l'installation.

En hydrologie, le débit est probablement une des variables aléatoires les plus étudiées. On cherche à prévoir les variations de débit pour diverses raisons. D'abord, l'étude du débit a son importance dans la gestion de la qualité de l'eau. Une diminution du débit augmente la concentration d'un polluant dans un cours d'eau, la charge du polluant étant constante ($concentration = charge/débit$). Ainsi, on s'intéresse aux débits d'étiages, moments où la concentration de polluants est importante. On étudie le débit aussi parce que l'on désire connaître le régime hydrologique pour la construction de ponceaux ou de canaux. Une route risque d'être inondée si le canal servant à écouler l'eau de drainage sous la route n'est pas de taille suffisante. On s'intéresse également à la variable débit pour des raisons de sécurité publique. On ne construit pas une habitation à un endroit qui risque d'être inondé au printemps. Parfois, un site qui peut nous sembler adéquat et hors de danger, comporte tout de même un certain risque d'être inondé. L'étude du débit permet alors de quantifier ce risque et ainsi de prendre une décision sécuritaire. Finalement, il apparaît évident que la construction d'un barrage sur une rivière serait impensable sans avoir au préalable effectué une étude du débit de la rivière. D'abord, il doit être suffisamment haut pour contenir des crues extrêmes. Il faut donc détenir assez d'information sur le débit de la rivière pour au moins connaître la probabilité d'occurrence des crues extrêmes. L'étude du débit intervient donc dans le dimensionnement d'un ouvrage hydroélectrique.

En hydrologie, on cherche souvent à associer une probabilité aux événements extrêmes. Cette probabilité a un rôle important à jouer dans le dimensionnement des diverses installations construites pour améliorer la qualité de vie des êtres humains. Ces installations doivent être construites en considérant le risque associé aux débits extrêmes. Par exemple, lors de la construction d'un pont qui enjambe une rivière, il est important de déterminer la

hauteur que celui-ci doit avoir en fonction de la probabilité des débits extrêmes. Il doit être construit en tenant compte du risque qu'un certain débit extrême se produise, sinon il risque d'être submergé et d'être endommagé. Comme une protection totale n'est rarement rentable, les installations sont construites avec un certain risque, en tenant compte de la probabilité de débordement. Il est donc très important de connaître la probabilité associée aux événements extrêmes.

La connaissance de la probabilité des événements extrêmes est importante non seulement dans le dimensionnement des ouvrages, mais aussi dans la gestion des cours d'eau. La gestion du niveau d'un réservoir d'une centrale hydroélectrique en est un exemple. On doit pouvoir subvenir à la demande hydroélectrique en tout temps. Avant la crue, on doit s'assurer que le niveau du réservoir n'est pas trop élevé, puisque dans l'éventualité d'un événement extrême, par exemple une pluie intense sur plusieurs jours ou une fonte de neige très rapide, on risque un débordement du réservoir, le bris des installations hydroélectriques et des dommages considérables sur les abords du cours d'eau dans lequel se déverse le réservoir. Après la crue, il faut s'assurer d'avoir conservé le niveau du réservoir suffisamment haut afin d'emmagasiner des réserves pour la période d'étiage. En connaissant la probabilité d'occurrence des événements extrêmes, il est possible d'améliorer la gestion d'un ouvrage pour la sécurité et la rentabilité.

1.2 Objectifs

L'étude de la probabilité des débits extrêmes, l'analyse de fréquence de crue, s'effectue généralement à l'aide de méthodes dites « paramétriques ». On considère alors une certaine loi statistique comme étant représentative de la distribution de probabilité des pointes de crue annuelles du cours d'eau d'intérêt. On verra plus loin que ces méthodes possèdent des inconvénients qui peuvent faire en sorte d'accroître l'incertitude liée à l'estimation des

probabilités. Beaucoup de travaux sont effectués dans le but d'accroître l'efficacité de ces méthodes. Depuis quelques années, les méthodes dites « non paramétriques » ont gagné de la popularité dans plusieurs domaines de la science où l'étude des probabilités est requise, l'hydrologie ne faisant pas exception. Ces méthodes ont l'avantage de ne pas nécessiter d'hypothèses sur la distribution de la population d'intérêt. L'intérêt marqué pour ces méthodes et les inconvénients causés par les méthodes traditionnelles nous a permis de formuler l'objectif principal du présent travail de maîtrise de la façon suivante:

> Proposer une méthode non paramétrique comme une alternative aux modèles paramétriques pour l'estimation de fréquence de crue.

Dans la présente étude, la méthode des noyaux est étudiée dans un contexte d'estimation de fréquence de crue. Depuis son introduction en hydrologie en 1983, la méthode des noyaux a connu des développements considérables et gagné de l'intérêt chez les hydrologues-statisticiens. Pour atteindre l'objectif principal, il a été nécessaire de définir les sous-objectifs suivants :

a) Évaluer la performance de la méthode des noyaux par rapport à l'ajustement des distributions paramétriques ;

b) Évaluer la performance de différentes méthodes de calcul du paramètre de lissage pour la méthode des noyaux ;

c) Proposer un type de noyau à utiliser selon le contexte ;

d) Étudier l'efficacité de la méthode des noyaux selon la taille de l'échantillon disponible.

Ce travail se divise en six chapitres. Le second chapitre traite des notions de base en estimation de fréquence de crue ainsi que des méthodes d'estimation classiques, c'est-à-dire l'ajustement de distributions statistiques. Au chapitre suivant, la méthode des noyaux est introduite et on étudie certaines de ses propriétés. Au chapitre quatre, une revue des principales méthodes de calcul du paramètre de lissage est présentée. On y étudie les propriétés de chacune des méthodes. Le chapitre cinq renferme les résultats des différentes comparaisons effectuées dans le cadre des objectifs définis précédemment. Le dernier chapitre est consacré à la discussion des résultats, aux conclusions et aux recommandations.

2. ESTIMATION DE FRÉQUENCE DE CRUE

Dans ce chapitre, des outils statistiques de base permettant de prévoir la réalisation de variables aléatoires, seront présentés.

2.1 Densité de probabilité et fonction de répartition

La fonction de densité de probabilité est une des caractéristiques essentielles décrivant le comportement d'une variable aléatoire. Si on connaît la loi de probabilité de la variable aléatoire associée à un processus hydrologique particulier, on est en mesure de connaître le comportement de cette variable et ainsi, éventuellement d'en prévoir les variations. La fonction de densité de probabilité nous permet de connaître les chances qu'une réalisation de la variable aléatoire X, se trouvant dans un certain intervalle $[a,b]$, se produise. En intégrant la fonction de densité de a à b, on obtient ainsi la probabilité d'occurrence d'un événement se trouvant dans l'intervalle $[a,b]$:

$$prob\{a \le X \le b\} = \int_a^b f(x)\,dx \qquad \forall\, a \le b \tag{2.1}$$

La fonction de densité f est non-négative en tout point. De plus, l'aire totale sous la courbe f, c'est-à-dire l'intégrale de la fonction de densité sur $\mathbb{R}$ est égale à 1:

$$f(x) \ge 0 \quad \forall x\ ; \qquad \int_{-\infty}^{\infty} f(x)\,dx = 1 \tag{2.2}$$

La fonction de répartition, ou la fonction de distribution cumulée, est définie comme le cumul de la fonction de densité :

$$F(x) = P\{X \leq x\} = \int_{-\infty}^{x} f(t)\, dt \qquad (2.3)$$

Cette fonction possède les caractéristiques suivantes :

(a) Non décroissante sur tout le domaine ;

(b) Non-négative et n'excédant pas 1 ; $0 \leq F(x) \leq 1 \quad \forall x$;

(c) $F(-\infty) = 0$ et $F(\infty) = 1$.

En hydrologie statistique, on utilise la fonction de répartition pour calculer des quantiles de période de retour T.

2.2 Quantile de période de retour

L'analyse de fréquence de crue permet d'établir une relation entre des événements hydrologiques extrêmes et leur probabilités au dépassement ou au non-dépassement. Il est important de connaître le risque qu'un événement d'une certaine amplitude se produise. Par exemple, lors de la conception d'un ouvrage hydroélectrique, on doit connaître la distribution de probabilité d'un événement extrême afin de dimensionner le barrage de façon sécuritaire et rentable. Si le barrage est sous-dimensionné, il risque d'être inondé plus souvent que prévu. Mais par contre, il n'est pas bénéfique de construire un barrage sur-dimensionné, résistant à une crue qui n'a aucune chance de se réaliser, qui engendre des coûts. Il est donc important de déterminer un débit optimal servant au dimensionnement du barrage, qui fera en sorte que ce dernier soit le plus résistant que possible à un coût moindre.

Pour la construction d'un barrage, on désire habituellement que celui-ci puisse résister à un certain débit maximal. On définit donc la variable aléatoire X comme étant le débit

maximum annuel et la valeur x_{max} comme étant le débit maximum critique pour le barrage. On s'intéresse alors à la probabilité que le débit critique soit dépassé c'est-à-dire :

$$prob\{X > x_{max}\} = 1 - F(x_{max}) \tag{2.4}$$

Afin de définir la période de retour et le quantile, on rappelle d'abord la définition de la loi de probabilité discrète de Bernoulli. Supposons que l'on réalise une expérience dont le résultat sera interprété soit comme un succès soit comme un échec. On définit alors la variable aléatoire Y en lui donnant la valeur 1 lors d'un succès et 0 lors d'un échec. La loi de probabilité de Y est alors la suivante, où p est la probabilité d'un succès:

$$prob\{Y = 1\} = p \quad \text{et} \quad prob\{Y = 0\} = 1 - p \tag{2.5}$$

Dans le cas du débit maximum annuel, on définit le cas où $X > x_{max}$ comme étant un « échec » et le cas où $X \leq x_{max}$ comme étant un « succès ». On peut donc définir une expérience de Bernoulli de la façon suivante :

$$prob\{X > x_{max}\} = p \; ; \quad prob\{X \leq x_{max}\} = 1 - p \tag{2.6}$$

On exécute une série d'épreuves indépendantes ayant chacune la probabilité p d'être un échec et ce, jusqu'à l'obtention d'un premier échec. Si on désigne le nombre d'épreuves nécessaires avant d'obtenir le premier échec par M, M est alors une variable aléatoire qui suit la loi géométrique :

$$prob\{M = m\} = p\,(1-p)^{m-1} \tag{2.7}$$

Si l'on considère la variable débit maximum annuel, M désigne le nombre d'années avant d'avoir un débit supérieur au débit maximal x_{max}. On peut montrer que:

$$E[M] = \frac{1}{p} = T \tag{2.8}$$

Le nombre moyen d'épreuves pour obtenir la première réalisation de $X > x_{max}$ est égale à l'inverse de la probabilité d'avoir $X > x_{max}$. La valeur T, que l'on nomme période de retour en hydrologie, représente donc le nombre moyen d'années qui s'écoulent entre des débits supérieurs au débit maximal x_{max}. Généralement, la variable T est interprétée en hydrologie comme étant le temps moyen entre deux événements où la valeur maximale est dépassée. Cette variable est une moyenne qui est souvent calculée sur une longue période (100 ans et plus). Si on définit la valeur maximale en fonction de la période de retour T, on obtient une variable x_T représentant le quantile de période de retour T et ainsi on a :

$$prob\{X > x_T\} = 1 - F(x_T) = \frac{1}{T} \tag{2.9}$$

Ainsi, on peut dire que le débit x_T possède une période de retour T si ce débit est dépassé en moyenne à chaque T années. Finalement, on peut écrire le quantile x_T de période de retour T de la façon suivante :

$$x_T = F^{-1}\left(1 - \frac{1}{T}\right) \tag{2.10}$$

où F^{-1} signifie la fonction inverse de la fonction de répartition. La fonction inverse attribue une valeur x à une probabilité plutôt que d'attribuer une probabilité à une valeur x :

$$prob = F(x) \; ; \qquad x = F^{-1}(prob) \tag{2.11}$$

Lors de la conception d'un barrage, on doit donc déterminer la période de retour qui nous convient. On peut alors dire que l'on est prêt à accepter la défaillance du barrage une fois tous les 100 ans par exemple. Le débit de période de retour 100 a donc une probabilité 0,01 d'être dépassé. Cette valeur ne nous permet pas de dire avec certitude qu'un débit supérieur à x_T se produira dans cent ans, mais plutôt que sur une longue période de temps, le débit x_T se réalisera en moyenne une fois sur cent.

L'analyse de fréquence de crue permet donc de trouver le débit x_T qui correspond à la période de retour désirée. Comme on l'a vu avec l'équation (2.10), pour pouvoir déduire le quantile à partir de la période de retour T, il faut connaître la fonction de répartition F et par le fait même, la fonction de densité f. Mais, comme dans la plupart des cas, la loi de probabilité de la variable aléatoire d'intérêt est inconnue, il faut tenter de la trouver ou bien l'estimer.

2.3 Ajustement de distributions statistiques

Le débit d'un cours d'eau dépend des précipitations qui se produisent aléatoirement. De plus, la nature étant relativement complexe, les caractéristiques statistiques du débit d'un cours d'eau sont associées à un grand nombre de processus physiques qui convertissent les précipitations en débits. Il est difficile de connaître chacun des processus qui interviennent et il est encore plus difficile d'en connaître le degré d'implication ou l'influence. On peut dans certains cas, avoir une idée des processus impliqués et connaître l'essentiel des relations qui existent entre ces processus et le débit. Cependant, en général les connaissances sont trop vagues pour permettre de déduire les caractéristiques statistiques du débit à partir des processus physiques et pour cette raison on fait appel à l'approche statistique.

Les caractéristiques stochastiques des précipitations et les processus physiques qui les transforment en débits sont complexes et résultent en une distribution complexe des débits. La modélisation d'un processus hydrologique est complexe et il y a toutes sortes d'erreurs d'introduites dans le modèle. Il apparaît donc difficile, voire même impossible, de connaître la forme exacte de la de probabilité empirique, c'est-à-dire l'ensemble des paramètres qui caractérisent la distribution des probabilités d'occurrence de la variable aléatoire. Par contre, il est possible d'obtenir une estimation des paramètres de la loi à partir d'un

échantillon de la population. On dit alors que l'on ajuste une loi de probabilité à cet échantillon.

2.3.1 Méthodes d'ajustement

Lorsque l'on effectue un ajustement d'une distribution statistique à un échantillon de données, les inférences que l'on peut faire sur la population d'où provient cet échantillon, peuvent être faussées si la loi sélectionnée n'est pas adéquate. Mais, il se peut aussi que l'incertitude provienne de l'estimation des paramètres de la loi. Il existe plusieurs méthodes pour estimer les paramètres d'une loi. Il est difficile de recommander l'utilisation d'une méthode particulière pour toutes les lois. Par contre, pour chaque loi, on peut généralement dégager une méthode qui est plus efficace que les autres. Dans cette section, trois méthodes d'estimation qui ont été utilisées dans la partie paramétrique du présent travail seront présentées.

2.3.1.1 Maximum de vraisemblance

La méthode du maximum de vraisemblance consiste à maximiser une certaine fonction que l'on nomme fonction de vraisemblance. La fonction de vraisemblance est la densité jointe de l'échantillon. Considérons un échantillon aléatoire $X_1, X_2, \ldots, X_n$, où les variables aléatoires X_i sont indépendantes et identiquement distribuées. Supposons que cet échantillon provient d'une distribution $f(x;\theta_1,\theta_2,\ldots,\theta_m)$ où $\theta_1,\theta_2,\ldots,\theta_m$ sont les m paramètres régissant la forme de la distribution. En vertu de l'indépendance des X_i, la densité jointe est la multiplication des densités de toutes les variables aléatoires X_i. On utilise le nom fonction de vraisemblance pour désigner la densité jointe :

$$L(\theta_1,\theta_2,\ldots,\theta_m;x_1,x_2,\ldots,x_n)=\prod_{i=1}^{n} f(x_i;\theta_1,\theta_2,\ldots,\theta_m) \qquad (2.12)$$

On peut obtenir l'estimateur $\hat{\theta}_p$ en dérivant la fonction de vraisemblance par rapport à θ_p et en égalant les dérivées à zéro. On a donc à résoudre le système d'équations suivant :

$$\frac{\partial L(\theta_1,\theta_2,...,\theta_m;x_1,x_2,...,x_n)}{\partial \theta_p} = 0 \qquad p = 1,2,...,m \tag{2.13}$$

Il est souvent plus pratique d'utiliser le logarithme de la fonction de vraisemblance, ce qui facilite généralement le calcul de la dérivée. Que l'on dérive la fonction log-vraisemblance ou la fonction de vraisemblance elle-même, le maximum obtenu sera le même dans les deux cas puisque :

$$\frac{d \ln L(\theta)}{d\theta} = \frac{1}{L}\frac{d\, L(\theta)}{d\theta} \tag{2.14}$$

2.3.1.2 Méthode des moments

Les moments caractérisent la forme d'une distribution. L'estimation des paramètres d'une distribution par la méthode des moments, consiste à égaler les moments empiriques (calculés à partir des données) aux moments théoriques de la distribution considérée. On note le moment théorique non-centré d'ordre *r* d'une distribution par μ'_r et on le calcule de la façon suivante:

$$\mu'_r = E[X^r] = \int_{-\infty}^{\infty} x^r f(x)\, dx \tag{2.15}$$

Le moment non-centré d'ordre 1 , μ'_1 , représente la moyenne de la distribution. Le moment d'ordre *r* centré par rapport à la moyenne est défini par :

$$\mu_r = E\left[\left(X - \mu'_1\right)^r\right] = \int_{-\infty}^{\infty} \left(x - \mu'_1\right)^r f(x)\, dx \tag{2.16}$$

Le moment centré d'ordre 2 représente la variance de la distribution. Les moments d'ordre 3 et 4 sont reliés respectivement aux coefficients d'asymétrie et d'aplatissement.

Si l'on considère un échantillon $\{x_1,...,x_n\}$ de taille n, le moment d'ordre r non-centré, dénoté par m'_r,est défini par :

$$m'_r = \frac{1}{n}\sum_{i=1}^{n} x_i^r \qquad (2.17)$$

De la même façon que dans le cas des moments théoriques, le moment non-centré d'ordre 1 est la moyenne de l'échantillon. Le moment d'ordre r de l'échantillon centré sur la moyenne est exprimé de la façon suivante :

$$m_r = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - m'_1\right)^r \qquad (2.18)$$

où m_2 représente la variance de l'échantillon, un estimateur biaisé de la variance de la population.

Les paramètres $\theta_1,\theta_2,...,\theta_m$ de la loi $f\left(x;\theta_1,\theta_2,...,\theta_m\right)$ sont obtenus en égalant les moments théoriques et les moments de l'échantillon. On doit donc résoudre le système à m équations suivant (il doit y avoir autant d'équations que de paramètres inconnus):

$$\mu'_r = m'_r \qquad r = 1,2,...,m \qquad (2.19)$$

2.3.1.3 Méthode des moments pondérés

Cette méthode est particulièrement intéressante pour les distributions pouvant être définies de façon explicite par leur fonction de distribution inverse $X(F)$. Les moments pondérés théoriques d'ordre r peuvent être définis comme suit (Greenwood *et al.* ; 1979):

$$\beta_r = E\left[XF^r\right] = \int_0^1 x(F)F^r dF \tag{2.20}$$

Les moments pondérés de l'échantillon d'ordre r sont donnés par:

$$b_r = \frac{1}{n}\sum_{i=1}^{n} \frac{(i-1)(1-2)...(1-r)}{(n-1)(n-2)...(n-r)} x_{(i)} \tag{2.21}$$

où $x_{(i)}$ est la i^e statistique d'ordre, c'est-à-dire le i^e élément de l'échantillon classé en ordre croissant, $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$.

Tout comme la méthode des moments classiques présentée précédemment, les paramètres $\theta_1, \theta_2, ..., \theta_m$ de la loi $f\left(x; \theta_1, \theta_2, ..., \theta_m\right)$ sont obtenus en égalant les moments pondérés théoriques et les moments pondérés de l'échantillon correspondants. On doit donc résoudre le système à m équations suivant :

$$\beta_r = b_r \qquad r = 1,2,...,m \tag{2.22}$$

2.3.2 Lois de probabilité

L'analyse de données hydrologiques extrêmes montre que généralement la distribution empirique possède une asymétrie positive, ce qui exclut la loi normale qui à un coefficient d'asymétrie nul. Les distributions possèdent souvent une borne inférieure, une borne supérieure, ou les deux. Dans le cas du débit d'une rivière, la distribution possède une borne inférieure, puisque le débit ne peut être négatif, mais ne possède aucune borne supérieure (Bobée et Ashkar ; 1991). Il importe donc de choisir une loi qui convient à notre échantillon.

Pour ajuster une distribution statistique, on doit d'abord faire l'hypothèse que les observations de l'échantillon sont indépendantes et identiquement distribuées. On doit donc

vérifier en plus de l'indépendance des observations, l'homogénéité, la stationnarité et la présence de valeurs aberrantes dans l'échantillon. Il existe des tests permettant de vérifier chacune de ces hypothèses (HYFRAN; 1998, Ondo *et al.* ; 1997 , Faucher *et al.* ; 1997), mais ils ne seront pas présentés dans cette étude. En effet, ce travail consiste plutôt à étudier les méthodes non paramétriques qui sont plus robustes au non respect de ces hypothèses. La brève présentation des méthodes paramétriques a pour unique but de donner marche à suivre pour les utiliser.

Parmi le grand nombre de distributions statistiques qui peuvent être utilisées, seulement trois parmi les principales seront présentées dans cette section, soit la log-Pearson type 3, la loi généralisée des valeurs extrêmes et la log-normale à 3 paramètres, ainsi que la méthode d'estimation des paramètres recommandée pour chacune d'elles.

2.3.2.1 Loi log-Pearson type 3 (LP3)

La loi log-Pearson type 3 (LP3) fait partie de la famille des distributions gamma. Cette loi a été recommandée aux États-Unis par le *Water Resources Council* (*WRC*) (Benson ; 1968) et en Australie par l'*Institute of Engineers* (*I.E.A*). La distribution log-Pearson type provient de la distribution Pearson type 3. Si la variable aléatoire $Y = \ln(X)$ est distribuée selon une loi Pearson type 3, alors la variable aléatoire X suit une distribution log-Pearson type 3. Les paramètres des deux lois sont les mêmes. La fonction de densité de la LP3 s'écrit donc de la façon suivante :

$$f(x) = \frac{\alpha^{\lambda}}{x\Gamma(\lambda)}(\ln x - m)^{\lambda-1} e^{-\alpha(\ln x - m)} \qquad (2.23)$$

où $\alpha \neq 0$, $\lambda > 0$ et m sont des paramètres de la distribution. Cette loi n'a pas de fonction inverse explicite. La méthode des moments est utilisée pour trouver les paramètres de cette distribution. Le moment non-centré d'ordre r (2.19) est donné par (Bobée ; 1975):

$$\mu'_r = e^{\frac{mr}{k}} \Big/ \left(1 - \frac{r}{\alpha k}\right)^{\lambda} \qquad r = 1,2,\ldots,m \tag{2.24}$$

En résolvant le système d'équation formé des trois premiers moments, en prenant le logarithme de chacun et en posant $\beta = \alpha k$, on trouve :

$$\frac{\ln\left[\left(1-1/\hat{\beta}\right)^3 / \left(1-3/\hat{\beta}\right)\right]}{\ln\left[\left(1-1/\hat{\beta}\right)^2 / \left(1-2/\hat{\beta}\right)\right]} = \frac{\ln m'_3 - 3\ln m'_1}{\ln m'_2 - 2\ln m'_1} \tag{2.25}$$

$$\hat{\lambda} = \frac{\ln m'_2 - 2\ln m'_1}{\ln\left[\left(1-1/\hat{\beta}\right)^2 / \left(1-2/\hat{\beta}\right)\right]} \tag{2.26}$$

$$m = k\left(\ln m'_1 + \hat{\lambda}\left(1-1/\hat{\beta}\right)\right) \tag{2.27}$$

On doit d'abord calculer la valeur de $\hat{\beta}$ avec une méthode itérative, puisque l'équation (2.25) est implicite en $\hat{\beta}$. Ensuite, on peut déduire facilement les valeurs de $\hat{\lambda}$ et de $\hat{m}$.

2.3.2.2 Loi généralisée des valeurs extrêmes

La loi généralisée des valeurs extrêmes (GEV) a été introduite pour convenir aux situations où l'on mesure des événements extrêmes, comme c'est le cas en hydrologie lorsque l'on s'intéresse au débit maximum annuel d'une rivière par exemple (Jenkinson ; 1955). La GEV regroupe une famille de lois des valeurs extrêmes, formée de la EV1, de la EV2 et de la EV3. Depuis 1975, la loi généralisée de cette famille (GEV), qui combine les trois autres, est couramment utilisée au Royaume-Uni, suite aux recommandations faites par le *Natural Environmental Research Council* (*NERC*) (Hall ; 1984). La fonction de densité de la GEV s'écrit de la façon suivante :

$$f(x) = \frac{1}{\alpha}\left[1 - \frac{k}{\alpha}(x-u)\right]^{1/k-1} \exp\left(-\left[1 - \frac{k}{\alpha}(x-u)\right]^{1/k}\right) \qquad (2.28)$$

où $\alpha > 0$, u et k sont respectivement les paramètres d'échelle, de position et de forme. La fonction de distribution de cette loi est définie de la façon suivante :

$$F(x) = \exp\left(-\left[1 - \frac{k}{\alpha}(x-u)\right]^{1/k}\right) \qquad (2.29)$$

Un avantage de cette loi est que l'on peut définir de façon explicite, la fonction de distribution inverse de la GEV :

$$x(F) = u + \frac{\alpha}{k}\left[1 - (-\ln F)^k\right] \qquad (2.30)$$

Il est donc possible d'utiliser la méthode des moments pondérés pour l'estimation des paramètres de la distribution . Le moment pondéré d'ordre r de la distribution GEV est donné par :

$$\beta_r = \frac{1}{1+r}\left(u + \frac{\alpha}{k}\left[1 - \frac{\Gamma(1+k)}{(1+r)^k}\right]\right) \qquad (2.31)$$

En considérant les moments pondérés d'ordre 0, 1 et 2 et en résolvant le système d'équations correspondant, on peut estimer les paramètres u, α et k :

$$\hat{k} = 7.8590c + 2.9554c^2 \quad \text{où} \quad c = \frac{2b_1 - b_0}{3b_2 - b_0} - \frac{\log 2}{\log 3} \qquad (2.32)$$

$$\hat{\alpha} = \hat{k}\left(\frac{2b_1 - b_0}{\Gamma\left(1+\hat{k}\right)\left(1 - 2^{-\hat{k}}\right)}\right) \qquad (2.33)$$

$$\hat{u} = b_0 + \frac{\hat{\alpha}}{\hat{k}}\left(\Gamma\left(1+\hat{k}\right)-1\right) \tag{2.34}$$

2.3.2.3 Loi log-normale à 2 paramètres

La distribution log-normale à 2 paramètres (LN2) est la dernière des trois distributions statistiques utilisées dans cette étude. Elle est déduite de la loi normale qui est la distribution la plus utilisée en statistique et elle est reconnue en hydrologie statistique. Si la variable aléatoire $Y = \ln(X)$ est distribuée selon une normale de moyenne μ_Y et de variance σ_Y^2, alors X suit une distribution LN2. La densité d'une loi LN2 s'exprime de la façon suivante :

$$f(x) = \frac{1}{x\,\sigma_Y\sqrt{2\pi}} \exp\left(-\frac{\left[\ln(x)-\mu_Y\right]^2}{2{\sigma_Y}^2}\right) \tag{2.35}$$

où μ_Y et σ_Y sont les paramètres de la distribution. Pour l'estimation des paramètres de cette distribution, la méthode du maximum de vraisemblance a été utilisée. Le logarithme de la fonction de vraisemblance de la distribution LN2 est donné par :

$$\ln L\left(\mu_Y,\sigma_Y\right) = -\frac{n}{2}\ln\left(2\pi\sigma_Y^2\right) - \sum_{i=1}^{n}\ln\left(x_i\right) - \frac{1}{2\sigma_Y^2}\sum_{i=1}^{n}\left[\ln\left(x_i\right)-\mu_Y\right]^2 \tag{2.36}$$

Si l'on annule les dérivées partielles correspondant aux deux paramètres, on obtient une expression pour l'estimation de chacun des paramètres de la distribution en résolvant le système de deux équations:

$$\hat{\mu}_Y = \frac{1}{n}\sum_{i=1}^{n}\ln\left(x_i\right) \tag{2.37}$$

$$\hat{\sigma}_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left[\ln\left(x_i\right)-\hat{\mu}_Y\right]^2 \tag{2.38}$$

3. MÉTHODES NON PARAMÉTRIQUES

Traditionnellement, l'analyse de fréquence de crue s'effectue à partir de méthodes paramétriques, par exemple par l'ajustement de des distributions présentées au chapitre précédent. On fait alors l'hypothèse que les observations de l'échantillon proviennent toutes d'une même population et que leur distribution est connue. Cependant, on ne connaît généralement pas la distribution théorique à partir de laquelle les observations sont tirées. Ainsi si l'on effectue l'estimation d'un quantile à partir d'un échantillon tiré d'une population dont on ignore la distribution, il a y une incertitude importante associée au choix de la loi. De plus, il est difficile de quantifier cette incertitude si on veut en tenir compte dans notre estimation.

Jusqu'à maintenant, on n'arrive pas tout à fait à s'entendre sur « la distribution » à utiliser dans un contexte hydrologique. Mais comme il est important de fixer un certain standard dans le but d'éviter des choix de distribution arbitraires, certaines distributions sont favorisées plus que d'autres. Comme on a vu précédemment, la loi généralisée des valeurs extrêmes (GEV) est la distribution qui est recommandée au Royaume-Uni pour estimer les débits extrêmes de crue lors de l'analyse de fréquence des crues. Aux États-Unis, il est plutôt convenu de considérer la Log Pearson type 3 (LP3) comme méthode de base pour toutes les agences fédérales américaines. Ces recommandations ont l'avantage d'uniformiser les études dans les agences de recherche à l'intérieur d'un pays et ainsi d'éviter les grandes divergences de résultat. Dans chacun des cas, la loi recommandée est probablement celle qui a donné les meilleurs ajustements en moyenne sur un certain nombre d'études. Mais, ces distributions ne sont pas nécessairement celles qui sont les plus efficaces dans toutes les situations. Prenons par exemple les États-Unis où on utilise la LP3.

On peut avoir les données d'une rivière en particulier qui s'ajuste mal à la LP3, peut-être est-il préférable d'utiliser alors une autre loi, la GEV ou la LN3 par exemple. La LP3 n'est donc pas nécessairement utilisable dans tous les cas possibles. On est contraint par la forme de la distribution que l'on utilise. Il est donc difficile d'accepter le standard s'il n'est pas toujours efficace.

Parfois, on peut posséder certaines informations sur la distribution de la population, comme le fait qu'elle soit bimodale ou unimodale ou bien on peut connaître certaines caractéristiques concernant les extrémités (lourdes ou légères). Cette information peut provenir d'expériences régionales, mais peut aussi provenir de l'échantillon lui-même. Dans ce cas, l'incertitude liée au choix de la distribution peut être réduite. Mais par contre, le problème peut devenir insoluble, lorsque l'on n'a aucune information concernant la distribution, ou plus précisément lorsque que l'on ne connaît pas la forme des extrémités où l'on veut estimer les quantiles. Peu importe la situation dans laquelle on se trouve, l'incertitude liée à la connaissance de la distribution d'une population, induit une source d'erreur qui s'ajoute à l'erreur liée à l'échantillonnage. Il apparaît donc évident que le choix des distributions paramétriques dépend d'une certaine information que l'on doit détenir sur la fonction de distribution impliquée.

Les méthodes non paramétriques permettent d'éviter certains problèmes qui sont reliés aux méthodes paramétriques. En fait, les méthodes non paramétriques possèdent l'avantage de ne pas nécessiter d'hypothèses sur la distribution de la population. La fonction de densité non paramétrique n'a pas de forme particulière puisqu'elle est déterminée directement à partir des données, elle est donc beaucoup plus flexible que les distributions statistiques traditionnelles. Cette flexibilité devient relativement importante dans les extrémités où l'on a parfois de la difficulté à ajuster une loi, celle-ci étant ajustée sur la partie principale de la distribution et par conséquent, accorde un poids moindre aux extrémités pourtant importantes dans l'étude de fréquence de crue. Le problème majeur que l'on peut craindre en ce qui concerne les méthodes non paramétriques, est leur capacité à extrapoler. Comme

la distribution est estimée empiriquement on peut être porté à douter de la qualité de l'estimation au-delà des observations. L'étude de ce dernier point représente justement un des objectifs du présent travail.

Pour ce qui est du standard qui est recherché en hydrologie, les méthodes non paramétriques pourraient permettre d'en déterminer un qui soit efficace dans tous les cas. À la lumière de ce travail, il est souhaitable que l'on puisse recommander une procédure non paramétrique à suivre pour l'estimation de fréquence de crue. La méthode standard pourra alors être utilisée sans crainte de tomber sur un cas particulier puisque les méthodes non paramétriques, avec leur flexibilité, ont la qualité de s'ajuster à tous les échantillons. En utilisant une méthode non paramétrique, nous ne sommes pas contraint par la forme de la distribution comme on peut l'être avec les méthodes d'ajustements paramétriques traditionnels.

Dans ce chapitre, une des méthodes non paramétriques des plus répandues sera présentée : la méthode des noyaux. Une bonne revue de la littérature sur la méthode des noyaux a été effectuée par Lall (1995) et par Izenman (1991). Dans ce chapitre, on fera la distinction entre deux méthodes, celle dite à noyau simple et celle à fenêtre variable. Mais dans un premier temps, on présentera la méthode de l'histogramme qui représente un bon moyen d'introduire le concept de noyau.

3.1 Histogramme

L'une des premières méthodes d'estimation de densité à être utilisée est l'histogramme. En effet, l'histogramme, en dénombrant la quantité d'observations se trouvant dans un certain intervalle, permet d'avoir une estimation de la probabilité qu'un événement, se situant dans une certaine classe, se produise. Pour un échantillon de taille n, en utilisant un intervalle de largeur h, on déduit la fonction de densité de probabilité de la façon suivante :

$$\hat{f}(x)h = prob\left\{x - \frac{h}{2} \leq X \leq x + \frac{h}{2}\right\} \tag{3.1a}$$

$$\hat{f}(x)h = \frac{nombre\ de\ valeurs\ \left\{X_i \in \left[x - h/2, x + h/2\right]\right\}}{n} \tag{3.2b}$$

$$\hat{f}(x) = \frac{1}{nh} \times\ nombre\ de\ valeurs\ \left\{X_i \in \left[x - h/2, x + h/2\right]\right\} \tag{3.3c}$$

On peut exprimer cette fonction de façon plus générale en définissant les classes comprises sur tout le domaine comme suit :

$$C_j : \left[x_0 + (j-1)h; x_0 + jh\right[\tag{3.4}$$

où C_j représente la j^e classe telle que $\sum_j C_j$ couvre tout le domaine de X et x_0 représente le début de la première colonne de l'histogramme. On peut alors maintenant exprimer la fonction de densité d'une façon plus générale (Härdle, 1991):

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} \sum_j I\left(X_i \in C_j\right) I\left(x \in C_j\right) \tag{3.5}$$

où $I(A)$ est une variable dichotomique prenant respectivement les valeurs 1 ou 0 si l'énoncé A est respecté ou non. Dans cette expression, le paramètre h qui représente la largeur des classes, est une variable déterminante dans l'estimation. Cette méthode possède deux grands inconvénients, la fonction de densité est discontinue et le choix de l'origine de la première classe est délicat. Silverman (1986) montre par un exemple, que deux choix d'origine conduisent à des résultats relativement différents. Il est donc recommandé d'utiliser d'autres méthodes que celle de l'histogramme dans un contexte d'estimation de densité ou de fréquence. Cette méthode peut être très utile dans une analyse exploratoire ou pour la présentation d'un échantillon de données univariées, mais elle présente certains problèmes dans le cas multivarié.

3.2 Méthode des noyaux

Le concept de noyau a d'abord été introduit par Rosenblatt (1956), mais c'est Cacoullos (1966) qui a été le premier à utiliser le terme « noyau » (*kernel*) pour désigner la fonction de densité que l'on utilise dans les méthodes non paramétriques, remplaçant du même coup le terme « fonction-poids » (*weight function*) qui était généralement utilisé depuis Rosenblatt. En hydrologie statistique, c'est à S. Yakowitz et K. Adamowski que l'on doit la méthode des noyaux, qu'ils ont introduit indépendamment à une conférence de l'AGU en 1983 (Adamowski et Feluch, 1983 ; Yakowitz, 1983), dans un contexte d'estimation de fréquence de crue.

L'approche non paramétrique d'estimation de densité consiste à considérer une certaine fonction pour chacune des observations d'un échantillon de données, contrairement à l'approche paramétrique qui permet d'ajuster une seule distribution sur l'ensemble des observations. Au lieu de considérer une fonction dichotomique qui représente le nombre d'observations se trouvant dans une classe comme dans le cas de l'histogramme, la méthode des noyaux utilise plutôt une fonction K. Ces fonctions que l'on nomme noyau (*kernel function*), représentent le poids de chacune des observations dans l'estimation. L'estimation non paramétrique de la densité d'un échantillon peut se voir comme le cumul des fonctions K de chaque observation sur tout le domaine (figure 3.1).

Donc, l'approche non paramétrique requiert la sélection d'une fonction noyau K, et nécessite aussi le calcul d'un paramètre h qui détermine le degré de lissage de l'estimation, paramètre dont les propriétés seront détaillées dans la section 3.2.4. L'expression de la fonction de densité non paramétrique est relativement similaire à celle de l'histogramme (3.5), à la différence que l'on remplace la fonction dichotomique $I(A)$ par la fonction noyau

K. En divisant les « poids » par *nh*, la fonction de densité devient donc une sorte de moyenne pondérée sur tout l'échantillon:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{3.6}$$

où n représente la taille de l'échantillon $\{x_1, x_2, ..., x_n\}$, h représente le paramètre de lissage et K représente le noyau. Contrairement à l'histogramme, la fonction de densité obtenue avec la méthode des noyaux est continue lorsque l'on considère un noyau continu. De plus, on n'a pas non plus le problème du choix de l'origine.

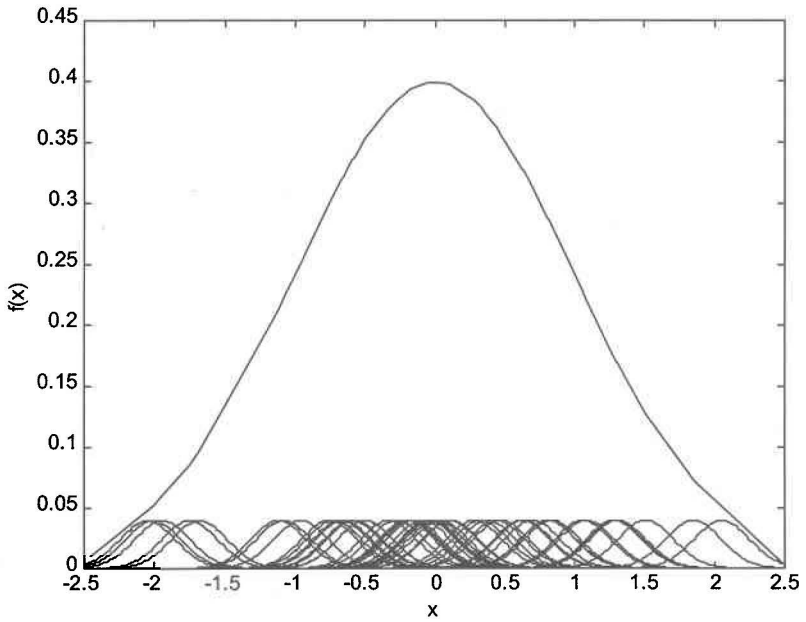


Figure 3.1 Illustration de la méthode des noyaux pour l'estimation d'une fonction de densité.

3.2.1 Notion de noyau *K*

La fonction noyau est l'élément qui détermine la forme des « bosses » qui, une fois cumulées, constituent l'estimation de la densité. Dans la littérature, on a généralement stipulé que le choix du noyau K n'a pas vraiment d'importance, qu'il est choisi de façon relativement arbitraire et que c'est plutôt le choix du paramètre de lissage h, qui détermine l'étendue du noyau de chaque côté de l'observation, qui est primordial. De nombreux travaux ont été effectués par plusieurs chercheurs dans le but de déterminer l'importance du choix du noyau et dans l'espoir de trouver un noyau optimal qui serait préférable d'utilisation à tous les autres. Epanechnikov (1969), en s'appuyant sur certains critères, a suggéré un noyau optimal qui porte maintenant son nom. Rao (1983) a montré que l'utilisation d'un noyau autre que le noyau optimal, ne menait qu'à une faible perte de précision. Lall *et al.* (1993), quant à eux, considèrent que le choix du noyau a une certaine importance, mais que l'influence sur l'ensemble de l'estimation est faible. Malgré tout, il serait intéressant d'évaluer la performance de certains types de noyaux par rapport à d'autres dans différentes circonstances, notamment dans l'extrapolation de la distribution empirique.

3.2.2 Propriété des noyaux

En théorie, toute distribution qui possède une fonction de probabilité définie peut être utilisée comme noyau dans l'estimation non paramétrique de densité. En partant de la définition même d'une fonction de densité, le noyau doit être positif en tout point et l'intégration de tout les points doit être égale à 1. Lorsque l'on utilise un noyau qui répond aux exigences des fonctions de densité, on en déduit à partir de l'équation 3.4 que $\hat{f}(x)$ est aussi une fonction de densité. De plus, $\hat{f}(x)$ hérite de toutes les propriétés de continuité et de différentiabilité de sa fonction noyau (Silverman , 1986). On verra plus loin qu'il est possible d'utiliser d'autres fonctions qui ne sont pas nécessairement des densités.

3.2.2.1 Noyau optimal basé sur le critère de l'*IMSE*

Epanechnikov (1969) cite des travaux effectués par d'autres chercheurs sur des fonctions noyau (3.6). Ces auteurs travaillaient généralement avec un noyau de forme arbitraire. Dans son travail, Epanechnikov, étudie les propriétés de la fonction de densité empirique sujette à des noyaux de forme arbitraires. Il étudie d'abord les propriétés asymptotiques de la fonction de densité empirique et ensuite il étudie l'erreur induite par l'estimation $\hat{f}$ sur la véritable fonction de densité f. À partir de cette fonction erreur, il tente de déterminer une forme de noyau optimal que l'on pourrait utiliser plutôt qu'un noyau arbitraire.

Tout d'abord, il impose certaines contraintes sur le noyau. En fait, il faut que le noyau soit symétrique (3.7), qu'il soit non-négatif sur tout le domaine D (3.7), que l'intégration sur tout son domaine D soit unitaire (3.8), qu'il ait une moyenne nulle (3.9) et une variance finie (3.10) et qu'il ait des moments donnés par (3.11):

$$K(x) = K(-x) \geq 0 \tag{3.7}$$

$$\int_D K(x)dx = 1 \tag{3.8}$$

$$\int_D x\,K(x)\,dx = 0 \tag{3.9}$$

$$\int_D x^2 K(x)dx \leq a \tag{3.10}$$

$$\int_D x^m K(x)dx \leq \infty \qquad 0 \leq m < \infty \tag{3.11}$$

Ces contraintes semblent n'être utilisées que pour simplifier les calculs d'optimisation du noyau (Silverman, 1986). Epanechnikov (1969) exhibe un noyau optimal sous ces

contraintes en minimisant l'erreur quadratique moyenne intégrée (*integrated mean square error, IMSE*) :

$$IMSE = \int_{-\infty}^{\infty} E\left[\left(\hat{f}(x) - f(x)\right)^2\right] dx \tag{3.12}$$

Si on développe cette expression, on trouve qu'elle est formée de deux expressions connues, la variance et le biais (voir annexe A):

$$IMSE = \int_{-\infty}^{\infty} \text{var}\left[\hat{f}(x)\right] + biais\left[\hat{f}(x)\right]^2 \; dx \tag{3.13}$$

Il s'agit alors ensuite de trouver une expression pour le biais et la variance intégrés de l'estimation de la fonction de densité (Silverman ; 1986) :

$$\int biais\left(\hat{f}(x)\right)^2 dx = \int \left\{E\left[\hat{f}(x)\right] - f(x)\right\}^2 \; dx$$

$$= \int \left\{\int \frac{1}{h} K\left(\frac{x-y}{h}\right) f(y) \; dy - f(x)\right\}^2 \; dx \tag{3.14}$$

Puisque l'on a de façon générale (Whittle ; 1958):

$$E\left[\hat{f}(t)\right] = \frac{1}{n}\sum E\left[K\left(X_i, t\right)\right] = \int K(x,t) f(x) \; dx \tag{3.15}$$

En effectuant le changement de variable $y = x - ht$ on obtient :

$$\int biais\left(\hat{f}(x)\right)^2 dx = \int \left(\int K(t) f(x - ht) \; dt - f(x)\right)^2 dx \tag{3.16}$$

En utilisant la propriété 3.6 qui stipule que l'intégration du noyau sur l'ensemble du domaine est unitaire, on obtient :

$$\int biais\left(\hat{f}(x)\right)^2 dx = \int \left(\int K(t)\left[f(x-ht) - f(x)\right]\, dt \right)^2 \, dx \qquad (3.17)$$

On peut développer ensuite le terme $f(x-ht)$ en série de Taylor autour du point x:

$$f(x-ht) = f(x) - ht\, f'(x) + \frac{1}{2}h^2 t^2 f''(x) + o\left(h^2\right)$$

$$\int biais\left(\hat{f}(x)\right)^2 dx = \int \left(-h\, f'(x) \int tK(t)\, dt + \frac{1}{2}h^2 f''(x) \int t^2 K(t)\, dt + \ldots \right)^2 dx$$

En utilisant la propriété (3.9) et en négligeant les termes d'ordre supérieur on a l'approximation suivante:

$$\int biais\left(\hat{f}(x)\right)^2 dx \approx \int \left(\frac{1}{2}h^2 f''(x) k_2 \right)^2 dx \qquad (3.18)$$

où $k_2 = \int t^2 K(t) dt$ est une constante qui représente la variance du noyau K. En intégrant par rapport à x, on obtient finalement une approximation du biais au carré intégré :

$$\int biais\left(\hat{f}(x)\right)^2 dx \approx \frac{h^4}{4} k_2^2 \int \left[f''(x)\right]^2 dx \qquad (3.19)$$

Pour ce qui est de la variance on a (Whittle ; 1958):

$$\mathrm{var}\left[\hat{f}(t)\right] = \frac{1}{n}\mathrm{var}\left[K\left(X_i, t\right)\right] = \frac{1}{n}\left[\int \left[K(x,t)\right]^2 f(x)\, dx - \left\{ \int K(x,t) f(x)\, dx \right\}^2 \right] \qquad (3.20)$$

En effectuant le changement de variable $y = x - ht$, on obtient :

$$\int \mathrm{var}\left[\hat{f}(x)\right] dx = \int \frac{1}{n}\left[\frac{1}{h} \int K^2(t) f(x-ht)\, dt - \left\{ \int K(t) f(x-ht)\, dt \right\}^2 \right] dx \qquad (3.21)$$

Comme l'intégration du premier terme par rapport à x est égale à un, on a :

$$\int \mathrm{var}\left[\hat{f}(x)\right] dx = \frac{1}{nh}\int K^2(t)\,dt - \int\left[\int K(t) f(x-ht)\,dt\right]^2 dx \qquad (3.22)$$

Dans l'expression précédante, on retrouve l'expression du biais :

$$\int \mathrm{var}\left[\hat{f}(x)\right] dx \approx \frac{1}{nh}\int K^2(t)\,dt - \frac{1}{n}\int\left\{biais\left[\hat{f}(x)\right] + f(x)\right\}^2 dx \qquad (3.23)$$

En utilisant l'expression (3.18), on peut faire une approximation du 2e terme (Silverman ; 1986) :

$$\int \mathrm{var}\left[\hat{f}(x)\right] dx \approx \frac{1}{nh}\int K^2(t)\,dt - \frac{1}{n}\int\left\{f(x) + o\left(h^2\right)\right\}^2 dx \qquad (3.24)$$

Finalement, en supposant que pour $n \to \infty$, $h \to 0$ et $nh \to \infty$, on obtient une approximation de la variance intégrée de l'estimation:

$$\int \mathrm{var}\left[\hat{f}(x)\right] dx \approx \frac{1}{nh}\int\left[K(t)\right]^2 dt \qquad (3.25)$$

En combinant les équations (3.19) et (3.25), on obtient une approximation de l'expression de l'erreur moyenne intégrée et en dérivant celle-ci par rapport à h et en égalant à 0, on obtient la valeur h_{opt} qui minimise l'erreur moyenne intégrée :

$$IMSE = \frac{1}{4}h^4 k_2^2 \int\left[f''(x)\right]^2 dx + \frac{1}{nh}\int\left[K(t)\right]^2 dt \qquad (3.26)$$

$$h_{opt,IMSE} = \left\{\frac{\int\left[K(t)\right]^2 dt}{nk_2^2\int\left[f''(x)\right]^2 dx}\right\}^{1/5} \qquad (3.27)$$

On remarque que l'expression du paramètre de lissage optimal au sens de l'erreur moyenne intégrée dépend encore de la fonction de densité « théorique » inconnue, à cause du terme $f''(x)$. On ne peut par conséquent calculer directement le paramètre de lissage optimal avec cette expression. Le chapitre 5 présentera certaines méthodes utilisées pour le calcul de h

dont l'une qui est basée sur la minimisation de l'erreur moyenne intégrée. En substituant dans l'équation 3.11 les valeurs des équations 3.23 et 3.16 obtenues pour un h optimal donné par 3.25, on a la valeur approximative de l'*IMSE* suivante (Silverman ; 1986):

$$IMSE_{h_{opt}} \approx \frac{5}{4n^{-4/5}} k_2^{2/5} \left\{ \int [K(t)]^2 dt \right\}^{4/5} \left\{ \int [f''(x)]^2 dx \right\}^{1/5} \qquad (3.28)$$

En examinant cette expression, on remarque que le seul moyen de réduire l'erreur moyenne intégrée, c'est de choisir un noyau K qui minimise le terme $\int [K(t)]^2 dt$ puisque l'on n'a aucun contrôle sur le terme $f''(x)$ et que le paramètre de lissage a déjà été optimisé. Le problème du choix du noyau optimal au sens de l'*IMSE* se résume donc à :

minimiser : $\int K(t)^2 dt$

sous les contraintes : (3.7), (3.8), (3.9) et (3.10)

Ce problème peut être résolu en considérant la forme polynomiale d'Euler (Epanechnikov ; 1969). La fonction résultant de cette optimisation porte le nom de noyau Epanechnikov et est donnée par:

$$K(y) = \begin{cases} \frac{3}{4\sqrt{5}} - \frac{3y^2}{20\sqrt{5}} & pour \quad -\sqrt{5} \le x \le \sqrt{5} \\ 0 & ailleurs \end{cases} \qquad (3.29a)$$

En effectuant le changement de variable $x = y/\sqrt{5}$, on obtient la forme suivante qui est généralement utilisée:

$$K(x) = \begin{cases} \frac{3}{4}(1-x^2) & pour \quad -1 \le x \le 1 \\ 0 & ailleurs \end{cases} \qquad (3.29b)$$

Il est possible d'évaluer l'efficacité des autres noyaux par rapport au noyau optimal. Il suffit de comparer la valeur $\int K_e(t)^2 dt$ du noyau Epanechnikov à la valeur $\int K(t)^2 dt$ du noyau

dont on veut calculer l'efficacité. Silverman (1986) définit l'efficacité d'un noyau K quelconque par rapport au noyau Epanechnikov K_e de la façon suivante :

$$eff(K) = \frac{3}{5\sqrt{5}} \left\{ \int t^2 K(t) dt \right\}^{-1/2} \left\{ \int K(t)^2 dt \right\}^{-1} \tag{3.30}$$

La constante $\frac{3}{5\sqrt{5}}$ est la valeur de $\int K_e(t)^2 dt$ du noyau Epanechnikov. Plus la valeur de $eff(K)$ se rapproche de 1, plus le noyau K est comparable au noyau optimal. L'efficacité à été calculée pour plusieurs noyaux et il semblerait que la plupart des noyaux soient comparables au noyau Epanechnikov. On a par exemple une efficacité de 0.9512 pour le noyau normal, de 0.9859 pour le noyau triangulaire et de 0.9295 pour le noyau rectangulaire (ces noyaux seront présentés au tableau 3.1 à la figure 3.2 de la section 3.2.3). Il n'apparaît donc pas nécessaire d'appuyer notre choix d'un noyau sur la minimisation de l'erreur moyenne intégrée, puisqu'il semble qu'on ne perde que très peu de précision en utilisant un autre noyau que le noyau optimal. Il serait donc préférable de choisir plutôt un noyau qui convient au type d'estimation que l'on veut effectuer.

3.2.2.2 Autres critères d'optimisation

On a vu que Epanechnikov (1969) a proposé un noyau qui est optimal au sens de l'erreur quadratique moyenne intégrée, c'est à dire le noyau qui minimise l'*IMSE*. Cette fonction d'erreur se calcule à partir de l'estimation de la fonction de densité. On peut donc dire que le noyau Epanechnikov est le noyau qui minimise l'erreur lors de l'estimation d'une densité de probabilité. En hydrologie, on s'intéresse habituellement à l'estimation des quantiles de période de retour T plutôt qu'à la fonction de densité elle-même. C'est pourquoi il n'est pas certain que le noyau Epanechnikov soit vraiment le noyau optimal à utiliser dans ce contexte. Il s'agirait donc d'optimiser plutôt une fonction d'erreur calculée à partir de l'estimation des quantiles. Le noyau optimal résultant devrait théoriquement être plus efficace que le noyau Epanechnikov. D'autres critères pourraient être optimisés comme par exemple un critère qui ne considère l'erreur que dans les extrémités de la fonction de

distribution ou bien qui confère un poids supérieur à l'extrémité d'intérêt. Dans la littérature, il n'y a apparemment peu d'autres études que celle d'Epanechnikov qui tente de déterminer un noyau optimal. La raison en est probablement que la recherche s'est concentrée sur les techniques de calcul du paramètre de lissage optimal plutôt que sur les noyaux. Comme la plupart des chercheurs dans le domaine stipulent que le choix du noyau importe peu, il semble y avoir peu d'intérêt à déterminer un noyau optimal si on peut en utiliser un autre sans une grande perte de précision. Mais comme il a été mentionné précédemment, le choix du noyau peut avoir une certaine importance selon le contexte.

3.2.3 Types de noyaux

Il existe donc une diversité de noyaux qui peuvent être utilisés sans crainte de perte de précision. On croit toujours que le choix du noyau n'est pas nécessairement important et qu'il suffit qu'il convienne aux exigences (3.7) à (3.11) pour qu'il soit efficace. Mais certains groupes de noyaux ont des caractéristiques que les autres n'ont pas et ils semblent plus préférables dans certaines situations plus que d'autres. Dans un contexte d'interpolation, il est vrai que la forme du noyau peu avoir une faible influence sur l'estimation. Toutefois, le domaine de variation du noyau peut avoir une certaine importance sur la capacité de la méthode à extrapoler. La méthode des noyaux a jusqu'à maintenant surtout été utilisée dans un contexte d'interpolation, ce qui pourrait expliquer le fait que dans la littérature on stipule généralement que le choix du noyau est sans importance. Dans les prochaines sections, on a distingué les noyaux par deux facteurs, le support et la symétrie, et on discute des situations où l'on a avantage à utiliser chacun des types. La figure 3.2 présente les différents noyaux considérés pour cette étude.

3.2.3.1 Noyaux à support fini et à support non-fini

Les noyaux à support fini sont des noyaux qui ne sont définis que sur un domaine restreint de $\Re$. Ce sont donc des noyaux qui sont bornés auxquels on attribue une valeur nulle à

l'extérieur de l'intervalle de définition. Par contre, les noyaux à support non-fini sont asymptotiques et sont donc non-nuls sur l'ensemble des réels $\mathbb{R}$:

$$\text{(a) } K_f(t) = 0 \qquad \text{pour} \qquad t \notin [-a; a]$$

$$\text{(b) } K_{nf}(t) \to 0 \qquad \text{pour} \qquad |t| \to \infty \tag{3.31}$$

où K_f et K_{nf} représentent un noyau à support fini et un noyau à support non-fini, respectivement et $[-a; a]$ représente le domaine de définition du noyau à support fini. Le tableau 3.1 contient les noyaux qui sont utilisés dans cette étude, classés selon le type de support.

Tableau 3.1. Noyaux classés selon le support.

Noyaux à support fini		**Noyaux à support non-fini**	
Epanechnikov	$\frac{3}{4}(1-t^2)$ pour $\lvert t\rvert<1$	Normal	$\frac{1}{\sqrt{2\pi}}e^{-(1/2)t^2}$
Rectangulaire	$\frac{1}{2}$ pour $\lvert t\rvert<1$	Cauchy	$\frac{1}{\pi(1+t^2)}$
Biweight	$\frac{15}{16}(1-t^2)^2$ pour $\lvert t\rvert<1$	EV1	$e^{-t-e^{-t}}$

Le fait qu'un noyau soit borné ou non borné peut être important dans l'estimation, surtout dans l'extrémité droite (ou gauche) de la fonction de densité. Par exemple, lorsque l'on est intéressé à estimer la queue de droite d'une fonction au delà des observations mises à notre disposition, c'est à dire que l'on est dans une situation d'extrapolation, les noyaux à support finis peuvent être inefficaces. Les estimations faites au delà des observations proviennent principalement des dernières données de l'échantillon et si à chacune de ces valeurs, on associe un noyau qui devient nul au delà de la largeur du paramètre de lissage, on limite considérablement la capacité d'extrapolation. Par contre, les noyaux asymptotiques

permettent de sortir plus loin en dehors de l'échantillon, puisqu'ils sont non-nuls au delà de la largeur du paramètre de lissage.

3.2.3.2 Noyaux symétriques et asymétriques

Dans la littérature, la plupart des noyaux utilisés dans un contexte d'estimation de fonction de densité sont symétriques. Même si selon les exigences établies par Epanechnikov, on doit utiliser des noyaux symétriques, l'utilisation de noyaux asymétriques pourrait être favorable à l'estimation dans la queue de droite si l'asymétrie favorise l'extrémité de droite du noyau en accroissant la longueur de celle-ci par rapport à celle de gauche. De cette façon, on attribue un poids supérieur au côté droit de la distribution et on peut ainsi augmenter le degré d'extrapolation. Evidemment, ces noyaux risquent de causer un biais dans la partie inférieure de l'estimation, contrairement aux noyaux symétriques qui conduisent à une estimation non biaisée à chaque observation, mais ils pourraient être utilisés lorsque l'on désire estimer des quantiles à grande période de retour. Lall *et al.* (1993) prétendent même que les noyaux asymétriques permettent de réduire le biais dans l'estimation pour les dernières observations de l'échantillon. Toutefois, il a été démontré (Lall *et al.* ; 1993) que l'utilisation d'un noyau symétrique dans le cas où l'on a une densité théorique symétrique, conduit à un biais nul à l'origine, si le noyau et la densité sont symétriques autour de 0. Il est donc préférable d'utiliser des noyaux symétriques dans ces cas.

On peut identifier deux types d'asymétrie, l'asymétrie positive et l'asymétrie négative. On différencie les deux types en examinant le coefficient d'asymétrie :

$$\int_{-a}^{a} \left\{ K_{as}(t) - E\left[K_{as}(t)\right] \right\}^3 dt \qquad (3.32)$$

où K_{as} est un noyau asymétrique et où $[-a; a]$ est sont domaine de variation s'il est borné. Si cette quantité (3.32) est positive, la fonction K est asymétrique positive, sinon elle est asymétrique négative. Bien sûr si l'expression 3.32 est nulle, le noyau est symétrique.

En analyse de fréquence de crue, on a des distributions empiriques qui possèdent une borne inférieure, c'est à dire que l'on a des débits qui sont généralement plus grand que 0. Le fait que la densité théorique ne soit bornée que d'un côté peut nous laisser croire qu'elle peut être parfois asymétrique. Donc, comme on est moins certain de la symétrie de la distribution théorique, il semble y avoir un risque moins grand à utiliser un noyau asymétrique. Parmi les noyaux présentés au tableau 3.1, seul le noyau EV1 est asymétrique autour de 0. Comme on peut le constater sur la figure 3.2, l'extrémité droite du noyau EV1 est relativement plus lourde que celle de gauche.

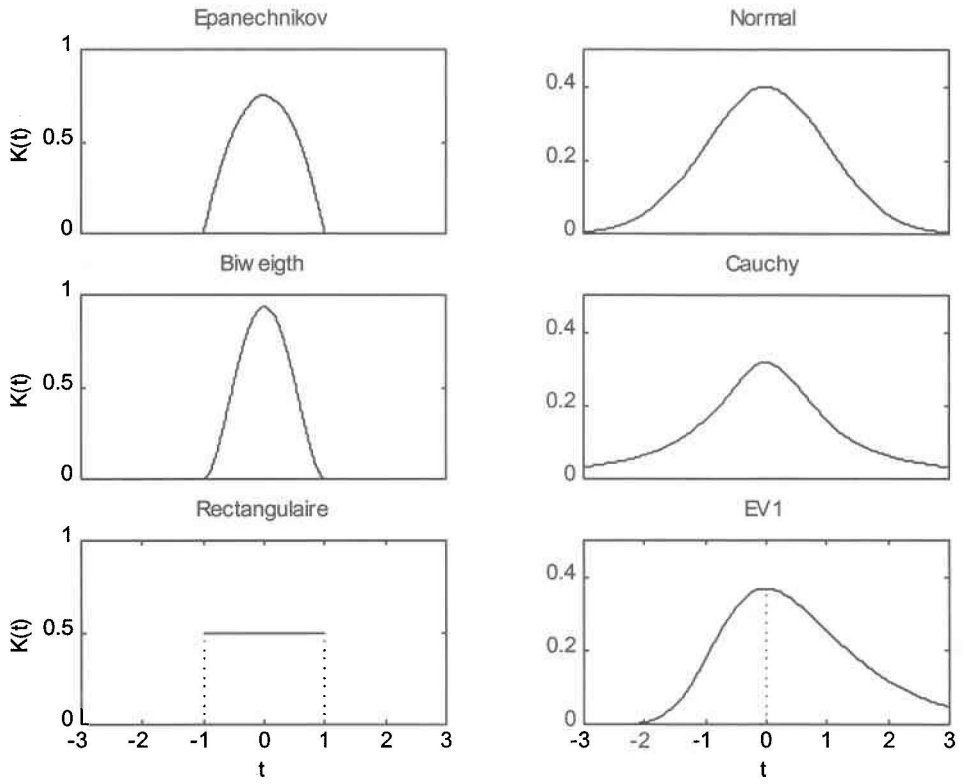


Figure 3.2 : Représentation graphique des six noyaux considérés dans l'étude.

3.2.4 Paramètre de lissage

Le paramètre de lissage h de la méthode des noyaux joue sensiblement le même rôle que dans le cas de l'histogramme, c'est-à-dire qu'il agit comme une sorte de fenêtre. Alors que le noyau détermine la forme des « bosses » de la figure 3.1, le paramètre de lissage quant à lui, en détermine la largeur. On peut considérer le même paramètre de lissage pour tout l'échantillon, on prend alors la même taille de fenêtre pour chaque observation. Cette approche, communément appelée méthode d'estimation à noyau fixe (*fixed kernel estimator*) eq (3.6), fait en sorte que toutes les observations ont le même noyau K.

Même si les avis sont partagés quant à l'importance du choix du noyau à utiliser pour l'estimation, tout le monde s'entend pour dire que le choix du paramètre de lissage est relativement plus important que le choix du noyau lui-même. L'importance de l'estimation du paramètre de lissage a eu pour conséquence de concentrer la recherche depuis plusieurs années, sur les techniques d'estimation du paramètre de lissage optimal. Par conséquent, il existe une grande variété de méthodes de calcul de h, les plus populaires feront l'objet du chapitre suivant. On a vu que l'expression (3.25) nous donnait la valeur du paramètre de lissage qui rend la moyenne des erreurs quadratiques intégrée (*IMSE*) minimale. Comme cette expression dépend de la fonction théorique $f(x)$ inconnue, la valeur de h doit être déduite empiriquement directement à partir de la série de données observées.

3.2.5 Propriétés du paramètre de lissage

C'est le paramètre de lissage qui détermine si l'estimation est tout à fait lisse ou bien qu'elle est plutôt très irrégulière. La figure 3.3 illustre bien le rôle que joue h dans le lissage de l'estimation. Lorsque l'on a une valeur de h faible, le degré de lissage est faible et on a une estimation irrégulière. À l'opposé, lorsque l'on a une valeur de h élevée, l'estimation est très lisse. Donc, lorsque $h \to 0$ l'estimation de la fonction est très bruitée, l'estimation en

un point provient principalement du point lui-même, la contribution des autres points étant très faible. Lorsque $h \to \infty$, tous les points tendent à contribuer de façon égale et l'estimation devient une constante sur tout l'échantillon. Un lissage trop faible donne trop d'importance à chacune des observations et peut faire en sorte que des valeurs singulières nuisent à l'estimation. Si on lisse trop, on risque de camoufler certaines particularités de la fonction théorique, comme une bimodalité par exemple. Il est donc crucial de choisir un paramètre de lissage qui convienne au type d'estimation.

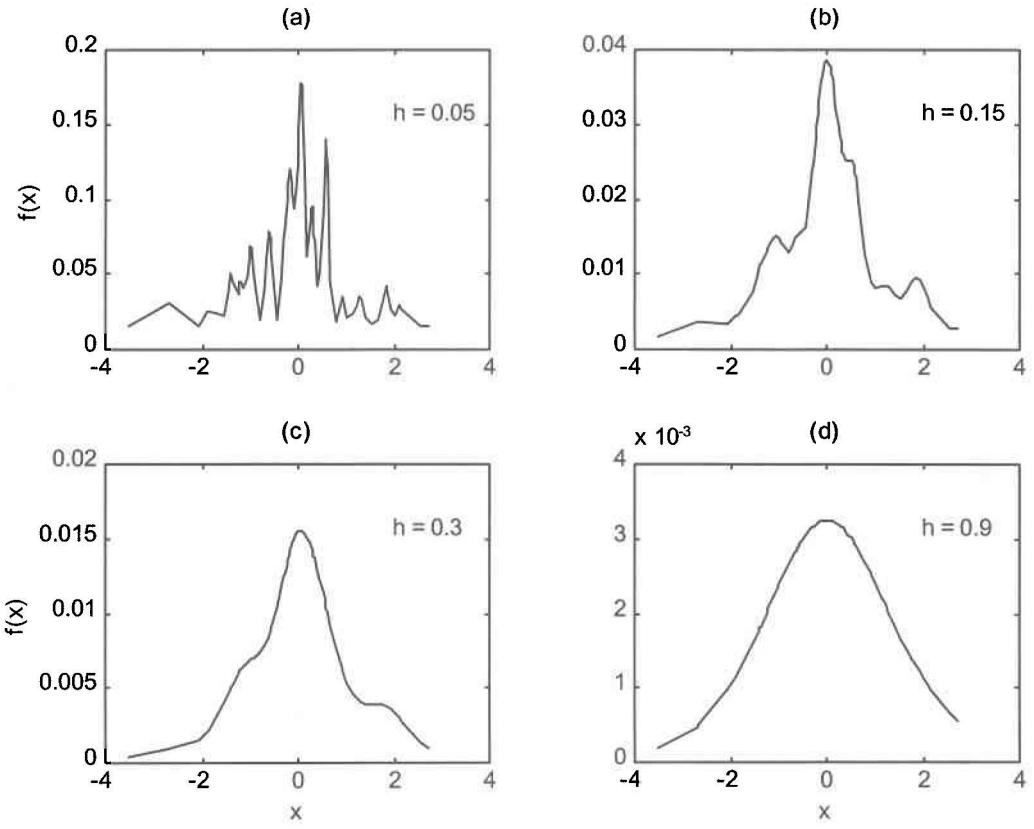


Figure 3.3 Illustration de l'effet du paramètre de lissage sur l'estimation de la fonction de densité.

En considérant un noyau quelconque, le noyau optimal d'Epanechnikov par exemple, on peut analyser le rôle que joue le paramètre de lissage dans la fonction d'erreur *IMSE* définie

en (3.26). Mais on utilisera plutôt la fonction *MSE* pour évaluer l'influence de *h* sur l'erreur localement :

$$MSE_f(x) \approx \frac{1}{nh} f(x) \int K(t)^2 dt + \frac{h^4}{4} \left\{ f''(x) \int t^2 K(t) dt \right\}^2 \qquad (3.33)$$

où $\int MSE\ dx = IMSE$. Il est important de remarquer que dans l'expression (3.33) la valeur de h n'a pas été remplacée par la valeur optimale h_{opt} contrairement à ce qui avait été fait pour l'*IMSE* donnée par (3.28), l'objectif étant d'analyser le comportement de l'erreur face à la valeur de *h*, ce dernier doit être variable.

On note d'abord que si $n \to \infty$, $h \to 0$ et $nh \to \infty$, alors *MSE* converge vers 0. L'efficacité de la méthode des noyaux repose donc en partie sur la taille de l'échantillon, plus l'échantillon est grand, meilleures sont nos chances d'avoir une erreur faible. L'expression (3.33) est formée de 2 termes, le premier représente la variance qui dépend de $f(x)$ et le deuxième le biais qui dépend de $f''(x)$. On remarque facilement que dans les zones où la densité $f(x)$ est élevée, on a une variance plus grande. De plus, lorsqu'on estime la fonction f à un endroit où on a une dérivée seconde positive, on a un biais positif (surestimation). Dans les cas où on a une dérivée seconde négative, on a une sous-estimation. L'impact du signe du biais sur l'erreur est nul, puisque le terme du biais est élevé à la puissance 2. Par conséquent, un biais négatif augmente l'erreur de la même façon qu'un biais positif. Par ailleurs, on remarque que le paramètre de lissage n'est pas du même ordre dans le cas de la variance que dans le cas du biais. Dans le cas de la variance, puisque le paramètre de lissage est à la puissance -1, on aurait tendance à choisir une valeur élevée de *h* pour réduire la variance.

Pour ce qui est du biais, une valeur faible de *h* le réduit étant donné la puissance de 4 du paramètre de lissage. On peut donc généraliser en disant que le terme de la variance sert à pénaliser les cas où on ne lisse pas assez, en accroissant la valeur de *MSE* et de la même

façon, le terme du biais a pour fonction de pénaliser les cas où on lisse trop notre estimation (Härdle; 1991).

À partir de l'équation (3.33), on peut déterminer la valeur de h qui rend l'erreur quadratique moyenne minimale. Comme dans le cas de l'*IMSE*, il suffit de dériver (3.33) par rapport à h et d'annuler l'expression obtenue:

$$h_{opt,MSE} = \left[\frac{f(x) \int K(t)^2 dt}{n \left(f''(x) \int t^2 K(t) dt \right)^2} \right]^{1/5} \quad \text{si} \quad f''(x) \neq 0 \qquad (3.34)$$

En examinant cette expression, on remarque que le paramètre h est proportionnel à la densité théorique $f(x)$ et par conséquent, que pour une valeur de densité faible, on a un paramètre optimal faible. Mais, dans les zones de faible densité, généralement les extrémités, on sait que la fenêtre h doit être suffisamment large afin d'éviter des discontinuités dans la fonction estimée ainsi que pour limiter l'influence des valeurs singulières. Il y a donc une contradiction entre l'expression (3.34) et le degré de lissage souhaité dans les extrémités. Par contre, en examinant le dénominateur de (3.34), on note que pour une dérivée seconde élevée le paramètre de lissage est faible, ce qui est souhaitable dans la région du mode de la distribution, zone où il y a une forte concentration d'observations et où une large fenêtre n'est pas nécessaire. Ces remarques permettent de mettre en évidence la différence entre le lissage dans les extrémités et le lissage dans la partie centrale de l'échantillon. Mais le paramètre optimal qui provient du *MSE* est le même pour tout l'échantillon. La section suivante introduit une méthode qui permet de considérer le paramètre de lissage variable d'une observation à l'autre.

3.3 Méthode à fenêtre variable

La méthode à noyau fixe possède l'inconvénient de considérer la partie centrale et les extrémités de la densité de la même façon. Dans les zones où la densité est élevée, c'est-à-

dire où les données sont relativement concentrées, le paramètre h peut être trop grand et par conséquent, la densité estimée risque d'être trop lissée. À l'opposé, dans les régions où les données sont moins fréquentes, c'est à dire les extrémités, on risque de " sous-lisser " la distribution, puisque le paramètre de lissage est trop faible. Dans ce cas particulier, on risque d'accorder trop d'importance aux valeurs extrêmes. La méthode d'estimation à fenêtre variable de Breiman *et al.* (1977) (*variable kernel estimator*) possède l'avantage de pouvoir utiliser un paramètre de lissage h variable selon la densité locale.

Selon Breiman *et al.* (1977), il est évident que la méthode d'estimation habituelle à noyau fixe ne peut répondre adéquatement aux variations de l'amplitude de la fonction de densité à estimer. Par exemple, si on se trouve dans une zone où la densité théorique est faible et qu'il n'y a qu'une seule observation dans l'échantillon, on se retrouvera alors avec un pic dans notre estimation alors que la densité est peut-être plus lisse dans cette région. Selon ces auteurs, aucune des méthodes existantes permet de prétendre avec certitude que le paramètre de lissage optimal h que l'on obtient est vraiment celui qui permet d'estimer la densité le plus parfaitement possible.

Avec la méthode à fenêtre variable, les paramètres de lissage dépendent de la distance d'une observation à ses k voisins les plus proches. Il propose donc d'utiliser un paramètre de lissage variable h_i que l'on estime par $a_k d_{ki}$. Le paramètre a_k est une composante multiplicative et d_{ki} représente la distance entre l'observation i et son k^e voisin le plus proche. De cette façon, dans les zones où il y a peu de données, c'est-à-dire des régions à faible densité, les valeurs de d_{ki} sont élevées, alors que l'on observe l'inverse dans le cas des régions à forte densité. La formule d'estimation de la fonction de densité (3.4) prend la forme suivante:

$$\hat{f}(x) = \sum_{i=1}^{n} \frac{1}{n a_k d_{ki}} K\left(\frac{x - x_i}{a_k d_{ki}}\right) \tag{3.35}$$

Les méthodes d'estimation des paramètres k et a_k seront présentées à la section 4.6. Breiman et *al.* (1977) ont comparé cette méthode à fenêtre variable à la méthode à noyau fixe traditionnelle, en utilisant des données simulées à partir de la distribution normale bivariée. En connaissant la distribution théorique des données, ils ont pu calculer les paramètres optimaux k et a_k en minimisant directement l'erreur de l'estimation. Dans tous les cas qu'ils ont étudiés, la méthode à fenêtre variable a été supérieure à celle à noyau fixe. Toutefois, ils ont remarqué que les voisins les plus proches qui ont été retenus pour l'estimation étaient très éloignés des observations, c'est-à-dire que les distances d_{ki} étaient relativement élevées. De plus, l'estimation pouvait encore être améliorée, même au-delà du 100^e voisin le plus proche (k = 100). Dans certaines simulations, la méthode à fenêtre variable semblait même insensible à la valeur de k.

3.4 Fonction de répartition et estimation d'un quantile

Dans l'estimation de la fréquence des crues, nous désirons estimer la fonction de répartition $F(x)$ (fonction de distribution cumulée). Pour ce faire, on peut considérer l'estimation non paramétrique $\hat{F}(x)$ suivante, qui est tout simplement l'intégrale de la fonction de densité non paramétrique (3.6):

$$\hat{F}(x) = \int_{-\infty}^{x} \sum_{i=1}^{n} \frac{1}{nh} K\left(\frac{t - x_i}{h}\right) dt \qquad (3.36)$$

En sortant de l'intégrale les termes indépendants de t et en effectuant le changement de variable $w = \frac{t - x_i}{h}$, on obtient :

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} K_I\left(\frac{x - x_i}{h}\right) \qquad (3.37)$$

où $K_I(u) = \int_{-\infty}^{u} K(w)\, dw$ et $u = \frac{x - x_i}{h}$

Il s'agit donc d'intégrer chacune des fonctions noyaux pour pouvoir estimer le fonction de répartition non paramétrique. Mais comme on l'a vu au chapitre 2, en hydrologie on doit souvent estimer des quantiles de période de retour T tels que.

$$F(x_T) = 1 - \frac{1}{T} \tag{3.38}$$

L'estimation d'un quantile peut être définie comme étant l'inverse de l'estimation de la fonction de répartition, mais il est assez difficile d'isoler la valeur x dans l'expression (3.35). On peut cependant, par un processus itératif, trouver le débit x_T de période de retour T avec une certaine précision.

Dans cette étude, on utilise la méthode itérative fort simple du demi-intervale (ou bisection). Elle consiste tout d'abord à considérer deux valeurs extrêmes de débit, en s'assurant que la valeur cible x_T se trouve entre ces deux valeurs. Pour ce faire, on calcule la valeur de la fonction de répartition pour les deux valeurs extrêmes et ensuite on s'assure que la valeur de la fonction de répartition correspondant à la période de retour T désirée, se trouve entre les deux :

$$F(x_{\text{inf}}) < F(x_T) < F(x_{\text{sup}}) \quad \text{où} \quad x_{\text{inf}} < x_T < x_{\text{sup}} \tag{3.39}$$

Ensuite, on divise l'intervalle $[x_{\text{inf}}, x_{\text{sup}}]$ en deux parties égales, x_k étant le centre de l'intervalle, et on compare $F(x_T)$ avec $F(x_k)$. Si $F(x_T)$ est inférieur à $F(x_k)$, on conserve x_{inf} et on change x_{sup} par la valeur du demi-intervale x_k, sinon on conserve x_{sup} et on change x_{inf} par la valeur du demi-intervalle :

$$\text{(a)}\ x_k = \frac{x_{\text{sup}} - x_{\text{inf}}}{2} \tag{3.40}$$

(b) Si $F(x_T) < F(x_k) \Rightarrow$ intervalle $[x_{\inf}, x_k]$

Si $F(x_T) > F(x_k) \Rightarrow$ intervalle $[x_k, x_{\sup}]$

De cette façon, on redéfinit un nouvel intervalle contenant la valeur x_T à estimer. On vérifie alors si le centre du nouvel intervalle se rapproche de la valeur x_T en comparant les valeurs de la fonction de répartition. On répète la procédure (3.38) jusqu'à ce que l'on atteigne la précision désirée :

$$\left| F(x_T) - F(x_k) \right| < \text{précision voulue}$$

Cette procédure permet de réduire l'intervalle de plus en plus jusqu'à ce que l'on tombe sur la bonne valeur. La valeur x_k du dernier intervalle représente alors la valeur x_T obtenue par interpolation.

On peut aussi définir l'estimation non paramétrique d'un quantile comme étant la moyenne non paramétrique de la fonction quantile empirique. Sheather et Marron (1990) ont étudié l'utilisation de ce type d'estimation dans un contexte d'interpolation. Moon et Lall (1994) ont quant à eux, développé une méthodologie applicable pour l'extrapolation. L'estimation de la fonction noyau des quantiles (*kernel quantile estimator*) est en fait une sorte de lissage de la fonction quantile empirique. La fonction quantile empirique est définie à partir des données de l'échantillon. Il existe un certain nombre de formules de probabilités empiriques que l'on peut utiliser pour représenter la fonction quantile empirique. Moon et Lall (1994) utilisent la formule de probabilité empirique d'Adamowski, qui sera présentée au chapitre suivant. L'estimation de la fonction noyau des quantiles proposée par Moon et Lall (1994) a été inspirée de l'estimateur non paramétrique pour la régression présenté par Gasser et Müller (1984). La fonction quantile est donc la suivante :

$$\hat{x}(p) = \sum_{i=1}^{n} \frac{1}{h} y_i \int_{s_{i-1}}^{s_i} K\left(\frac{p-u}{h}\right) du \qquad (3.41)$$

où $s_i = (p_i + p_{i+1})/2$; $s_0 = 0$; $s_n = 1$, y_i sont les observations de l'échantillon et où les p_i proviennent d'une formule de probabilité empirique. Cette méthode d'estimation est utilisée dans le cadre de la méthode *plug-in* de Gasser *et al.* (1991) qui sera présentée au chapitre suivant. La procédure itérative décrite précédemment (3.38) a été utilisée pour estimer les quantiles pour toutes les autres méthodes qui seront présentées au chapitre 4.

4. CALCUL DU PARAMÈTRE DE LISSAGE

Il apparaît maintenant évident que le choix du paramètre de lissage est crucial lors de l'estimation avec la méthode des noyaux. Hormis le choix du noyau, le choix de h gouverne seul l'estimation. Un mauvais choix peut induire un sous-lissage ou un sur-lissage conduisant à une sous-estimation ou à une surestimation. Dans certains cas, la fonction est très sensible au paramètre de lissage, il en résulte qu'une faible variation de h peut conduire à une grande erreur.

Il est important de bien identifier *a priori* la raison pour laquelle on effectue un lissage. Si l'estimation non paramétrique est effectuée pour une analyse exploratoire, par exemple pour avoir une idée de la forme de la distribution empirique d'un échantillon, il n'est pas vraiment nécessaire d'appliquer une des méthodes qui seront présentées dans ce chapitre. Il suffit alors d'utiliser une méthode subjective, basée sur une distribution de référence, laquelle sera présentée à la section suivante. Au lieu de calculer un paramètre de lissage « optimal », on peut faire varier la valeur de ce dernier afin de déterminer le degré de lissage désiré. Ces recommandations ne sont valables que dans le cas où l'on désire faire l'analyse exploratoire d'un échantillon. Lorsque l'on considère la méthode des noyaux dans un contexte d'estimation, que ce soit pour l'interpolation que pour l'extrapolation, il est important d'appliquer une des méthodes présentées dans ce chapitre, ou toute autre méthode ayant fait ses preuves.

Il semble que le choix approprié du paramètre de lissage soit relié au type d'estimation que l'on désire effectuer. Que l'on soit dans un contexte d'interpolation ou d'extrapolation, on est porté à croire que le rôle du paramètre de lissage est différent. On tentera de trouver une

réponse à cette interrogation dans les études de comparaisons effectuées dans le cadre de cette étude.

Dans ce chapitre, une revue des méthodes de calcul du paramètre de lissage les plus utilisées sera présentée. Le chapitre est divisé en cinq parties. D'abord, on discutera de l'utilisation d'une distribution standard pour le calcul subjectif de h. Ensuite, on poursuit avec certaines considérations théoriques nécessaires pour la compréhension des sections 4.3 et 4.4. correspondant respectivement aux méthodes de calcul de h basées sur la fonction de densité et celles basées sur la fonction de répartition. La section 4.5 est consacrée à la présentation d'une méthode de calcul de h à partir de l'estimation des quantiles. Finalement, les méthodes d'optimisation pour la méthode à fenêtre variable feront l'objet de la dernière section du chapitre.

4.1 Utilisation d'une distribution standard

Lorsque le choix du paramètre de lissage n'est pas critique, par exemple dans les cas où l'on effectue l'analyse exploratoire d'un échantillon, on peut utiliser une distribution standard afin d'estimer les termes inconnus dans l'équation (3.27). On peut supposer que les données de l'échantillon proviennent d'une distribution normale de moyenne μ et de variance σ^2, ce qui permet de déterminer la valeur du terme $\int f''(x)^2\ dx$. Ensuite, afin de simplifier les calculs, on utilise un noyau normal pour l'estimation de f qui est supposée normale. On doit donc d'abord remplacer f'' par ϕ'', la dérivée seconde de la densité d'une distribution normale centrée réduite :

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}; \qquad \phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{4.1}$$

$$\phi''(x) = \frac{x^2 - 1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \qquad ; \phi''\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma^2\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \left[\left(\frac{x-\mu}{\sigma}\right)^2 - 1\right]$$

La moyenne n'est qu'un paramètre de position, tandis que l'écart-type joue un rôle dans la forme de la distribution. On désire donc que h soit relié à l'écart-type. Ainsi, on pose la moyenne égale à zéro. Par conséquent, en considérant une distribution de variance σ^2, on doit diviser le terme $\phi''(x/\sigma)$ par σ, puisque $f(x) = \sigma^{-1}\phi(x/\sigma)$. On a alors :

$$\int [f''(x)]^2 \, dx = \frac{1}{\sigma^2} \int \left[\phi''\left(\frac{x}{\sigma}\right)\right]^2 dx \tag{4.2}$$

En effectuant le changement de variable $z = x/\sigma$ et en sortant de l'intégrale le terme σ compris dans le terme ϕ'', on obtient :

$$\int [f''(x)]^2 \, dx = \frac{1}{\sigma} \int [\phi''(z)]^2 \, dz \tag{4.3}$$

En développant le terme à la puissance deux, on peut manipuler chacun des termes afin de trouver des formulations particulières :

$$\int [f''(x)]^2 \, dx = \frac{1}{\sigma^5} \int \frac{(z^2-1)^2}{\sqrt{2\pi}} e^{-z^2} dz \tag{4.4}$$

$$= \frac{1}{\sigma^5} \int \frac{z^4}{\sqrt{2\pi}} e^{-z^2} dz - \frac{2}{\sigma^5} \int \frac{z^2}{\sqrt{2\pi}} e^{-z^2} dz + \frac{1}{\sigma^5} \int \frac{1}{\sqrt{2\pi}} e^{-z^2} dz$$

$$= \frac{1}{\sigma^5\sqrt{2\pi}} \left[\int z^4 \phi(z\sqrt{2}) dz - \int 2z^2 \phi(z\sqrt{2}) dz + \int \phi(z\sqrt{2}) dz \right]$$

En effectuant le changement de variable $u = z\sqrt{2}$, on obtient :

$$\int\left[f''(x)\right]^2\,dx = \frac{1}{\sigma^5\sqrt{\pi}}\left[\int\frac{u^4}{8}\phi(u)\,du - \int\frac{u^2}{2}\phi(u)\,du + \int\frac{1}{2}\phi(u)\,du\right] \quad (4.5)$$

En considérant que $\text{var}[Y] = \int y^2\,\phi(y)\,dy = 1$ pour une loi normale centrée réduite, et en considérant que l'intégrale de la densité ϕ sur tout son domaine est égale à 1, les deux derniers termes de (4.5) s'annulent et on a finalement :

$$\int\left[f''(x)\right]^2\,dx = \frac{3}{8\sigma^5\sqrt{\pi}} \quad (4.6)$$

puisque le 4^e moment non-centré $\mu'_4 = \int u^4\phi(u)\,du$ est égal à 3 pour la loi normale. En remplaçant $\int\left[f''(x)\right]^2\,dx$ donné par l'expression (4.6) dans l'équation (3.27), le h optimal théorique qui minimise l'*IMSE*, en considérant le noyau K comme étant le noyau normal de variance unitaire, est le suivant :

$$h_{opt,IMSE} = \left\{\frac{1}{2\sqrt{\pi}}\right\}^{1/5}\left\{n\frac{3}{8\sigma^5\sqrt{\pi}}\right\}^{-1/5} = \sigma\left(\frac{4}{3n}\right)^{1/5} \quad (4.7)$$

Il est donc possible d'obtenir une estimation du paramètre de lissage optimal en ne considérant que l'écart-type et la taille de l'échantillon. Cette expression est efficace lorsque l'échantillon provient réellement d'une distribution normale, mais elle risque de causer un sur-lissage dans d'autres situations, par exemple si la population est multimodale. De plus, Silverman (1986), a remarqué que plus la population est asymétrique, plus l'expression (4.7) donne un paramètre de lissage élevé. On risque donc de sur-lisser lorsque la population est asymétrique.

Il est possible d'obtenir de meilleurs résultats en utilisant une mesure de l'étendue de l'échantillon plus robuste que l'écart-type. L'écart inter-quartile R permet, entre autres, d'être moins sensible aux valeurs singulières. Si on modifie l'équation (4.7) en utilisant plutôt R que σ et en utilisant le fait que $R = 1.34\sigma$ dans le cas d'une distribution normale, on obtient :

$$h_{opt,IMSE} = \frac{0.79\ R}{n^{1/5}} \quad \text{où} \quad R = X_{0,75} - X_{0,25} \tag{4.8}$$

où $X_{0,25}$ et $X_{0,75}$ représentent la valeur de l'échantillon correspondant au 25^{e} et 75^{e} percentile, respectivement. Pour les distributions à extrémités lourdes et pour les distributions asymétriques, l'expression (4.8) permet de réduire l'importance du sur-lissage par rapport à l'équation (4.7). Par contre, dans le cas de distributions multimodales, l'utilisation de l'écart inter-quartile a pour conséquence d'augmenter la valeur du paramètre de lissage par rapport à l'utilisation de l'écart-type. Il a finalement été convenu d'utiliser les expressions (4.7) et (4.8) conjointement, afin de combiner les avantages de chacune d'elles :

$$h_{opt,IMSE} = \frac{1.06}{n^{1/5}} \min\left(\sigma, \frac{R}{1.34}\right) \tag{4.9}$$

Cette expression performe relativement bien dans le cas où la distribution de la population est unimodale et symétrique, mais elle est moins efficace dans le cas des distributions multimodales et asymétriques.

Cette façon de calculer le paramètre de lissage peut être vraiment efficace dans certaines situations particulières, c'est-à-dire pour un certain nombre de distributions empiriques. De plus, elle est simple d'utilisation et rapide. Mais, comme dans la nature on est souvent confronté à des populations qui proviennent d'un mélange de plusieurs distributions, l'équation (4.9) n'est pas toujours appropriée. Elle peut toutefois servir à calculer une valeur initiale pour les autres méthodes itératives d'optimisation de h plus efficaces et plus complexes.

4.2 Considérations théoriques

Avant de se lancer dans l'étude des différentes méthodes de calcul du paramètre de lissage, il est important d'établir une distinction entre les méthodes basées sur la fonction de densité

et les méthodes basées sur la fonction de répartition. On a déterminé, au chapitre précédant, l'expression de l'erreur quadratique moyenne, c'est-à-dire l'équation (3.31) :

$$MSE_f(x) \approx \frac{1}{nh} f(x) \int K(t)^2 dt + \frac{h^4}{4} \left\{ f''(x) \int t^2 K(t) dt \right\}^2$$

Cette expression, lorsque elle est minimisée permet de déterminer la valeur optimale du paramètre de lissage h. L'équation (3.31) représente l'erreur quadratique moyenne de la **fonction de densité**, puisqu'elle est calculée à partir des différences entre la fonction de densité théorique et son estimation (équation (3.12) avec $IMSE = \int MSE\, dx$):

$$MSE_f(x) = E\left[\left(\hat{f}(x) - f(x)\right)^2\right]$$

Par conséquent, le paramètre de lissage que l'on déduit de l'optimisation de cette expression, est adéquat dans le cadre de l'estimation non paramétrique d'une densité. Mais il n'est pas certain qu'il soit convenable d'utiliser ce paramètre dans le cadre de l'estimation d'une fonction de répartition. Généralement, les paramètres de lissage calculés à partir de critères concernant la fonction de densité, sont tout de même utilisés pour l'estimation de la fonction de répartition et vice-versa, en tenant compte du fait que la fonction de répartition est obtenue par intégration de la fonction de densité. Mais par contre les propriétés diffèrent d'un cas à l'autre. Lall *et al.* (1993) soutiennent qu'il y a une différence fondamentale entre ces deux types d'estimation et qu'il faut les considérer séparément, contrairement à ce qui avait été observé par Jones (1990). C'est pourquoi il serait intéressant de dériver une expression de MSE pour la fonction de répartition et de la comparer à celle obtenue pour la fonction de densité. On calcule l'erreur quadratique moyenne de la fonction de répartition de la façon suivante :

$$MSE_F(x) = E\left[\left(\hat{F}(x) - F(x)\right)^2\right] \qquad (4.10)$$

Azzalini (1981) a démontré que l'erreur quadratique moyenne de la **fonction de répartition** est donnée par :

$$MSE_F(x) \approx \frac{1}{n}\{F(x)(1-F(x)) - uh\} + vh^4 \qquad (4.11)$$

où

$$u = f(x)\left\{ a - \int_{-a}^{a} K_I(t)^2 \, dt \right\}; \qquad v = \left\{ \frac{1}{2} f'(x) \int_{-a}^{a} t^2 K(t) \, dt \right\}^2$$

où $[-a, a]$ représente le domaine de variation du noyau K et où K_I représente le noyau K intégré. Si on compare l'expression (4.11) à l'équation (3.31), on voit d'abord qu'il existe certaines similitudes. Le premier terme de chacune des expressions est similaire, ils sont tout les deux divisés par n, sont à la puissance 1 et ils dépendent tout les deux de $f(x)$. La principale différence est que le premier terme de $MSE_f(x)$ est fonction de $\int K(t)^2 \, dt$, plutôt que de $\int_{-a}^{a} K_I(t)^2 \, dt$ dans le cas du premier terme de $MSE_F(x)$. Par ailleurs, le deuxième terme de chacune des expressions est fonction de h^4 et de $\left[\int t^2 K(t) dt\right]^2$. Mais $MSE_f(x)$ dépend de $f''(x)$ et $MSE_F(x)$ dépend de $f'(x)$; ces différences peuvent s'expliquer par le fait que F est obtenu par l'intégration de f (propriété 2.3), mais elles laissent croire qu'il n'est peut-être pas adéquat d'utiliser le paramètre optimal calculé à partir de la fonction de densité dans l'estimation de la fonction de répartition. Il en serait évidemment de même pour la réciproque.

4.3 Méthodes basées sur la fonction de densité *f*

Dans cette section, deux méthodes d'estimation du paramètre de lissage sont présentées, la méthode des moindres carrés et la méthode du maximum de vraisemblance. Elles sont appliquées en estimant la fonction de densité à partir des observations de l'échantillon. Ces méthodes permettent de calculer le paramètre de lissage optimal directement à partir des observations. Ces deux méthodes sont appliquées en utilisant le concept de validation

croisée qui consiste à retirer une valeur de l'échantillon lors de l'estimation de la fonction à ce point. La méthode de validation croisée s'apparente à la méthode du Jackknife. La méthode du Jackknife est souvent utilisée pour l'estimation de la variance des paramètres d'une population tandis que la validation croisée est surtout utilisée pour l'estimation d'une fonction d'erreur (Efron et Gong ; 1983). Par ailleurs, la méthode biaisée avec validation croisée aurait aussi pu être considérée (Marron ; 1988, Park et Marron ; 1990).

4.3.1 Moindres carrés avec validation croisée

La méthode des moindres carrés avec validation croisée (*least squares cross-validation*) a été suggérée par Rudemo (1982) et Bowman (1984). Elle consiste à trouver la valeur de h qui minimise la fonction d'erreur quadratique intégrée (*ISE*) :

$$ISE_h = \int \left(\hat{f}(x) - f(x)\right)^2 dx \qquad (4.12a)$$

$$= \int \hat{f}(x)^2 \, dx \; - \; 2 \int \hat{f}(x) f(x) \, dx \; + \; \int f(x)^2 \, dx \qquad (4.12b)$$

En examinant cette expression on voit que le premier terme peut être calculé directement à partir de l'échantillon et que le dernier ne dépend pas de h. Le second terme quant à lui doit être estimé. On peut donc réécrire (4.12b) de la façon suivante :

$$ISE_h - \int f(x)^2 \, dx \; = \; \int \hat{f}(x)^2 \, dx - 2 \int \hat{f}(x) f(x) \, dx \qquad (4.13)$$

Le dernier terme de (4.13) peut être estimé en utilisant la méthode de validation croisée. On retire donc la i^e observation de l'échantillon afin d'estimer la fonction en x_i. Il s'avère utile d'introduire la fonction $\hat{f}_{-i}(x)$, l'estimation de la fonction densité construite à partir de tous les points sauf l'observation i :

$$\hat{f}_{-i}(x) = \frac{1}{(n-1)h} \sum_{\substack{j \neq i \\ j=1}}^{n} K\left(\frac{x - x_j}{h}\right) \tag{4.14}$$

On estime donc le dernier terme de (4.13) par la moyenne des estimations de la fonction de densité en retirant l'observation *i* pour chacune d'elles :

$$\int \hat{f}(x) f(x)\, dx = \frac{1}{n} \sum_{i=1}^{n} \hat{f}_{-i}(x_i) \tag{4.15}$$

En incorporant (4.15) dans l'expression (4.13), on obtient une estimation de l'erreur quadratique intégrée et en y additionnant $\int f(x)^2$, on obtient une estimation non-biaisée. Ceci s'illustre bien en considérant l'espérance de (4.15) :

$$E\left[\frac{1}{n} \sum_{i=1}^{n} f_{-i}(X_i)\right] = E\left[\hat{f}_{-n}(X_n)\right] \tag{4.16}$$

Comme l'espérance de $\hat{f}$ ne dépend que du noyau et du paramètre de lissage, mais est indépendante de la taille de l'échantillon (Silverman ; 1986) :

$$E\left[\hat{f}_{-n}(X_n)\right] = E\left[\int \hat{f}_{-n}(x) f(x)\, dx\right] = E\left[\int \hat{f}(x) f(x)\, dx\right] \tag{4.17}$$

On peut donc dire que minimiser l'erreur quadratique intégrée par rapport à *h* revient à minimiser l'expression suivante, en remplaçant le second terme de (4.12b) par (4.15) :

$$MCVC_h = \int \hat{f}(x)^2\, dx - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{-i}(x_i) \tag{4.18}$$

Une forme pratique pour le terme $\int \hat{f}(x)^2\, dx$ de l'équation (4.18), est obtenue en utilisant le concept de convolution. Le premier terme de l'expression (4.18) peut être développé de la manière suivante :

$$\int \hat{f}(x)^2\,dx = \frac{1}{(nh)^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int K\left(\frac{x-x_i}{h}\right)K\left(\frac{x-x_j}{h}\right)dx \qquad (4.19)$$

En effectuant le changement de variable $u = (x - x_i)/h$, on obtient :

$$\int \hat{f}(x)^2\,dx = \frac{1}{n^2 h}\sum_{i=1}^{n}\sum_{j=1}^{n}\int K(u)\,K\left(u + \frac{x_i - x_j}{h}\right)du \qquad (4.20)$$

Dans l'expression (4.20), l'intégrale représente la convolution du noyau K avec lui-même lorsqu'il est symétrique, que l'on désigne généralement par $K^{(2)}$:

$$\int \hat{f}(x)^2\,dx = \frac{1}{n^2 h}\sum_{i=1}^{n}\sum_{j=1}^{n}K^{(2)}\left(\frac{x_i - x_j}{h}\right) \qquad (4.21)$$

Finalement, la méthode des moindres carrés avec validation croisée consiste à trouver la valeur de h qui minimise l'estimation de l'erreur carrée intégrée suivante :

$$MCVC_h = \frac{1}{n^2 h}\sum_{i=1}^{n}\sum_{j=1}^{n}K^{(2)}\left(\frac{x_i - x_j}{h}\right) - \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\frac{1}{h}K\left(\frac{x_i - x_j}{h}\right) \qquad (4.22)$$

Le paramètre de lissage optimal obtenu par la méthode des moindres carrés avec validation croisée est le suivant :

$$\hat{h}_{MCVC} = \arg\min_{h}(MCVC_h) \qquad (4.23)$$

La fonction $\arg\min_{z}(g(z))$ désigne la valeur de z correspondant à la valeur minimale de la fonction g. Désignons par ISE_{opt}, l'erreur quadratique intégrée qui correspond au paramètre de lissage qui minimise $\int(\hat{f} - f)^2$, dans l'hypothèse où la fonction f est connue. C'est-à-dire que l'on considère le meilleur h possible au sens de l'*ISE*. On dénote ensuite par ISE_{MCVC} l'erreur quadratique intégrée calculée avec le paramètre de lissage obtenu en optimisant (4.22). Stone (1984) a démontré qu'à mesure que la taille de l'échantillon

augmente, l'expression $MCVC_h$ (4.22) représente une bonne estimation de l'erreur quadratique intégrée (4.12a) correspondant au meilleur choix de h:

$$\frac{ISE_{MCVC}}{ISE_{opt}} \to 1 \qquad \text{lorsque } n \to \infty \tag{4.24}$$

Silverman (1986) note que la méthode des moindres carrés avec validation croisée n'est pas très efficace lorsque l'échantillon est formé de données discrètes ou arrondies. Pour de faibles valeurs de h, la fonction $MCVC_h$ est très sensible au degré de discrétisation de l'échantillon. En fait, l'expression (4.22), peut conduire au résultat $h = 0$ dans ces cas. Silverman (1986) propose donc de chercher une valeur de h comprise dans l'intervalle $\left[0{,}25h_{opt,IMSE}\,,\,1{,}5h_{opt,IMSE}\right]$ où $h_{opt,IMSE}$ est le paramètre optimal donné par (4.9).

4.3.2 Maximum de vraisemblance avec validation croisée

La méthode du maximum de vraisemblance présentée à la section 2.3.1.1 peut être utilisée avec la méthode de validation croisée (*maximum likelihood cross-validation*) pour calculer le paramètre de lissage optimal. Tout d'abord, il importe de définir le test du rapport de vraisemblance qui permet de déterminer si la fonction de densité et son estimation non paramétrique sont semblables. Ce test permet de conclure que le paramètre de lissage est adéquat si la statistique $f(x)/\hat{f}(x)$ se rapproche de 1 et qu'il ne l'est pas lorsque cette statistique est voisine de 0. La méthode consiste donc, pour trouver la valeur optimale de h, à minimiser l'espérance du logarithme du rapport de vraisemblance. On a donc :

$$E\left[\log\left(\frac{f(x)}{\hat{f}(x)}\right)\right] \to 0 \qquad \text{lorsque } \hat{f} \to f \tag{4.25}$$

Par conséquent, le paramètre de lissage optimal est celui qui minimise ce que l'on nomme mesure d'information de *Kullback-Leibler*, $I_{KL}\left(f,\hat{f}\right)$:

$$I_{KL}\left(f,\hat{f}\right)=\int \log\left(\frac{f(x)}{\hat{f}(x)}\right) f(x)\,dx \tag{4.26}$$

Mais il est impossible d'utiliser directement cette méthode puisque la mesure d'information de *Kullback-Leibler* dépend de la fonction de densité théorique inconnue f. On peut utiliser le principe de la méthode du maximum de vraisemblance en effectuant de la validation croisée. La logique de cette méthode est la même que dans le cas de la méthode des moindres carrés avec validation croisée (*MVVC*).

On suppose que l'on dispose d'un échantillon de données indépendantes $\{Y_1,Y_2,...,Y_m\}$ supplémentaires en plus de l'échantillon $\{X_1,X_2,...,X_n\}$ à partir duquel on veut déterminer *h*. La fonction de vraisemblance de ces observations, $\prod_{i=1}^{m}\hat{f}\left(Y_i\right)$, lorsqu'elle est maximisée, permet de déterminer la valeur de h adéquate pour l'échantillon $\{X_1,X_2,...,X_n\}$. Mais comme on ne dispose généralement pas d'un échantillon supplémentaire, on doit appliquer la méthode de validation croisée. On effectue alors l'estimation de f à partir de l'échantillon $\{X_1,X_2,...,X_n\}$ mais en retirant la i^e observation (4.14). On utilise ces estimations pour calculer la fonction de vraisemblance de X. Comme on l'a mentionné dans la section 2.3.1.1, il est plus
facile d'utiliser la fonction log-vraisemblance :

$$MVVC_h=\frac{1}{n}\log\left[\prod_{i=1}^{n}\hat{f}_{-i}\left(x_i\right)\right] \tag{4.27}$$

En développant (4.27) en utilisant l'expression (4.14), on obtient l'estimation suivante de la fonction de vraisemblance :

$$MVVC_h=\frac{1}{n}\sum_{i=1}^{n}\log\left[\hat{f}_{-i}\left(x_i\right)\right] \tag{4.28a}$$

$$MVVC_h = \frac{1}{n}\sum_{i=1}^{n}\log\left[\frac{1}{(n-1)h}\sum_{\substack{j\neq i\\ j=1}}^{n}K\left(\frac{x_i - x_j}{h}\right)\right] \tag{4.28b}$$

$$MVVC_h = \frac{1}{n}\sum_{i=1}^{n}\log\left[\sum_{j\neq i}K\left(\frac{x_i - x_j}{h}\right)\right] + \log\left[\frac{1}{h(n-1)}\right] \tag{4.28c}$$

Le paramètre de lissage optimal selon la méthode du maximum de vraisemblance avec validation croisée est celui qui maximise la fonction *MVVC* :

$$\hat{h}_{MVVC} = \arg\max_h\left(MVVC_h\right) \tag{4.29}$$

Il est possible de démontrer que l'estimation de la fonction de densité calculée avec le paramètre de lissage obtenu à partir de l'équation (4.28), se rapproche de la véritable densité f au sens de l'information de *Kullback-Leibler*. En utilisant le fait que les X_i sont indépendantes et identiquement distribuées, on peut considérer que l'espérance de la moyenne des fonctions $\log\left[\hat{f}_{-i}(X_i)\right]$ est égale à l'espérance de chacune de ces fonctions :

$$E\left[\frac{1}{n}\sum_{i=1}^{n}\log\left[\hat{f}_{-i}(X_i)\right]\right] = E\left[\log\left[\hat{f}_{-i}(X_i)\right]\right] \tag{4.30}$$

En utilisant la définition de l'espérance mathématique donnée en (4.17) et en négligeant l'effet de la validation croisée on peut estimer à partir de (4.27) que :

$$E\left[MVVC_h\right] = E\left[\frac{1}{n}\sum_{i=1}^{n}\log\left[\hat{f}_{-i}(X_i)\right]\right] \approx E\left[\int\log\left[\hat{f}(x)\right]f(x)\,dx\right] \tag{4.31}$$

On peut montrer en utilisant la définition de I_{KL} donnée par (4.26) que (4.31) peut s'écrire de la manière suivante :

$$E\left[MVVC_h\right] \approx -E\left[I_{KL}\left(f,\hat{f}\right)\right] + \int\log\left[f(x)\right]f(x)\,dx \tag{4.32}$$

Ce résultat permet de conclure que l'expression $-MVVC_h$ est, à une constante près, un estimateur non-biaisé de l'erreur de *Kullback-Leibler*, puisque le dernier terme de (4.32) ne dépend pas de h.

Certains problèmes rencontrés avec l'utilisation de cette méthode ont été soulevés dans la littérature. D'abord, comme dans le cas de la méthode des moindres carrés, la méthode du maximum de vraisemblance avec validation croisée n'est pas très efficace pour les échantillons de données discrètes. La valeur de *MVVC* tend vers l'infini à mesure que le paramètre de lissage tend vers zéro. Parfois, il peut suffire d'avoir deux observations complètement identiques dans l'échantillon pour que la fonction *MVVC* conduise à une valeur nulle pour h. On a aussi remarqué que cette fonction est très sensible aux valeurs singulières. Dans le cas des noyaux à support fini, c'est-à-dire les noyaux qui ne sont définis que dans un certain domaine, si une certaine observation de l'échantillon, $X_{(a)}$, est située à une distance supérieure à h de toutes les autres observations, l'estimation de la fonction de densité à ce point $\hat{f}_{-a}(x_{(a)})$ est alors nulle. Par conséquent la fonction log-vraisemblance tend vers $-\infty$, quelle que soit la valeur de h. Lors de l'optimisation de la fonction *MVVC*, la valeur de h tend à être élevée afin d'éviter ce problème. Ce qui peut conduire à un problème de sur-lissage, l'estimation en un point étant influencée par un trop grand nombre d'observations.

4.4 Méthodes basées sur la fonction de répartition

Les méthodes d'estimation du paramètre de lissage basées sur la fonction de répartition sont beaucoup moins nombreuses et moins utilisées que celles basées sur la fonction de densité. Certains travaux effectués dans cette direction sont ceux de Adamowski (1985), Azzalini (1981), Lejeune et Sarda (1992) et Altman et Léger (1995). Dans cette section, deux méthodes d'estimation du paramètre de lissage à partir de la fonction de répartition sont présentées, le critère d'Adamowski et la méthode « plug-in » de Altman et Léger.

4.4.1 Méthode d'Adamowski

La méthode d'Adamowski (1985) s'appuie sur un concept relativement simple. Elle consiste à trouver le paramètre de lissage qui minimise la somme des carrés des différences entre l'estimation par la méthode des noyaux de la fonction de répartition et une estimation de la fonction de distribution empirique.

La méthode consiste donc à ranger les observations en ordre croissant, à calculer ensuite la valeur de la probabilité empirique p_j pour chacune des observations et de minimiser l'erreur quadratique entre l'estimation non paramétrique et les p_j :

$$AC_h = \sum_{j=1}^{n} \left[\hat{F}(x_j) - p_j \right]^2 \tag{4.33}$$

où $\hat{F}(x_j)$ est la fonction de répartition non paramétrique (3.37). Le paramètre de lissage optimal au sens de la méthode d'Adamowski est le suivant :

$$\hat{h}_{AC} = \arg\min_h \left(AC_h \right) \tag{4.34}$$

Comme il existe une grande variété de formules de probabilité empirique au non-dépassement, la question est de savoir quelle formule utiliser. Adamowski (1985) a utilisé une formule de probabilité qu'il avait lui-même dérivé (Adamowski ; 1981). Cependant, il est possible d'utiliser une autre formule. Le choix de la formule de probabilité empirique sera discuté dans la section suivante.

À partir de l'expression (4.34), certaines modifications ont été apportées par d'autres chercheurs afin de tenter d'augmenter l'efficacité de la méthode (Lall *et al.* ; 1993). D'abord, le concept de validation croisée à été ajouté, en supprimant x_j de l'échantillon lors du calcul de l'estimation $\hat{F}(x_j)$. Ensuite, on a tenté d'améliorer l'estimation pour les

quantiles d'ordres supérieurs, en minimisant (4.34) seulement sur la partie supérieure de l'échantillon plutôt que sur l'ensemble de l'échantillon. On a donc déterminé un ordre r, variant de $0{,}05n$ à n (l'ordre n représentant l'échantillon complet), à partir duquel uniquement les statistiques d'ordre supérieures étaient considérées dans l'optimisation de l'expression (4.34). Ces deux modifications à la méthode d'Adamowski n'ont apparemment pas amélioré l'efficacité de la méthode, les meilleurs résultats étant obtenus avec l'échantillon complet (Lall *et al.* ; 1993).

4.4.1.1 Formules de probabilité empirique au non-dépassement

Les formules de probabilités empiriques (*plotting position formulae*) sont souvent utilisées pour visualiser la distribution des observations d'un échantillon. Elles permettent aussi de détecter la présence de valeurs singulières. Il existe une grande quantité de formules de probabilité empirique qui peuvent s'exprimer sous la forme générale suivante :

$$p_j = \frac{j-\alpha}{n+1-2\alpha} \qquad \text{avec} \quad 0 \le \alpha \le 1 \tag{4.35}$$

où j est la j^e valeur de l'échantillon rangé en ordre croissant et α est une constante qui dépend du type de distribution et p_j correspond à une fonction de probabilité au non-dépassement. La formule proposée par Adamowski est obtenue pour $\alpha = 0{,}25$ et s'exprime de la façon suivante :

$$p_j = \frac{j-0{,}25}{n+0{,}5} \tag{4.36}$$

Une comparaison des formules de probabilité empirique a été effectuée par Adamowski (1981). Il utilise des données simulées à partir des distributions Gumbel et Pearson type 3 et il compare ainsi les résultats obtenus avec les formules de probabilité empirique aux fonctions de répartition théoriques des données. Il désirait tirer de ces comparaisons, des conclusions quant à la capacité de chacune des formules à estimer la fonction de répartition

de données provenant de populations distribuées selon une loi Gumbel ou Pearson type 3. Il a comparé l'efficacité de 11 formules de probabilité empirique en plus de celle qu'il a développé. Il arrive à la conclusion que la formule de Weibull ($\alpha = 0$), celle dont l'utilisation est généralement conseillée, est relativement inefficace pour l'analyse de fréquence de crue lorsque les données sont distribuées selon une distribution à deux paramètres. Par ailleurs, pour les faibles probabilités au non-dépassement, la formule de Gringorten ($\alpha = 0{,}44$) est celle qui est la plus efficace tandis que celle d'Adamowski ($\alpha = 0{,}25$) est celle qui conduit à la meilleure précision pour les valeurs élevées. Pour ce qui est de la distribution à trois paramètres (P3), la formule d'Adamowski donne des résultats comparables à celle de Cunnane ($\alpha = 0{,}4$) et à celle de Chegodajew ($\alpha = 0{,}3$), qui sont généralement suggérées pour la P3.

Mais dans le contexte où la formule de probabilité empirique est utilisée dans la présente étude, il serait intéressant d'étudier la performance de chacune de ces fonctions, de façon à pouvoir en recommander une ou pour évaluer l'importance du choix de la formule à utiliser. Moon et Lall (1994) ont effectué une analyse de sensibilité au choix de la formule de probabilité empirique en comparant les formules de Weibull, d'Adamowski d'Hazen, correspondant respectivement à $\alpha = 0$, 0,25, 0,5. Ils arrivent à la conclusion que le choix de la formule n'a pas une grande influence sur l'estimation. Ils ont aussi remarqué que la variation induite par l'usage des différentes formules demeure relativement plus faible que celle introduite lors d'une estimation paramétrique. Dans la présente étude, en considérant que le choix de la formule de probabilité empirique n'est pas vraiment critique, aucune analyse de ce genre n'a été effectuée.

4.4.1.2 Remarques sur la méthode d'Adamowski

Certains auteurs ont remarqué que cette méthode conduit souvent à de faibles valeurs de *h*. Ce phénomène a aussi été observé avec les échantillons que nous avons considérés. La figure 4.1 est un exemple où le paramètre de lissage qui minimise le critère d'Adamowski a

une valeur qui se rapproche de zéro. Un échantillon de taille 50, simulé à partir d'une distribution LP3, a été utilisé pour cet exemple, les caractéristiques de cet échantillon sont présentées au tableau 4.1.

On a calculé la valeur de la fonction (4.34) pour différentes valeurs de h variant de 0.01 à 100, en utilisant un noyau Epanechnikov pour l'estimation non paramétrique et la formule de probabilité empirique d'Adamowski. On remarque sur l'agrandissement de la figure 4.1 que le minimum de la fonction d'Adamowski est atteint pour toutes les valeurs de h inférieures à 0.1, c'est-à-dire à partir de la valeur correspondant au plus petit des écarts de toutes les observations prises deux-à-deux (d_m du tableau 4.1). En fait, le minimum de la fonction est atteint lorsque le paramètre de lissage est si faible que l'estimation non paramétrique en un point n'est déterminée que par l'observation elle-même. Il n'y a donc en pratique aucun lissage.

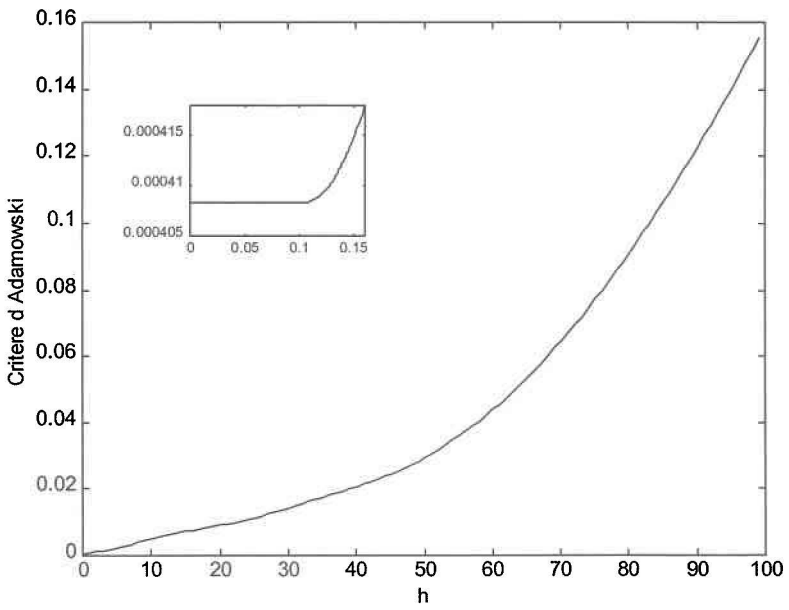


Figure 4.1 : Méthode d'Adamowski pour un échantillon simulé, en utilisant la formule de probabilité empirique d'Adamowski et un noyau d'Epanechnikov.

Tableau 4.1. Caractéristiques de l'échantillon simulé à partir de la loi LP3 considérée pour l'exemple d'utilisation de la méthode d'Adamowski.

Taille n	Moyenne $\bar{x}$	Écart-type s	Minimum $x_{(1)}$	Maximum $x_{(n)}$	Écart min d_m
50	205.5	68.5	117.0	507.2	0.093

Il est facile de montrer que lorsque $h < d_m$, la fonction de répartition non paramétrique $\hat{F}$ (3.37) est tout simplement une formule de probabilité empirique. Par exemple, pour un noyau Epanechnikov, on a la fonction de répartition non paramétrique suivante :

$$\hat{F}(x_j) = \frac{1}{n}\sum_{i=1}^{n}\left\{K_I^E\left(\frac{x_j - x_i}{h}\right) I_1\left(\left|\frac{x_j - x_i}{h}\right| \le 1\right) + I_2\left(\frac{x_j - x_i}{h} > 1\right)\right\} \qquad (4.37)$$

où K_I^E est le noyau Epanechnikov intégré (3.35), $I_1(A)$ et $I_2(B)$ sont des variables dichotomiques prenant respectivement les valeurs 1 et 0 selon que les énoncés A et B sont respectés ou non. En remplaçant $(x_j - x_i)/h$ par t pour alléger la notation, on a :

$$\hat{F}(x_j) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{3}{4}\left(\frac{2}{3} + t - t^3\right) I_1(|t| \le 1) + I_2(t > 1)\right\} \qquad (4.38)$$

Lorsque $h < d_m$, seule l'observation x_j elle-même se trouve dans l'intervalle $[x_j - h;\, x_j + h]$. Par conséquent, la variable I_1 ne vaut 1 que pour $i = j$. La variable I_2 quant à elle, prend la valeur 1 pour toutes les observations inférieures à x_j. On peut donc simplifier l'expression (4.38) de la façon suivante :

$$\hat{F}_{h<d_m}(x_j) = \frac{1}{n}\left[\frac{3}{4}\left(\frac{2}{3}\right) + (j-1)\right] \qquad (4.39)$$

Finalement, on obtient une expression indépendante du paramètre de lissage h :

$$\hat{F}_{h<d_m}(x_j) = \frac{j - 0.5}{n} \qquad (4.40)$$

Lorsque h est très faible, la fonction de répartition non paramétrique est donc une formule de probabilité empirique de Hazen correspondant à $\alpha = 0.5$ dans l'expression (4.35). Ce résultat s'applique dans le cas du noyau Epanechnikov, mais il est aussi valable dans le cas des autres noyaux présentés au tableau 3.1. La démonstration pour ces noyaux est présentée en annexe A.

Si on utilise la formule de probabilité empirique d'Adamowski, on peut montrer à partir de (4.36) et (4.40) que le critère d'Adamowski (4.33) prend la forme suivante lorsque h est très faible :

$$AC_{h<d_m} = \sum_{j=1}^{n}\left[\frac{j-0.5}{n} - \frac{j-0.25}{n+0.5}\right]^2 \qquad (4.41)$$

En réduisant les deux termes au même dénominateur et en développant le carré, on obtient :

$$AC_{h<d_m} = \frac{1}{\left(n^2+0.5n\right)^2}\sum_{j=1}^{n}\left[\frac{1}{16}\left(n^2+2n+1\right)+\frac{1}{4}\left(j^2-j-nj\right)\right] \qquad (4.42)$$

Les termes indépendants de j peuvent être sortis de la somme :

$$AC_{h<d_m} = \frac{1}{\left(n^2+0.5n\right)^2}\left[\frac{n(n+1)^2}{16}+\frac{1}{4}\sum_{j=1}^{n}j^2-\frac{(n+1)}{4}\sum_{j=1}^{n}j\right] \qquad (4.43)$$

En tenant compte des relations $\sum_{j=1} j = n(n+1)/2$ et $\sum_{j=1} j^2 = n(n+1)(2n+1)/6$, on obtient une expression qui n'est fonction que de la taille de l'échantillon n :

$$AC_{h<d_m} = \frac{1}{\left(n^2+0.5n\right)^2}\left[\frac{n(n+1)^2}{16}+\frac{n(n+1)(2n+1)}{24}-\frac{n(n+1)^2}{8}\right] \qquad (4.44)$$

Finalement, pour $h < d_m$, le critère d'Adamowski, est indépendant de la valeur de h et du type de noyau utilisé pour l'estimation non paramétrique et prend la forme suivante :

$$AC_{h<d_m} = \frac{(n+1)(n-1)}{48n(n+0.5)^2} \tag{4.45}$$

Comme on l'a vu précédemment, l'optimisation de la fonction d'Adamowski conduit à un paramètre de lissage relativement faible et le minimum de la fonction AC_h prend la valeur donnée par l'expression (4.45). Comme il n'y a aucun lissage, il y a peu d'intérêt à utiliser cette méthode.

Pour remédier à ce problème, Sarda (1993) propose d'utiliser le concept de validation croisée. Une comparaison de la méthode d'Adamowski (*AC*) classique et de la méthode d'Adamowski avec validation croisée (*ACVC*) a été effectuée en optimisant les deux critères pour divers échantillons simulés à partir d'une distribution LP3. Les résultats n'ont pas permis de conclure que l'utilisation de la validation croisée permet d'améliorer significativement la méthode d'Adamowski. Par ailleurs, Altman et Léger (1995) ont démontré que les deux méthodes sont asymptotiquement équivalentes :

$$\sup_{h \in H_n} \left| \frac{AC_h - ACVC_h}{IMSE_F} \right| \to 0 \tag{4.46}$$

où $IMSE_{F(x)} = E\left[\int \left(\hat{F}(x) - F(x)\right)^2 dx\right]$, $H_n = \left[C_1 n^{-a}, C_2 n^{-b}\right]$, $\frac{1}{4} \le b \le a \le \frac{1}{2}$ et C_1, C_2 sont des constantes.

On arrive donc à la conclusion que la méthode d'Adamowski conduit généralement à un paramètre de lissage inférieur à la plus faible distance entre les observations prises deux-à-deux ($h < d_m$). Comme il n'y a en pratique aucun lissage et que l'utilisation de la validation croisée n'apporte pas d'amélioration, la méthode d'Adamowski n'a pas été considérée dans la présente étude. En ce qui concerne les méthodes basées sur la fonction de répartition, seule la méthode *plug-in* de Altman et Léger présentée à la section suivante a été utilisée dans l'étude de comparaison.

4.4.2 Méthode « *plug-in* » de Altman et Léger

Une méthode automatique suggérée par Altman et Léger (1995), permet d'estimer h en minimisant l'erreur quadratique moyenne intégrée de la fonction de répartition $IMSE_F$. La valeur théorique de ce paramètre s'exprime de la façon suivante (Altman et Léger ; 1995):

$$h_{opt,IMSE_F} = \left[\frac{2\int f(x)^2 dx \int xK(x)K_I(x)dx}{n\int f'(x)^2 f(x)dx \left[\int x^2 K(x)dx\right]^2} \right]^{1/3} \tag{4.47}$$

où K_I est l'intégration du noyau. La valeur minimale de l'erreur quadratique moyenne intégrée de la fonction de répartition qui est obtenue avec la valeur optimale de h (4.47) est la suivante (Sarda; 1993) :

$$IMSE_F(x) \approx \frac{1}{n}\int F(x)[1-F(x)]f(x)dx - \frac{2h}{n}\int f(x)^2 dx \int xK(x)K_I(x)dx$$

$$+\frac{h^4}{4}\int f'(x)^2 f(x)dx \left[\int x^2 K(x)dx\right]^2 + \frac{Ch^2}{n} \tag{4.48}$$

où C est une constante positive. Altman et Léger (1995) proposent donc une méthode qui permet d'estimer chacun des termes de l'expression (4.48) qui comprennent la fonction inconnue f ou sa dérivée. Dénotons par $A(F)$ et $B(F)$ les deux termes inconnus :

$$A(F) = \int f(x)^2 dx \quad \text{et} \quad B(F) = \int f'(x)^2 f(x)dx \tag{4.49}$$

On peut estimer le terme $A(F)$ de la même façon que dans le cas de la méthode des moindres carrés (4.21). Hall et Marron (1987) proposent plutôt d'utiliser le concept de validation croisée pour estimer $A(F)$:

$$\hat{A}(F) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\frac{1}{h_A}K_A^{(2)}\left(\frac{x_i - x_j}{h_A}\right) \tag{4.50}$$

où K_A et h_A sont le noyau et le paramètre de lissage associés à l'estimation de $A(F)$. Pour l'autre terme inconnu $B(F)$, Altman et Léger (1995) ont introduit l'estimateur suivant :

$$\hat{B}(F) = \frac{1}{n^3 h_B^4} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} K'_B\left(\frac{x_i - x_j}{h_B}\right) K'_B\left(\frac{x_i - x_k}{h_B}\right) \tag{4.51}$$

où K_B et h_B sont le noyau et le paramètre de lissage associés à l'estimation de $B(F)$ et où $K'_B(t)$ est la dérivée du noyau $K_B(t)$ par rapport à t. En introduisant les estimateurs (4.50) et (4.51) dans l'expression (4.47), on obtient l'estimation du paramètre $h_{opt,IMSE_F}$ suivante:

$$\hat{h}_{opt,IMSE_F} = \left[\frac{2\,\hat{A}(F)\int xK(x)K_I(x)dx}{n\,\hat{B}(F)\left[\int x^2 K(x)dx\right]^2}\right]^{1/3} \tag{4.52}$$

4.4.2.1 Estimation des paramètres h_A et h_B

L'estimation de $A(F)$ et de $B(F)$ requiert la connaissance des paramètres h_A et h_B associés aux noyaux K_A et K_B. Altman et Léger (1995) suggèrent d'utiliser le noyau d'Epanechnikov dans les deux cas. Il ont montré que l'utilisation de ce noyau permet d'augmenter le degré de convergence. Pour ce qui est du paramètre de lissage, ils considèrent le même paramètre pilote pour l'estimation de $A(F)$ et de $B(F)$, soit $\alpha = h_A = h_B = n^{-0.3}$. Toutefois, ce paramètre ne tient pas compte de la variabilité de l'échantillon, il est fonction de la taille *n*, mais il est le même pour tous les échantillons de taille *n*, sans tenir compte de l'échelle des données. L'étude de simulation de Altman et Léger (1995) à été effectuée à l'aide d'échantillons simulés à partir d'une distribution normale centrée réduite $N(0,1)$, ou bien à partir d'un mélange de distributions normales. Dans tous les cas, le paramètre pilote constituait une fenêtre raisonnable pour l'estimation de $A(F)$ et de $B(F)$. Mais dans le cas de données hydrologiques comme les débits, les observations de l'échantillon sont d'une toute autre échelle et le paramètre α est beaucoup trop faible. Il est donc important de trouver une paramètre pilote qui tienne compte de la

variabilité de l'échantillon. Hall et Marron (1987) ont précisé que le paramètre α devait satisfaire les critères suivants :

$$n\alpha \to \infty \quad \text{et} \quad n^{1/4}\alpha \to 0 \quad \text{lorsque } n \to \infty \tag{4.53}$$

En prenant $\alpha = n^p$, on peut montrer que pour satisfaire (4.53), p doit être compris entre -1 et -0.25. On doit donc choisir α de la manière suivante :

$$\alpha = n^p \quad \text{pour} -1 < p < -0.25 \tag{4.54}$$

Pour tenir compte de la variabilité de l'échantillon, nous avons décidé d'utiliser plutôt :

$$\alpha_v = \hat{\sigma} n^p \quad \text{pour} -1 < p < -0.25 \tag{4.55}$$

Comme on l'a mentionné précédemment, Altman et Léger (1995) ont considéré une valeur de p de -0.3 pour leurs travaux. Dans notre cas, nous avons d'abord étudié la sensibilité de l'estimation au paramètre p. 50 échantillons tirés à partir d'une loi log-Pearson type 3 ont été générés et on a estimé les quantiles de période de retour 10, 20, 50, 100, 200 et 1000 ans avec α_v utilisant un paramètre p variable. Les paramètres de la loi de génération seront présentés au chapitre suivant. L'erreur quadratique moyenne (EQM) a été calculée à partir des quantiles théoriques de la log-Pearson type 3 pour chaque valeur de p. Cette procédure a été effectuée pour des échantillons de taille n = 10, 20, 50 100. Les résultats pour les échantillons de taille 50 sont présentés à la figure 4.2. Les courbes représentent l'erreur quadratique moyenne selon la valeur du paramètre p. On a considéré log(EQM) dans le but de réduire l'échelle afin de faciliter l'interprétation des résultats. Pour des périodes de retour de 10, 20 et 50 ans, le minimum de la courbe n'est pas très bien défini contrairement aux périodes de retour 100, 200 et 1000 ans.

On pourrait considérer à peu près toutes les valeurs comprises entre -1 et -0.8 pour une période de retour de 10 ans. Toutefois, on a décidé de retenir la valeur p = -0.9 pour T = 10, comme on peut le constater au tableau 4.2 qui contient les valeurs de p qui ont été

considérées dans la présente étude. Ces valeurs ont été obtenues avec l'optimisation de l'estimation des quantiles pour 25 échantillons seulement. Il aurait peut-être été nécessaire de considérer un plus grand nombre de réplications.

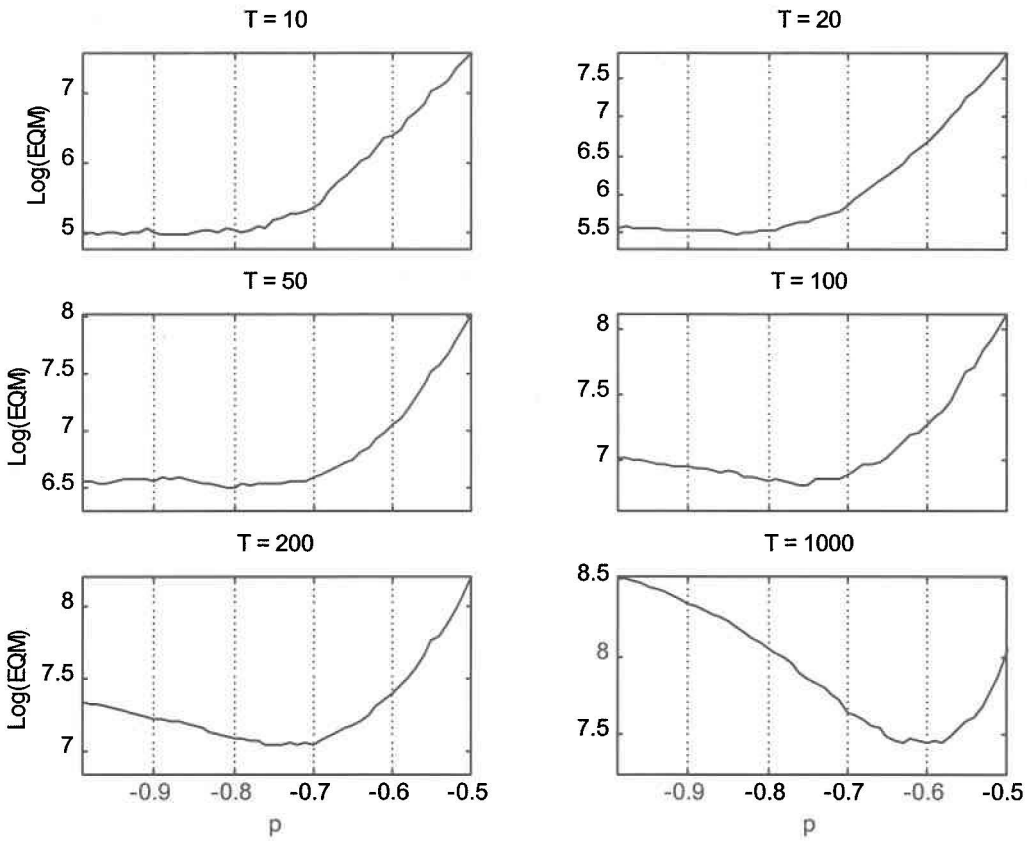


Figure 4.2 : Analyse de sensibilité du paramètre *p* pour l'estimation des quantiles de période de retour de 10, 20, 50, 100, 200 et 1000 ans. Résultats provenant de 50 échantillons de taille *n* = 50.

On remarque que la valeur de l'exposant p diminue avec la période de retour, ce qui implique que le paramètre pilote α_v augmente avec le degré d'extrapolation. Par conséquent, le fait de considérer le paramètre p variable en fonction de la période de retour T peut certainement favoriser l'extrapolation. Pour des périodes de retour qui ne figurent

pas dans le tableau 4.2, on peut effectuer une interpolation ou utiliser l'équation de régression suivante :

$$p = -1.0493 + 0.0652 \log(T) \tag{4.56}$$

Finalement, lorsque l'on utilise la méthode de Altman et Léger dans un contexte d'estimation de fréquence de crue, on peut considérer l'équation (4.55) pour calculer la valeur du paramètre pilote α servant à estimer le paramètre de lissage optimal à l'aide de l'expression (4.52) :

$$\alpha_v = \hat{\sigma} n^{-1.0493+0.0652 \log(T)} \tag{4.57}$$

Tableau 4.2 : Valeurs du paramètre *p* considérées selon la période de retour.

Période de retour	10 ans	20 ans	50 ans	100 ans	200 ans	1000 ans
Paramètre *p*	-0.9	-0.85	-0.8	-0.75	-0.7	-0.6

4.5 Méthode basée sur l'estimation des quantiles

On a vu à la section 4.2 qu'il existe des différences entre les méthodes d'estimation du paramètre de lissage basées sur l'erreur quadratique moyenne de la fonction de densité et celles basées sur la fonction de répartition. On a pourtant l'habitude de les utiliser dans toutes les situations, sans tenir compte du type d'estimation que l'on veut effectuer, même dans un contexte d'estimation de fréquence de crue. Comme en hydrologie statistique, on s'intéresse généralement à l'estimation des quantiles de période de retour *T*, la méthode décrite dans cette section présente un intérêt certain, puisqu'elle permet de calculer le paramètre de lissage optimal directement à partir de l'estimation des quantiles et est donc plus adaptée au problème rencontré en pratique.

4.5.1 Méthode *plug-in* (Gasser *et al.* ; 1991)

Cette méthode permet d'estimer itérativement la valeur du paramètre de lissage qui minimise l'erreur quadratique moyenne (*MSE*) des quantiles x_T de période de retour *T*. Il est important de comprendre que le paramètre de lissage optimal obtenu à l'aide de cette méthode n'est pas du même ordre de grandeur que les paramètres obtenus à l'aide des autres méthodes présentées aux sections 4.3 et 4.4, où le paramètre de lissage était utilisé pour les débits. Ces dernières permettaient d'estimer la fonction de répartition et seulement par la suite, on pouvait déduire la valeur du quantile correspondant en appliquant la relation (2.10). Mais dans le cas de la méthode *plug-in*, on considère plutôt le paramètre *h* comme fenêtre pour la fonction de répartition. Ainsi, les quantiles de période de retour *T* sont dérivés à partir de l'estimation de la fonction des quantiles empirique $\hat{x}(p)$:

$$\hat{x}(p) = x(p) + \varepsilon_i \tag{4.58}$$

Les résidus ε_i sont des variables aléatoires indépendantes. La fonction quantile non paramétrique telle que définie par Gasser et Müller (1984), peut s'exprimer comme étant la convolution de la fonction quantile empirique tout en considérant un noyau *K* :

$$\hat{x}(p) = \sum_{i=1}^{n} \frac{1}{h} y_i \int_{s_{i-1}}^{s_i} K\left(\frac{p-w}{h}\right) dw \tag{4.59}$$

où $s_i = \left(p_i + p_{i+1}\right)/2$ pour $i = \left\{1,\ldots,n-1\right\}$, $s_0 = 0$, $s_n = 1$, p est la probabilité correspondant à la valeur de la période de retour pour laquelle on veut estimer le quantile, telle que $p \in \left[0,1\right]$ et les y_i sont les observations de l'échantillon. Les p_i sont calculés à l'aide d'une fonction de probabilité empirique telle que présentée à la section 4.4.1.1 (équation 4.35).

La figure 4.3 illustre la façon dont on estime les quantiles à partir de l'équation (4.59). Un échantillon de taille 50 provenant d'une LP3 a été considéré pour cet exemple ainsi qu'un

noyau d'Epanechnikov pour l'estimation non paramétrique. Les paramètres de la loi de génération seront présentés au chapitre suivant. La figure de gauche représente le noyau impliqué dans l'estimation de la valeur de la fonction quantile pour la 25e observation de l'échantillon $\hat{x}(p_{25})$. La bande hachurée représente l'intégration du noyau pour la 22e observation. La méthode consiste à multiplier chacune des surfaces sous la courbe par la valeur de débit empirique correspondante. L'estimation de la fonction quantile pour p_{25} est obtenue avec le cumul de chacun de ces produits sur l'ensemble de l'échantillon. Dans cet exemple, seules les observations 18 à 32 sont impliquées dans le calcul de l'estimation pour p_{25}, la valeur du noyau étant nulle pour les autres. On peut noter que la fonction noyau illustrée sur cette figure n'est pas à l'échelle, elle a été multipliée par une constante afin d'améliorer la présentation. Le graphique de droite représente quant à lui, l'estimation de la fonction quantile pour des probabilités se rapprochant de 1, illustrant ainsi le problème causé par le fait que la distribution soit bornée et que le noyau soit incomplet. Ce problème est repris à la section suivante où le concept de noyau limite est introduit.

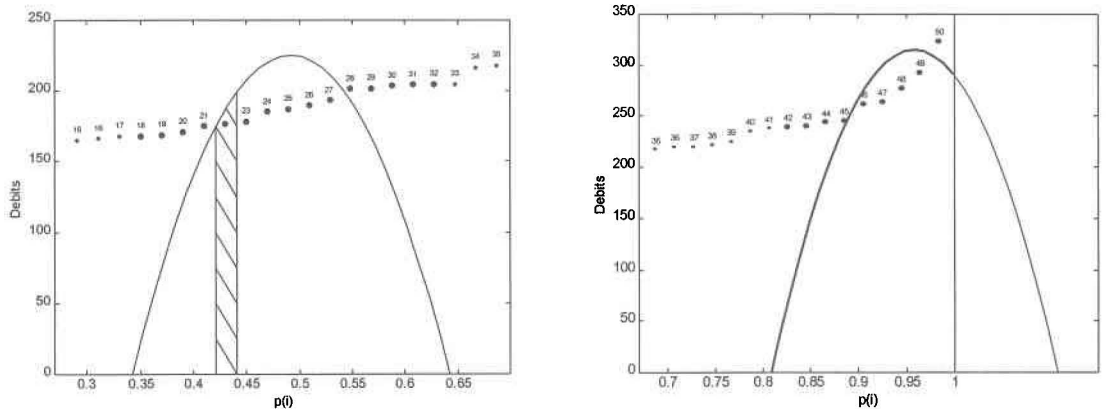


Figure 4.3 : Illustration de l'estimation d'un quantile à partir de l'expression non paramétrique.

La méthode de Gasser *et al.* permet de calculer une estimation de la valeur du paramètre de lissage qui minimise l'erreur quadratique moyenne de la fonction quantile (Müller ; 1991):

$$MSE_{x(p)} \approx \frac{\sigma^2}{nh}\int K(t)^2\,dt + \frac{h^4}{4}\left[x''(p)\right]^2\left\{\int K(t)t^2\,dt\right\}^2 \qquad (4.60)$$

où $\sigma^2 = \text{var}(\varepsilon_i)$ et $x''(p)$, la seconde dérivée de $x(p)$ doivent être estimés. Cette expression est très semblable à l'équation (3.31) où la fonction *MSE* avait été calculée à partir de la fonction de densité plutôt qu'à partir de la fonction quantile. Le paramètre de lissage qui minimise l'expression (4.60) est donné par Gasser et Müller (1984) :

$$h_{MSE_{x(p)}} = \left[\frac{1.5}{n}\frac{k_1}{k_2}\frac{\sigma^2}{\int_0^1\left[x''(p)\right]^2 dp}\right]^{0.2} \qquad (4.61)$$

Les termes k_1 et k_2 sont des constantes qui dépendent du noyau considéré. Le premier terme est deux fois l'intégrale sur tout le domaine de définition du carré de la fonction noyau, tandis que le second représente la variance multipliée par quatre :

$$k_1 = 2\int_{-a}^{a} K(t)^2\,dt \;; \qquad k_2 = 4\int_{-a}^{a} K(t)t^2\,dt \qquad (4.62)$$

Ensuite, on peut estimer $x''(p)$ par $\hat{r}_2(p;h_2)$ de la manière suivante :

$$\hat{r}_2(p;h_2) = \frac{1}{h_2^3}\sum_{i=1}^{n}\left[y_i\int_{s_{i-1}}^{s_i} D_x\left(\frac{p-u}{h_2}\right)du\right] \qquad (4.63)$$

où D_x est le noyau optimal du 4e ordre nécessaire pour estimer la deuxième dérivée de la fonction $x(p)$ (Müller ; 1991) qui sera présenté à la section 4.5.1.1 et h_2 représente le paramètre de lissage optimal correspondant. Gasser *et al.* (1991) ont montré que le fait de considérer $h_2 = hn^{1/10}$ mène à un taux de convergence de l'ordre de $n^{-1/2}$. L'estimateur pour la variance inconnue σ^2 qui a été considéré pour cette méthode est celui proposé par Gasser *et al.* (1986) :

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=2}^{n-1} c_i^2 \tilde{\varepsilon}_i^2 \tag{4.64}$$

Les pseudo-résidus $\tilde{\varepsilon}_i$ sont obtenus en considérant trois points consécutifs p_{i-1}, p_i, p_{i+1}. On trace une droite reliant les deux points les plus éloignés et on calcule ensuite la distance entre le point central et la droite de régression :

$$\tilde{\varepsilon}_i = a_i x(p_{i-1}) + b_i x(p_{i+1}) - x(p_i) \tag{4.65}$$

où $\quad a_i = \dfrac{p_{i+1} - p_i}{p_{i+1} - p_{i-1}}$; $\quad b_i = \dfrac{p_i - p_{i-1}}{p_{i+1} - p_{i-1}}$; $\quad c_i^2 = \dfrac{1}{a_i^2 + b_i^2 - 1}$

En incorporant les équations (4.62) à (4.65) dans l'expression (4.61), on obtient l'estimateur suivant :

$$\hat{h}_i = \left[\frac{1.5}{n} \frac{k_1}{k_2} \frac{\hat{\sigma}^2}{\int \hat{r}_2\left(p; \hat{h}_{i-1} n^{1/10}\right)^2 dp} \right]^{0.2} \tag{4.66}$$

Comme le terme $\hat{r}_2(p;h_2)$ dépend de la valeur $\hat{h}$, on doit estimer la paramètre de lissage optimal de façon récursive en considérant l'estimation $\hat{h}_{i-1}$ pour estimer le paramètre $\hat{h}_i$. Gasser *et al.* (1991) proposent de considérer la procédure suivante :

(a) On pose $h_1 = \dfrac{1}{n}$

(b) On calcule $\hat{h}_i$ à l'aide de l'équation (4.66) pour $i = 1,\ldots,11$

(c) On considère $\hat{h}_{plug-in} = \hat{h}_{11}$

La valeur optimale est obtenue lorsque l'on a estimé $\hat{h}_i$ 11 fois. Ce nombre d'itérations a été déterminé par une analyse des propriétés de convergence effectuée par Gasser *et al.*

(1991). Selon les résultats obtenus par Moon et Lall (1994), trois ou quatre itérations suffisent pour estimer h, la valeur de $\hat{h}_i$ ne variant que très peu à partir de quatre itérations.

Cet approche a été proposée dans le but d'améliorer l'estimation dans les extrémités des distributions et l'utilisation de noyaux d'ordre supérieurs pour accroître le taux de convergence de $MSE_{x(p)}$.

4.5.1.1 Utilisation de noyaux limites

Le graphique de droite de la figure 4.3 permet de visualiser le problème causé par le fait que la fonction de probabilité est bornée. La fonction à estimer étant discontinue à $p = 1$, les noyaux associés aux dernières observations sont incomplets. Sur la figure, on remarque qu'environ le tiers du domaine du noyau n'est associé à aucune observation et par conséquent, n'intervient pas dans l'estimation du quantile à p_{49}. La discontinuité de la fonction dans les extrémités a pour conséquence de causer un biais dans l'estimation non paramétrique des quantiles dans les queues. En considérant un noyau à support fini plutôt qu'un noyau asymptotique, on limite quelque peu l'effet de frontière.

Il serait préférable de considérer dans les extrémités un noyau différent de celui utilisé pour la partie centrale de la distribution. Müller (1991) propose un «noyau limite» (*boundary kernel*) qui correspond au noyau Epanechnikov utilisé pour estimer la partie centrale de la fonction. Ce noyau est construit de façon à réduire le biais pour l'estimation des extrémités. Le noyau Epanechnikov peut alors être utilisé pour l'interpolation, tandis que le noyau limite sert à extrapoler.

Müller (1991) considère le cas général où l'on veut estimer la v^e dérivée de la fonction d'intérêt. Considérons le noyau K comme étant μ fois différentiable, tel que $\mu \geq 0$. L'expression de la variance pour la $(v+\mu)^e$ dérivée de la fonction peut s'exprimer de la façon suivante :

$$\mathrm{var}\left(\hat{x}^{(\nu)},q\right) \approx \frac{\sigma^2}{nh^{2(\nu+\mu)+1}} \int_{-1}^{q} \left[K^{(\mu)}(q,u)\right]^2 du \qquad (4.67)$$

où q représente le support du noyau K. On considère un noyau différent selon que l'on se trouve dans l'extrémité de gauche ou de droite de la distribution. Pour l'extrémité de gauche, on pose $q = p/h$, tandis que pour l'extrémité de droite on prend $q = (1-p)/h$:

$K_+(p/h;t)$ est le noyau limite utilisé pour l'extrémité de gauche ; $p \in [0,h]$

$K_-((1-p)/h;t)$ est le noyau limite utilisé pour l'extrémité de droite ; $p \in [1-h,p]$

Le noyau optimal selon Müller (1991) est celui qui minimise la variance (4.67). En négligeant les constantes, le noyau optimal pour l'extrémité de gauche est celui qui minimise :

$$\int_{-1}^{q} \left[K_+^{(\mu)}(q,u)\right]^2 du \qquad (4.68)$$

avec les conditions suivantes :

(a) $K_+(q,\cdot):[-1,q] \to \Re$; $\quad K_+(\cdot,p):[0,1] \to \Re$

(b) $K_+^{(j)}(q,-1) = K_+^{(j)}(q,q) = 0 \qquad 0 \le j < \mu$

Le noyau optimal correspondant à l'extrémité droite peut être déterminé de façon analogue. Comme on cherche à estimer la fonction elle-même et non une de ses dérivées, on pose $\nu = 0$, et le minimum de (4.68) sujet aux contraintes (a) et (b) est le suivant :

$$K(q,t) = (1+t)^{\mu}(q-t)^{\mu}\left[g_0(q,t) + g_1(q,t)t\right] \qquad (4.69)$$

$$\text{où} \quad g_0 = \left(\frac{1}{1+q}\right)^{2\mu+1} \frac{(-1)^{\mu}(2\mu+1)!}{\mu!^2}\left[1+(2\mu+3)\left(\frac{1-q}{1+q}\right)^2\right]$$

$$g_1 = \left(\frac{1}{1+q}\right)^{2\mu+1} \frac{(-1)^{\mu}(2\mu+1)!}{\mu!^2}(4\mu+6)\frac{1-q}{(1+q)^2}$$

En posant $\mu = 1$ et $q = 1$ dans (4.69), le noyau $K(1, t)$ correspond au noyau Epanechnikov utilisé pour la partie centrale de la distribution. Pour $0 \le q \le 1$, on considère le noyau suivant, qui est le noyau optimal du 4e ordre nécessaire pour estimer la deuxième dérivée de la fonction $x(p)$:

$$K(q,t) = \begin{cases} 0.75(1-t^2) & p \in [h, 1-h];\ |t| < 1 \\ \dfrac{6(1+t)(q-t)}{(1+q)^3}\left[1+5\left(\dfrac{1-q}{1+q}\right)^2+10\dfrac{1-q}{(1+q)^2}t\right] & p \in [0,h];\ p \in [1-h,1];\ |t| < 1 \\ 0 & \forall p;\ |t| > 1 \end{cases} \tag{4.70}$$

Le noyau limite défini en (4.70) est illustré à la figure 4.4 en comparaison avec le noyau d'Epanechnikov standard. La comparaison des deux noyaux s'effectue dans l'extrémité de droite, mais un raisonnement analogue existe pour l'extrémité de gauche. En étant défini entre $p-h$ et 1, le noyau limite confère un poids supérieur aux dernières observations de l'échantillon et un poids moindre pour les observations situées près de $p-h$. De cette façon, il est plus efficace pour l'extrapolation qu'un autre noyau qui ne tient pas compte de l'effet de frontière. Pour un paramètre de lissage relativement grand, certaines observations ont un poids négatif, l'intégrale du noyau sur tout le domaine devant être égale à 1. À l'inverse, pour un paramètre de lissage se rapprochant de $1-p$, le noyau limite s'apparente au noyau d'Epanechnikov pour s'y confondre lorsque $h = 1-p$.

En utilisant ce noyau, on ne perd pas toute la partie qui sort du domaine de variation de la fonction à estimer, c'est à dire la partie du noyau d'Epanechnikov située à droite de $p = 1$. Le noyau a donc été modifié pour que son domaine de variation coïncide avec celui de la distribution à estimer, de façon à réduire le biais dans les extrémités.

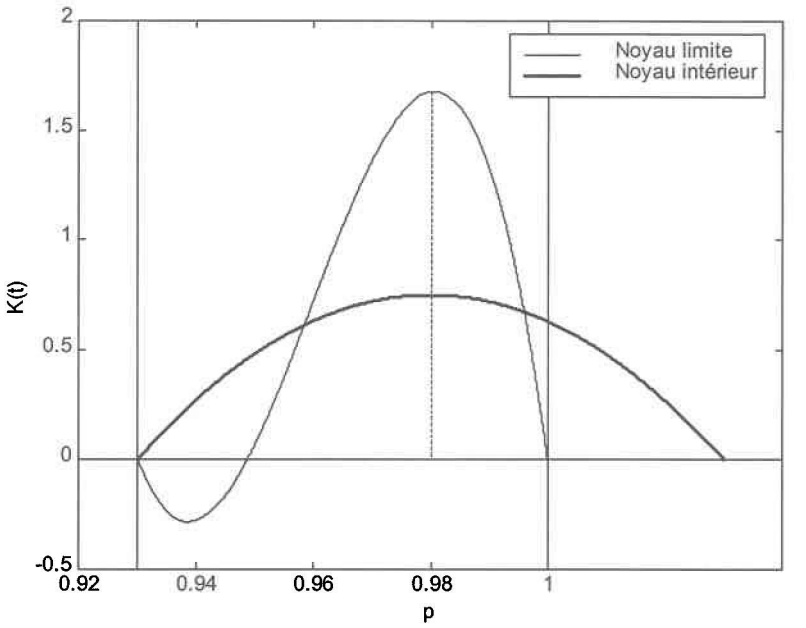


Figure 4.4 : Illustration du noyau limite et du noyau d'Epanechnikov standard dans les extrémités.

4.6 Méthodes à fenêtre variable

Deux méthodes permettant d'estimer la valeur des paramètres a_k et k de la méthode à fenêtre variable présentée à la section 3.3, sont décrites dans cette section. La première méthode consiste à déterminer les paramètres qui minimisent une certaine fonction qui reflète la qualité de l'adéquation, tandis que la seconde est la méthode du maximum de vraisemblance adaptée pour la méthode à fenêtre variable, qui a été proposée par Adamowski (1989).

4.6.1 Critère de la qualité de l'adéquation de Breiman *et al.* (1977)

Comme dans la pratique la distribution théorique n'est pas connue, il est impossible de se baser sur les calculs d'erreur pour identifier les paramètres optimaux k et a_k. Breiman *et al.* (1977) proposent donc plutôt une méthode qui s'appuie sur un critère d'adéquation (*goodness of fit*). Ces auteurs définissent les variables suivantes qui sont utilisables dans un contexte multivarié :

$$W_j = \exp\left\{-n f(x_j) V(d_{1j})\right\} \qquad j = 1,\ldots,n \tag{4.71}$$

où $f(x_j)$ représente la véritable fonction de distribution des données évaluée à x_j, où d_{1j} est la distance entre la j^e observation et son premier voisin le plus proche ($k = 1$) et où $V(d_{1j})$ représente le volume d'une sphère à M dimensions, de rayon d_{1j}, c'est-à-dire:

$$V\left(d_{1j}\right) = \frac{\pi^{M/2} d_{1j}^{M}}{\Gamma\left(M/2+1\right)} \tag{4.72}$$

Dans le cas univarié, on considère $M = 1$ et en estimant la fonction f avec la méthode à fenêtre variable (3.35), les statistiques $\hat{W}_j$ sont définies de la façon suivante :

$$\hat{W}_j = \exp\left\{-2n \hat{f}(x_j) d_{1j}\right\} \tag{4.73a}$$

$$\text{où} \quad \hat{f}\left(x_j\right) = \sum_{i=1}^{n} \frac{1}{n a_k d_{ki}} K\left(\frac{x_j - x_i}{a_k d_{ki}}\right) \tag{4.73b}$$

La valeur d_{ki} est la distance de l'observation x_i a son k^e voisin le plus proche. L'expression (4.73b) a été obtenue à partir de l'équation (3.33) pour $x = x_j$. On considère les statistiques $\hat{W}_j$ rangées en ordre croissant $\hat{W}_{(1)}, \hat{W}_{(2)}, \ldots, \hat{W}_{(n)}$ distribuées approximativement selon une loi uniforme (Breiman *et al.* ; 1977). La méthode de Breiman *et al.* (1977), s'appuyant sur un critère d'adéquation, consiste à minimiser le carré de la

somme des erreurs au carré, c'est-à-dire la différence entre la valeur empirique des variables W_j et leurs valeurs théoriques obtenues à partir de la distribution uniforme :

$$\hat{S} = \sum_{j=1}^{n} \left(\hat{W}_{(j)} - \frac{j}{n} \right)^2 \tag{4.74}$$

Les valeurs optimales de k et a_k sont obtenues en minimisant l'expression (4.74) :

$$\left[\hat{k}, \hat{a}_k \right]_{breiman} = \arg\min_{k, a_k} \left(\hat{S} \right) \tag{4.75}$$

Breiman *et al.* (1977) proposent une procédure en quatre étapes pour trouver le minimum de la fonction (4.74). Cette procédure n'est pas présentée en détail dans le présent travail, mais elle s'appuie sur le fait que la valeur optimale de k est obtenue pour :

$$\frac{a_k \bar{d}_k^2}{\sigma_{d_k}} \approx \text{constante} \tag{4.76}$$

où $\bar{d}_k$ et σ_{d_k} sont la moyenne et l'écart-type des distances de chacune des observations à leur k^e voisin le plus proche. La méthodologie qui a été préconisée pour cette étude comprend deux étapes. On trouve d'abord la valeur de a_k qui minimise (4.74) pour chacune des valeurs de k. On détermine ensuite le couple (k, a_k) pour lequel l'expression (4.75) est vérifiée. Cette façon de procéder est peut-être moins rapide que celle proposée par Breiman *et al.* (1977), mais elle permet de déterminer avec plus de précision le minimum de la fonction (4.74).

4.6.2 Fonction maximum de vraisemblance d'Adamowski (1989)

Une méthode basée sur l'optimisation de la fonction vraisemblance a été proposée par Adamowski (1989). Le choix du paramètre k, le rang du voisin le plus proche à considérer, est un peu subjectif dans cette méthode. Adamowski (1989) s'appuie sur une remarque faite par Breiman *et al.* (1977) pour le choix de la valeur k. Suite à la dérivation de la propriété

(4.76), Breiman *et al.* ont établi empiriquement que la valeur de k optimale correspondait à un coude sur la courbe de $\overline{d}_k$ en fonction de k. Un exemple typique de ce type de courbe est présenté au chapitre 5. En fait, on retient la valeur de k suivant le coude de la courbe, c'est à dire suivant un changement brusque de la courbe. Ensuite, on peut dériver la valeur de a_k qui maximise la fonction de vraisemblance :

$$L\left[a_k \middle| x_1, x_2, \ldots, x_n\right] = \sum_{i=1}^{n} \log\left[f\left(x_i \middle| a_k\right)\right] \tag{4.77}$$

où $f\left(x_i \middle| a_k\right)$ est obtenu à partir de (4.73b). Si on dérive cette fonction par rapport à a_k et qu'on l'égale à zéro, on obtient :

$$a_k + \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\sum_{j\neq i}^{n}\frac{x_i - x_j}{d_{kj}^2}K'\left(\frac{x_i - x_j}{a_k d_{kj}}\right)}{\sum_{j\neq i}^{n}\frac{1}{d_{kj}}K\left(\frac{x_i - x_j}{a_k d_{kj}}\right)}\right] = 0 \tag{4.78}$$

où K' est la dérivée du noyau K. La valeur optimale de k obtenue graphiquement et la valeur de a_k correspondante sont obtenues de la façon suivante :

$$\begin{aligned} &\hat{k} \mapsto \frac{a_k \overline{d}_k}{\sigma_{d_k}} \approx \text{constante} \\ &\hat{a}_k = \arg\max_{a_k}\left(L\left[a_k \middle| x_1, x_2, \ldots, x_n\right]\right) \end{aligned} \tag{4.79}$$

La procédure de calcul de la valeur optimale de k étant plutôt visuelle, il a fallu automatiser l'identification du coude et le choix du k correspondant afin d'éviter d'avoir à examiner la figure pour chacun des échantillons d'une étude de simulation. Une manière intuitive de procéder consiste à calculer la dérivée seconde et à déterminer le point où elle est maximale. Mais comme il est impossible d'évaluer analytiquement la seconde dérivée, puisqu'il n'existe pas de relation entre $\overline{d}_k$ et k, il faut donc s'en remettre à une procédure numérique. Comme les valeurs de k sont entières, on a un pas constant pour l'axe des x, on a donc $n-1$

segments de même largeur x. On calcule la valeur de la dérivée de chacun des segments en soustrayant les $\overline{d}_k$ du début et de la fin de chaque segment et en divisant par la largeur x (dans le cas présent 1). Le critère permettant d'établir la position du coude dans la courbe est donc la valeur maximale de tout les rapports de deux dérivées consécutives. Les rapports de dérivées représentent la seconde dérivée. On peut résumer la procédure de la façon suivante :

a) Pour $i = 1$ à n-1, calculer les dérivées:

$$A_i = \frac{\overline{d}_{i+1} - \overline{d}_i}{k_{i+1} - k_i} \tag{4.80}$$

b) On fait le rapport des dérivées consécutives :

$$R_i = \frac{A_{i+1}}{A_i} \tag{4.81}$$

c) On considère la valeur maximale des R_i comme estimation de k :

$$\hat{k} = \arg\max_k \left(R_i\right) \tag{4.82}$$

La figure 4.5 est un exemple typique de courbe de $\overline{d}_k$ en fonction de k. Cet exemple provient d'un échantillon de taille n = 50, il y a donc 49 valeurs possibles pour k. Mais seulement 14 valeurs de k sont illustrées sur la figure de façon à alléger la présentation. Le point encerclé sur la courbe correspond à la valeur du coude calculée à l'aide de la procédure précédante. On voit clairement que l'accroissement de la valeur de la distance moyenne est beaucoup plus élevé entre les voisins k_{i+1} et k_{i+2} que entre les voisins k_i et k_{i+1}, c'est à dire que :

$$\left[\overline{d}_{i+2} - \overline{d}_{i+1}\right] > \left[\overline{d}_{i+1} - \overline{d}_i\right] \tag{4.83}$$

Sur l'ensemble de la courbe, c'est à cet endroit que l'écart entre deux dérivées consécutives est le plus grand et à un pas constant pour k, la valeur k_{i+1} est identifiée comme étant la valeur optimale.

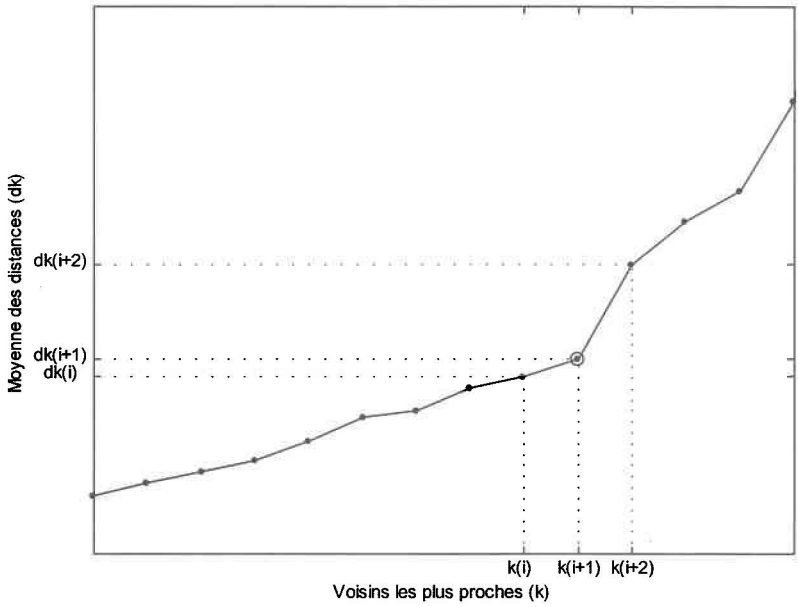


Figure 4.5 : Calcul de la valeur optimale de k par l'identification du coude à partir de la dérivée seconde.

4.6.3 Adaptation des méthodes à noyau fixe

Il est possible d'utiliser les méthodes de calcul du paramètre de lissage pour la méthode à noyau fixe présentées aux sections 4.3 et 4.4 dans un contexte d'estimation avec la méthode à fenêtre variable. L'optimisation des critères s'effectue alors sur deux variables corrélées, k et a_k au lieu de ne considérer que le paramètre h. Dans l'étude de simulation présentée au chapitre 5, la méthode des moindres carrés avec validation croisée pour noyau fixe a été considérée dans le cas à fenêtre variable. Dans ce cas, l'estimation de la fonction d'erreur *ISE* est la suivante :

$$MCVC_{a_k,k} = \frac{1}{n^2 a_k}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{d_{ki}}K^{(2)}\left(\frac{x_i - x_j}{a_k d_{ki}}\right) - \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n}\frac{1}{a_k d_{ki}}K\left(\frac{x_i - x_j}{a_k d_{ki}}\right) \tag{4.84}$$

Cette expression provient de l'équation (4.22) présentée à la section 4.3.1, en remplaçant h par $a_k d_{ki}$. Les paramètres optimaux $\hat{k}$ et $\hat{a}_k$ sont ceux qui minimisent l'estimation de l'erreur *ISE* énoncée à l'équation (4.84) :

$$\left[\hat{k}, \hat{a}_k\right]_{MCVC} = \arg\min_{k,a_k}\left(MCVC_{a_k,k}\right) \tag{4.85}$$

Cette façon de procéder diffère quelque peu de la méthodologie proposée par Fan *et al.* (1996) ou de celle de Staniswalis (1989) pour adapter la méthode des moindres carrés à une estimation locale du paramètre de lissage.

Une procédure adaptative proposée par Abramson (1982) et reprise par Silverman (1986) permet d'effectuer une estimation locale du paramètre de lissage en considérant le développement en série de Taylor du carré du biais de la fonction $\hat{f}(x)$. Brièvement, la méthode consiste à transformer le paramètre de lissage global λ_p en une séquence de paramètres locaux λ_i de la façon suivante :

$$\lambda_i = \lambda_p\left(\frac{\hat{f}_{\lambda_p}(x_i)}{g}\right)^{-1/2} \tag{4.86}$$

où $\hat{f}_{\lambda_p}$ représente l'estimation de la fonction de densité en utilisant le paramètre de lissage global et où g est la moyenne géométrique des $\hat{f}_{\lambda_p}$. Cette méthode n'a toutefois pas été considérée dans l'étude de simulation effectuée dans ce travail, mais elle a tout de même retenu l'attention pour la comparaison à la méthode des moindres carrés avec validation croisée à fenêtre variable à partir des conclusions tirées par Sharma *et al.* (1998). Cette comparaison ainsi que les remarques de Sharma *et al.* (1998) sont discutés à la section 5.5.6.

5. COMPARAISONS

Dans ce chapitre les résultats des différentes comparaisons des méthodes paramétriques et non paramétriques mentionnées au chapitre précédent sont présentés, ainsi que toutes les remarques constructives faites lors de l'étude de ces méthodes. Ce chapitre est divisé en sept sections traitant respectivement de :

- La méthodologie utilisée pour effectuer les comparaisons ;
- Une étude sur le nombre de réplications à considérer ;
- La comparaison de l'efficacité des différents noyaux ;
- Une étude de l'importance de la taille de l'échantillon ;
- La comparaison de toutes les méthodes de calcul du paramètre de lissage h ;
- La comparaison de la méthode à fenêtre fixe et de la méthode à fenêtre variable ;
- La comparaison de la méthodes des noyaux aux méthodes paramétriques.

5.1 Méthodologie

Afin de pouvoir tirer des conclusions générales, les comparaisons ont été effectuées par le biais d'une étude de simulation. Au lieu de considérer un échantillon de débits provenant d'une rivière en particulier, les données ont été obtenues par simulation d'un certain nombre d'échantillons à partir d'une distribution statistique. De cette façon, il est possible de connaître les valeurs théoriques des débits de période de retour T (à partir de la distribution parente) et ainsi pouvoir établir des critères de comparaisons adéquats, représentant l'efficacité des méthodes à estimer les valeurs théoriques. Les données ont été simulées à partir d'une distribution log-Pearson type 3 (section 2.3.2.1) à l'aide de la méthode de génération de Johnk (Johnk ; 1964, Devroye ; 1986). Les paramètres de la distribution ont été déterminés à partir de l'ajustement aux données de débits maximum annuels de la

rivière Harricana à Amos avec la méthode des moments présentée à la section 2.3.2.1. Le tableau 5.1 contient la valeur des paramètres utilisés pour la simulation ainsi que certaines caractéristiques de l'échantillon de la rivière Harricana. Le choix de la distribution parente a été relativement arbitraire. En fait, étant donné que la LP3 est souvent utilisée pour les débits, elle représente un choix raisonnable pour la simulation. On a aussi voulu créer des conditions favorables pour les méthodes paramétriques afin de pouvoir les comparer à la méthode des noyaux dans un contexte idéal. Pour ce qui est des paramètres, l'utilisation des données de la rivière Harricana pour l'ajustement a pour but d'obtenir des valeurs de débits simulés raisonnables.

Tableau 5.1 : Caractéristiques des données de la rivière Harricana (en m^3/s) et valeur des paramètres d'ajustement de la distribution log-Pearson type 3.

Rivière Harricana	
Nombre d'observations	69 années (1915-1983)
Moyenne de l'échantillon	191.32
Écart-type de l'échantillon	47.96
Coefficient d'asymétrie	0.86055
Coefficient de variation	0.25069
Paramètre α de la distribution LP3	119.72
Paramètre λ de la distribution LP3	846.88
Paramètre m de la distribution LP3	-1.85

On a donc simulé r échantillons de taille n fixe à partir de la LP3 avec les paramètres du tableau 5.1. Chacune des méthodes paramétriques et non paramétriques a été appliquée aux r échantillons et les statistiques suivantes ont été calculées pour chacunes des périodes de retour T d'intérêt, en tant que critères de comparaison :

$$\bar{\hat{x}}_{T_m} = \frac{1}{r}\sum_{i=1}^{r} \hat{x}_{i,T_m} \qquad m = 1,\ldots,6 \tag{5.1}$$

$$B_{T_m} = \frac{1}{r}\sum_{i=1}^{r}\left(\hat{x}_{i,T_m} - x_{T_m}\right) \qquad m = 1,\ldots,6 \tag{5.2}$$

$$RMSE_{T_m} = \sqrt{\frac{1}{r}\sum_{i=1}^{r}\left(\hat{x}_{i,T_m} - x_{T_m}\right)^2} \tag{5.3}$$

où $\hat{x}_{i,T_m}$ représente l'estimation du quantile de période de retour T_m pour l'échantillon i, x_{T_m} représente la valeur théorique du quantile de période de retour T_m déterminée à partir de la loi LP3 et $T_m = \{10,20,50,100,200,1000\}$ représente les périodes de retour considérées, en années (Tableau 5.2). Les équations (5.1) à (5.3) représentent respectivement la moyenne des estimations, le biais et la racine de l'erreur quadratique moyenne pour une période de retour T_m. Le biais sera plutôt présenté sous forme de pourcentage de valeur théorique :

$$\left(B_{T_m}\right)_r = \frac{B_{T_m}}{x_{T_m}} \times 100 \tag{5.4}$$

où $(B_{T_m})_r$ représente le biais relatif. Comme lors de la simulation on a le contrôle de la taille d'échantillon, l'étude par simulation présente l'avantage sur l'utilisation de données réelles. Comme on l'a vu lors de la définition des objectifs au chapitre 1, on cherche à évaluer la performance de ces méthodes selon la taille de l'échantillon utilisé. Il est donc relativement facile de faire varier la taille des échantillons pour les simulations et ainsi déterminer l'importance de n pour la méthode des noyaux. On a donc simulé r échantillons de taille n = 10, 25, 50 et 100, pour un total de $4r$ échantillons à utiliser avec les diverses méthodes présentées précédemment.

Tableau 5.2 : Périodes de retour considérées dans l'étude et probabilités au non dépassement correspondantes.

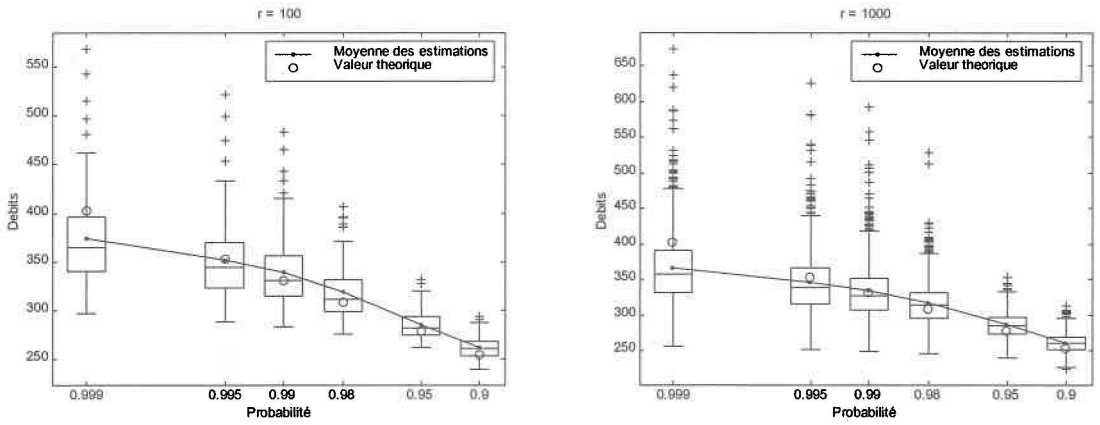
Période de retour (T)	10 ans	20 ans	50 ans	100 ans	200 ans	1000 ans
Probabilité (F)	0.9	0.95	0.98	0.99	0.995	0.999

5.2 Nombre de réplications

Dans un premier temps, il a fallu déterminer le nombre de réplications r à utiliser pour chaque taille d'échantillon. Dans la littérature, on peut fréquemment voir des études de simulation considérant 1000 réplications. Ce nombre est généralement raisonnable. Mais comme pour certaines méthodes le temps d'exécution des algorithmes menant au calcul du paramètre de lissage optimal est relativement élevé, il est nécessaire de limiter le nombre de réplications, sans toutefois risquer de perdre de la précision. Dans leur étude de simulation, Lall *et al.* (1993) considèrent 100 réplications, puisqu'ils ont observé que l'utilisation de 1000 réplications menait à des résultats semblables à ceux obtenus avec 100 échantillons. Avant de décider de considérer 100 réplications pour notre étude, il a fallu déterminer si le passage de 1000 à 100 réplications avait une influence significative sur la valeur de l'estimation globale.

Il a donc été nécessaire de simuler des groupes d'échantillons, en faisant varier le nombre de réplications r de 100 à 1000. Il importe de mentionner que ces échantillons n'ont été utilisés que pour l'étude du nombre de réplications, c'est à dire qu'ils n'ont pas été utilisés dans la comparaison des méthodes. Ainsi, dix groupes d'échantillons de taille $n = 50$ ont été générés à partir de la loi LP3 (tableau 5.1) tels que $r = \{100, 200, 300, 400, 500, 600, 700, 800, 900, 1000\}$. Le biais et l'erreur quadratique moyenne sont les mesures qui ont été considérés pour comparer les résultats obtenus avec les différents groupes.

Dans un premier temps, on a effectué une comparaison du biais et du *RMSE* pour les groupes de 100 et de 1000 réplications. Pour la comparaison, la méthode du maximum de vraisemblance avec validation croisée (*MVVC*) a été utilisée avec un noyau normal. La figure 5.1 présente les résultats de l'estimation des quantiles de période de retour 10, 20, 50, 100, 200 et 1000 ans pour le groupe de 100 échantillons et le groupe des 1000 réplications. Il est important de mentionner qu'une échelle logarithmique est utilisée en abscisse. On remarque sur cette figure, en comparant l'estimation à la valeur théorique, que la différence entre la moyenne des estimations pour 100 et 1000 réplications est relativement faible.



(a) (b)

Figure 5.1 : Estimations des quantiles pour 100 et 1000 réplications de taille $n = 50$, en utilisant la méthode du maximum de vraisemblance (*MVVC*) et un noyau normal.

La figure 5.2 quant à elle, présente la valeur du biais, de l'erreur quadratique moyenne, la moyenne des estimations et la variance des estimations pour les 2 groupes d'échantillons. Le biais du groupe de 100 échantillons est inférieur à celui du groupe de 1000 réplications pour des périodes de retour de 200 et 1000 ans. Par contre, pour des périodes de retour de 50 et 100 ans, le groupe de 1000 réplications a un biais inférieur. Pour des périodes de retour de 10 et 20 ans, les 2 groupes ont un biais similaire. Malgré tout, ces différences demeurent relativement faibles, l'écart maximal entre les 2 courbes étant inférieur à 2 %.

Pour ce qui est du *RMSE*, la différence entre les 2 courbes est relativement faible, comme on peut le constater sur la figure 5.2 (b). Pour des périodes de retour de 100 et 200 ans, on obtient la même valeur de RMSE pour les 2 groupes. Pour les autres quantiles le RMSE du groupe de 100 échantillons est légèrement inférieur à celui du groupe de 1000 échantillons.

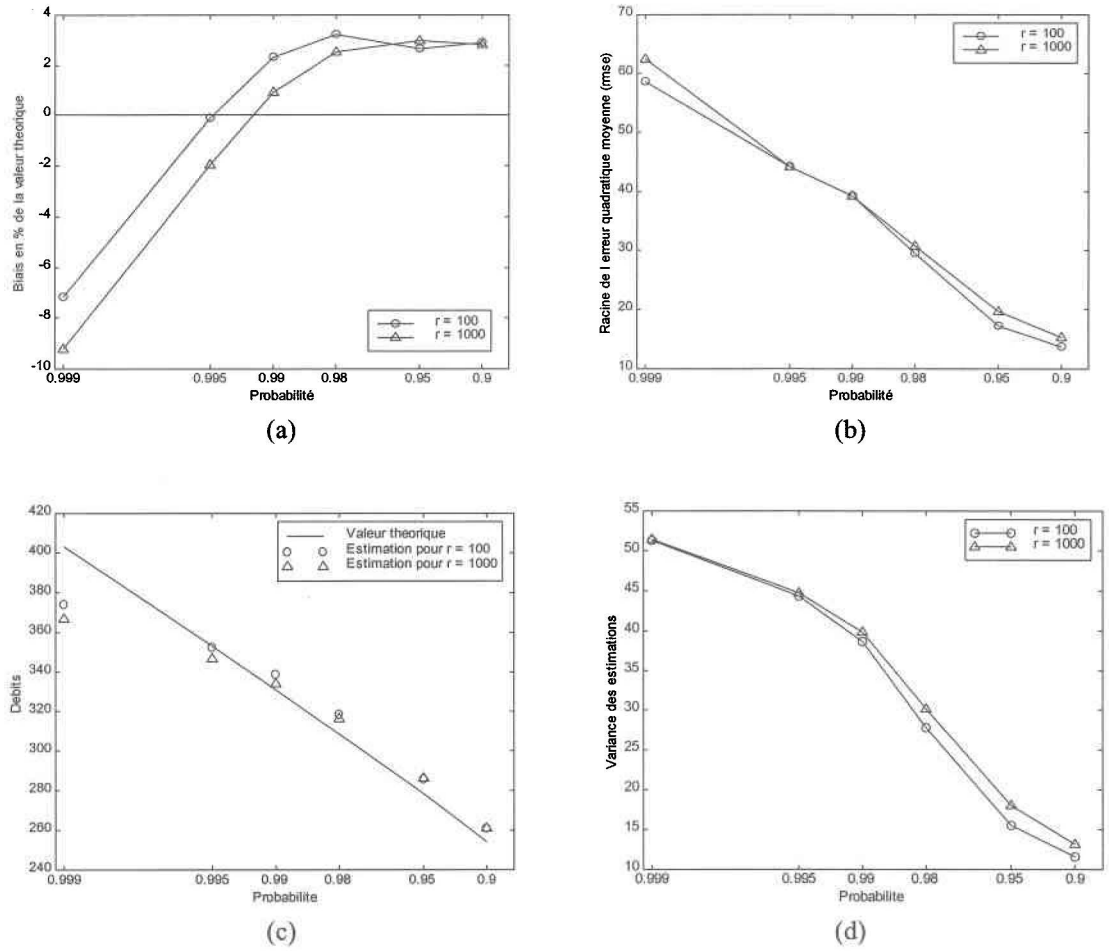


Figure 5.2 : Biais (a), *RMSE* (b), moyenne des estimations (c) et variance des estimations (d) pour 100 et 1000 réplications de taille n = 50, en utilisant la méthode du maximum de vraisemblance (*MVVC*) et un noyau normal.

Ces résultats ne permettent toutefois pas vraiment de conclure que l'usage de 100 réplications est réellement adéquat pour la comparaison des méthodes. Comme l'objectif de l'étude de simulation est de comparer les méthodes de calcul du paramètre de lissage h,

l'étude du nombre de réplications doit se faire dans cette optique. On en vient donc à comparer le biais et l'EQM des différentes méthodes. À titre indicatif, il apparaît raisonnable de considérer deux méthodes parmi celles qui sont utilisées et d'effectuer un test d'égalité des moyennes. Ce test préliminaire est effectué uniquement dans le but de déterminer le nombre de réplications à considérer systématiquement par la suite. La procédure n'est pas assez complète et rigoureuse pour permettre de tirer des conclusions sur les différences entre les méthodes.

Afin de déterminer si les résultats obtenus avec deux groupes différents sont équivalents, l'espérance et la variance du biais et de l'erreur quadratique moyenne ont du être calculées. Voici les expressions de l'espérance et de la variance du biais :

$$E\left[\hat{B}_{T_m}\right] = E\left[\bar{\hat{x}}_{T_m}\right] - x_{T_m} \tag{5.5a}$$

$$V\left[\hat{B}_{T_m}\right] = \frac{1}{r} V\left[\hat{x}_{i,T_m}\right] \tag{5.5b}$$

Pour ce qui est de l'erreur quadratique moyenne, on calcule l'espérance et la variance de la façon suivante :

$$E\left[E\hat{Q}M\right] = E\left[\hat{x}_{i,T_m}^2\right] - 2x_{T_m} E\left[\hat{x}_{i,T_m}\right] + x_{T_m}^2 \tag{5.6a}$$

$$V\left[E\hat{Q}M\right] = \frac{1}{r}\left\{V\left[\hat{x}_{i,T_m}^2\right] + 4\hat{x}_{T_m}^2 V\left[\hat{x}_{i,T_m}\right] - 2\hat{x}_{T_m} Cov\left[\hat{x}_{i,T_m}^2, \hat{x}_{i,T_m}\right]\right\} \tag{5.6b}$$

Considérons deux méthodes quelconques, M_1 et M_2 dont on veut comparer le biais, par exemple. On doit effectuer cette comparaison pour une période de retour en particulier. On peut, pour simplifier la notation, omettre l'indice T_m dans les développements qui vont suivre, toujours en considérant que la procédure est applicable pour toutes les périodes de

retour. On peut raisonnablement supposer que le biais de chacune des méthodes est distribué selon une loi normale :

$$B_1 \approx N\left(\mu_1, \sigma_1^2\right) \quad \text{et} \quad B_2 \approx N\left(\mu_2, \sigma_2^2\right) \tag{5.7}$$

L'hypothèse nulle et l'hypothèse alternative du test peuvent donc être posées ainsi :

$$H_0: \; B_1 = B_2 \;\Leftrightarrow\; B_1 - B_2 = 0$$

$$H_1: \; B_1 \neq B_2 \;\Leftrightarrow\; B_1 - B_2 \neq 0 \tag{5.8}$$

En présence de telles hypothèses, le test classique de comparaisons des moyennes de Student s'impose. Sous l'hypothèse nulle, la différence des deux moyennes (5.5a), divisée par l'estimation de l'écart-type (5.5b), suit une loi de Student avec $r_1 + r_2 - 2$ degrés de liberté, $t_{r_1+r_2-2}$. Considérons $B_{M;i}$, la différence entre l'estimation de l'échantillon i obtenue à partir de la méthode M et la valeur théorique d'un certain quantile, alors la moyenne des $B_{M;i}$ représente le biais B_M tel que défini en (5.2). Sous l'hypothèse H_0, on a que :

$$t_0 = \frac{\hat{B}_1 - \hat{B}_2}{\sqrt{\frac{1}{r}\left(S_1^2 + S_2^2\right)}} \tag{5.9}$$

est distribué selon une Student à $2(r-1)$ degrés de liberté, $t_{\alpha/2;2(r-1)}$. Les variables $\hat{B}_M$ et S_M^2 représentent respectivement, la moyenne et la variance des $B_{M;i}$ pour la méthode M.

L'objectif de ce test est de trouver la valeur de la variable r, le nombre de réplications, qui fait en sorte que l'on rejette l'hypothèse nulle, c'est-à-dire que les deux méthodes ne peuvent pas être jugées équivalentes. On fait l'hypothèse de valeurs typiques pour les B_M et les S_M^2 pour qu'ils soient indépendants du nombre de réplications r. Ensuite on exige que la différence soit supérieure à un certain seuil α. On doit donc isoler la variable r dans l'équation (5.9). En considérant que l'on rejette l'hypothèse nulle si $|t_0| > t_{\alpha/2;2(r-1)}$, on a :

$$r_\alpha = \left(\frac{t_{\alpha/2;2(r-1)} \sqrt{S_1^2 + S_2^2}}{\hat{B}_1 - \hat{B}_2} \right)^2 \qquad (5.10)$$

La valeur critique $t_{\alpha/2;2(r-1)}$ dépend de r, mais on peut supposer que le nombre de réplications soit supérieur à 100 et ainsi remplacer $t_{\alpha/2;2(r-1)}$ par la valeur approximative 1.96 si on considère un seuil $\alpha = 5\%$. Cette valeur est la valeur que l'on obtient lorsque l'on fait tendre le nombre de degrés de liberté vers l'infini. À noter que la précision de la valeur critique n'influence pas significativement les résultats du test. Pour être plus conservateurs, on aurait pu considérer le nombre de réplications comme étant supérieur à 60, par exemple. Avec une valeur critique correspondante de 1.98, la différence pour le calcul de r_α demeure relativement faible ($r_\alpha(r \geq 60) = 1.02\, r_\alpha(r \geq 100)$).

Le nombre de réplications r_α à considérer pour discriminer la méthode du maximum de vraisemblance (*MVVC*) et la méthode des moindres carrés (*MCVC*) a été calculé à l'aide de l'équation (5.10) et ce, à partir des échantillons simulés dont il a été question précédemment. Les résultats présentés au tableau 5.3 démontrent qu'il semble raisonnable, à un niveau de signification de 5%, de considérer au moins une cinquantaine de réplications pour en arriver à effectuer des comparaisons valables des différentes méthodes. Par conséquent, il est relativement conservateur de considérer 100 réplications pour les diverses comparaisons qui vont suivre.

Tableau 5.3 : Nombre de réplications à considérer à un niveau de signification de 5%.

Période de retour (T)	10 ans	20 ans	50 ans	100 ans	200 ans	1000 ans
$r_{0.05}$ (*MVVC* vs. *MCVC*)	6	10	24	41	47	53

Si on considère plutôt un nombre de réplications inférieur à 53, un problème se présentera lors de la comparaison des méthodes. Le faible nombre de réplications ferait en sorte que la variance des estimations des deux méthodes serait trop élevée pour arriver à discriminer les

deux méthodes à un certain seuil. Comme l'objet de ce travail est de dégager les différences entre les méthodes, il est important de tenir compte de ce facteur.

5.3 Types de noyaux

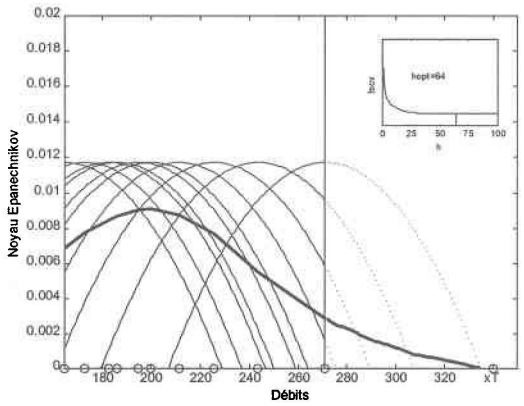
Comme on l'a mentionné au chapitre 3, la performance de six noyaux différents a été étudiée (tableau 3.1). Un des objectifs du présent travail est de comparer les résultats obtenus en utilisant un noyau fini avec ceux obtenus en utilisant un noyau non-fini. On désire également comparer les noyaux symétriques et les noyaux asymétriques. Les critères utilisés pour cette comparaison, sont ceux qui ont été définis à la section 5.1. Par ailleurs, il est important de noter ici que toutes les comparaisons qui seront effectuées par la suite, sont toutes faites à partir des mêmes données synthétiques (100 réplications pour des tailles d'échantillon $n = 10, 25, 50, 100$), qui sont présentées en annexe C.

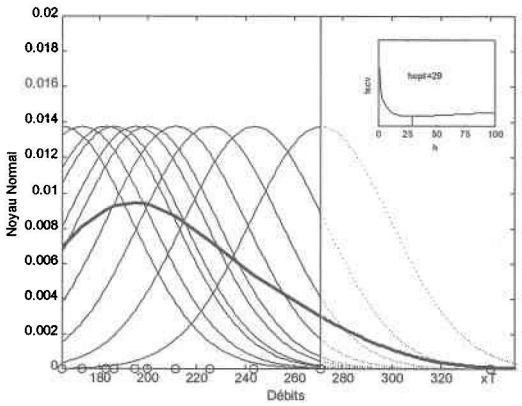
5.3.1 Domaine de définition

Comme on peut le constater au tableau 3.1, trois noyaux à support fini et trois noyaux à support non-fini ont été considérés. On peut noter que le domaine de variation des noyaux à support fini, du noyau Epanechnikov, du noyau rectangulaire et du noyau biweight est le même ($|t| < 1$). Au chapitre 3, on avait examiné le fait qu'un noyau soit borné ou non puisse avoir une influence sur l'estimation dans un contexte d'extrapolation. Dans cette section cette hypothèse sera vérifiée.

5.3.1.1 Remarques sur le domaine de définition

La figure 5.3 illustre bien la différence entre les deux types de noyau. Supposons que nous sommes dans le cas où nous désirons calculer la valeur de la fonction de répartition F pour un débit x_T. Si on prend une valeur x_T élevée, c'est à dire située complètement à l'extérieur de l'échantillon, on se retrouve dans une situation d'extrapolation.





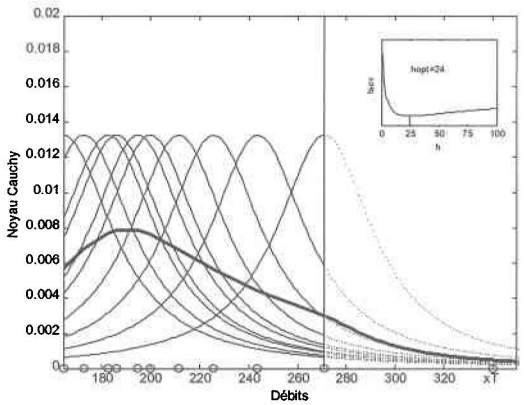


Figure 5.3 : Illustration de la différence entre les noyaux à support finis et ceux à support non-finis pour l'extrapolation.

La figure 5.3 (a) montre qu'aucune probabilité n'est associée à des valeurs supérieures à x_T avec un noyau d'Epanechnikov, puisque même le poids de la dernière observation n'est pas assez important pour couvrir x_T. On pourrait donc craindre une sous-estimation systématique des grands quantiles avec le noyau d'Epanechnikov. Par contre, sur la figure 5.3 (b) et la figure 5.3 (c), on remarque qu'au moins deux noyaux, associés aux deux dernières observations, pourraient permettre d'obtenir une estimation de la fonction F pour un débit x_T. Bien sûr dans cet exemple, étant donné la faible influence des observations de l'échantillon à x_T, on peut mettre en doute la qualité de l'estimation avec un noyau normal. Mais cet exemple n'a été présenté que pour illustrer la faiblesse des noyaux bornés

(Epanechnikov), lorsqu'utilisés dans un contexte d'extrapolation, par rapport aux noyaux non bornés (Cauchy et normal).

Parmi les noyaux à support non-fini, c'est le noyau Cauchy qui a les extrémités les plus lourdes, c'est à dire qu'il attribue des poids plus élevés aux valeurs situées dans les extrémités que les autres noyaux (figure 5.4). Effectivement, on remarque au tableau 3.1, que le noyau Cauchy tend vers zéro très lentement à cause du facteur $1/t^2$. Le fait que le noyau Cauchy ait non seulement une variance infinie, mais aussi une espérance infinie indique jusqu'à quel point les extrémités de ce noyau sont lourdes. L'utilisation des noyaux à queues lourdes diminue le risque de sous-estimation lors de l'extrapolation, mais augmente du même coup le risque de surestimation. Ce type de noyau permet de tenir davantage compte de l'ensemble de l'échantillon, contrairement aux noyaux à queues faibles ou à support fini, qui favorisent plutôt les dernières observations. L'utilisation des noyaux à queues lourdes peut véritablement causer une surestimation puisque l'on augmente l'apport de chacune des observations pour l'extrapolation et par conséquent, on augmente les valeurs de la fonction de densité dans l'extrémité de droite.

Pour l'étude de l'influence du type de support, deux noyaux ont été considérés. La comparaison a donc été effectuée à l'aide d'un représentant de chacun des deux types, en l'occurrence, le noyau Epanechnikov et le noyau normal. La figure 5.4 montre le résultat de l'ajustement du groupe d'échantillons simulés de taille 50 à l'aide de la méthode du maximum de vraisemblance (*MVVC*) pour chacun des noyaux.

Lorsque l'on observe la figure du biais 5.4 (a),on peut considérer que la différence entre les deux courbes n'est pas significative. Par contre, l'utilisation du noyau normal conduit à une erreur quadratique moyenne 5.4 (b) légèrement inférieure à celle du noyau Epanechnikov surtout pour les quantiles supérieurs (200 et 1000 ans). Toutefois, la différence semble

relativement trop faible pour que l'on puisse exclure l'utilisation d'un noyau à support fini par rapport à un noyau à support non-fini.

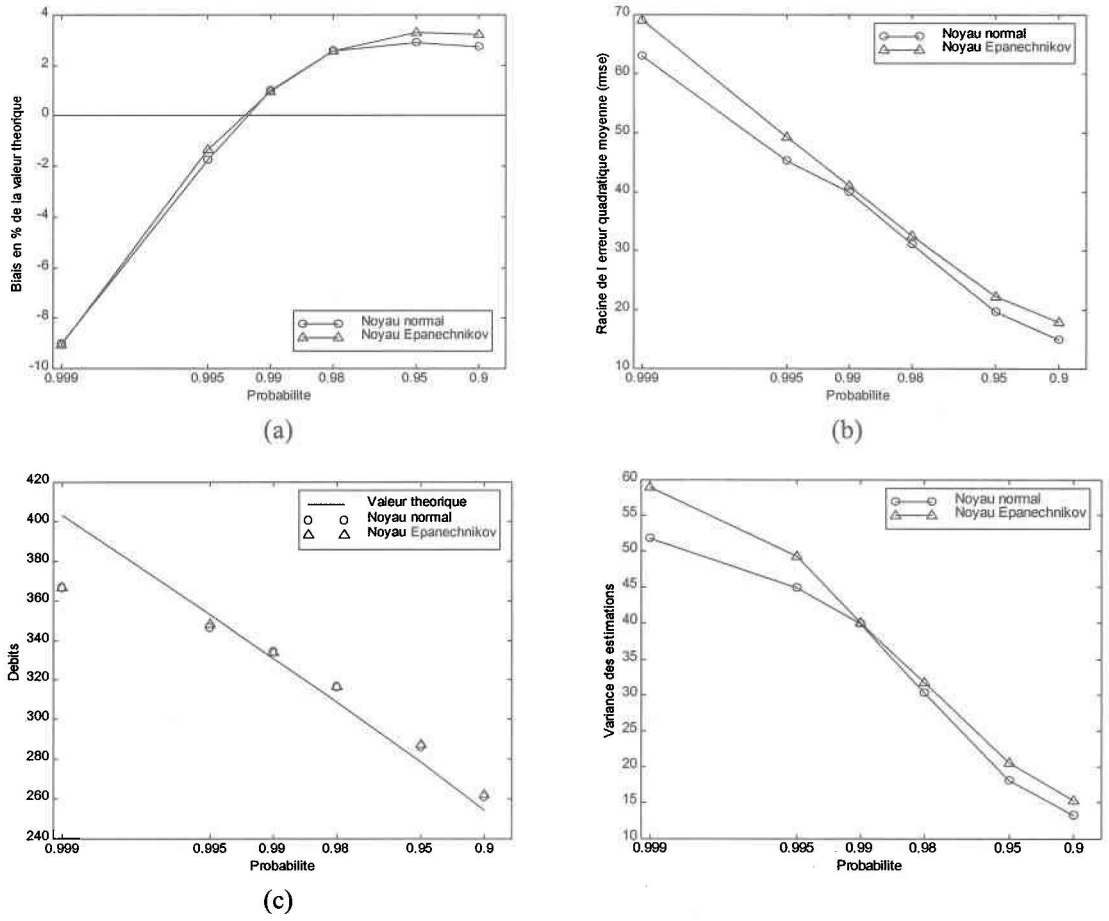


Figure 5.4 : Comparaison des noyaux finis et non-finis en utilisant la méthode du maximum de vraisemblance (*MVVC*) avec *n*=50.

Ces résultats s'expliquent de par le fait que lors du calcul du paramètre de lissage, la méthode utilisée conduit à des paramètres plus grands dans le cas des noyaux bornés que dans le cas des noyaux non-bornés. En considérant une fenêtre plus grande, on augmente le niveau d'extrapolation des noyaux à support fini. La figure 5.5 présente la dispersion des paramètres de lissage calculés pour chaque noyau. Ici encore, la méthode du maximum de vraisemblance (*MVVC*) avec 100 échantillons de taille $n = 50$, a été considérée.

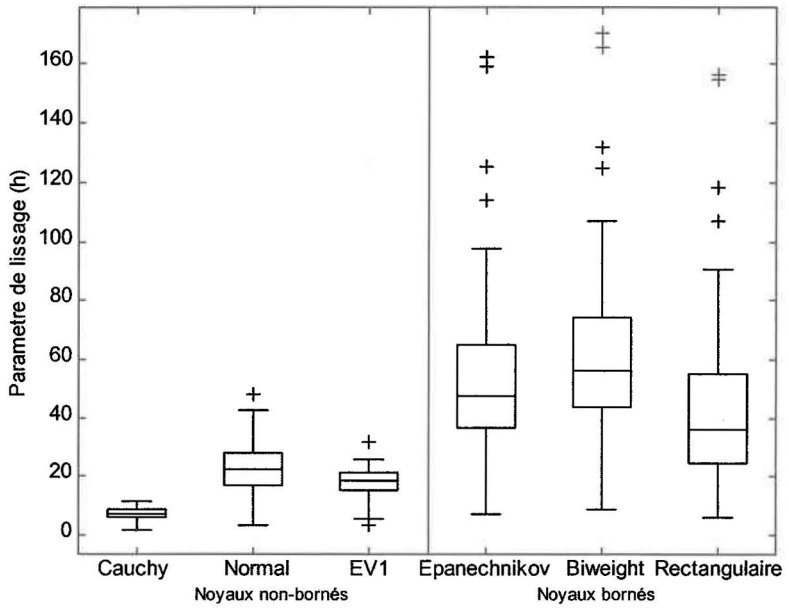


Figure 5.5 : Comparaison des paramètres de lissage selon le type de support, en utilisant la méthode du maximum de vraisemblance (*MVVC*) avec n=50.

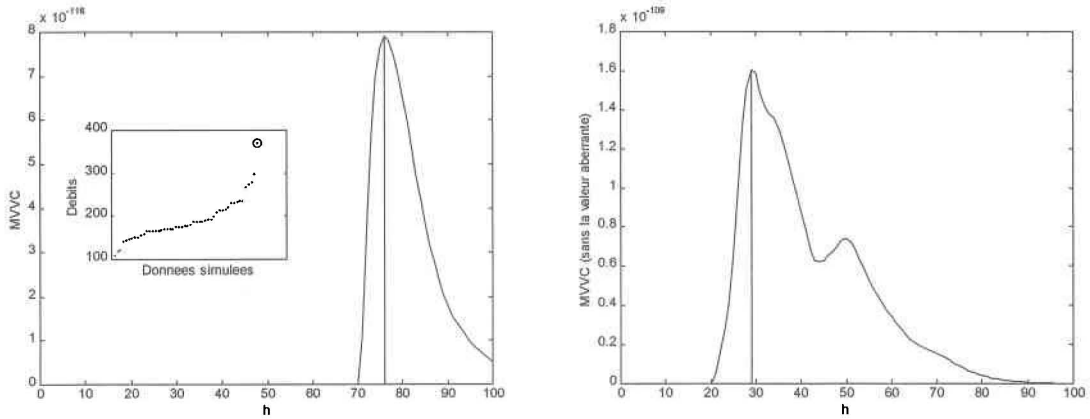
Le noyau qui, parmi les six noyaux considérés, a les queues les plus lourdes, le Cauchy, se voit attribuer des paramètres de lissage nettement plus faibles que les autres noyaux non-bornés. Évidemment, plus les queues d'un noyau sont lourdes, moins le paramètre de lissage a besoin d'être élevé pour pouvoir estimer les quantiles supérieurs.

Pour les noyaux non-bornés (figure 5.5 (a)) la méthode du maximum de vraisemblance donne des paramètres de lissage nettement inférieurs à ceux obtenus en considérant des noyaux bornés (figure 5.5 (b)). Ce phénomène s'explique par la présence de valeurs extrêmes dans les échantillons simulés. Comme la méthode du maximum de vraisemblance est appliquée en utilisant le concept de validation croisée, la fonction de vraisemblance est nulle lorsqu'une observation de l'échantillon se trouve, par rapport au reste de l'échantillon, à une distance supérieure à la longueur du paramètre de lissage. Si par exemple, on considère un paramètre de lissage qui est plus faible que la distance entre la plus grande

observation et la précédente, la valeur de la fonction de densité en ce point sera nulle si on applique la validation croisée. Par le fait même, la valeur de la fonction de vraisemblance est nulle pour ce paramètre de lissage. Ce phénomène s'observe aussi dans le cas de la méthode des moindres carrés avec validation croisée (*MCVC*). En fait, dès que l'on effectue de la validation croisée, l'utilisation d'un noyau à support fini fait en sorte que le paramètre de lissage optimal est élevé.

5.3.1.2 Importance des valeurs extrêmes

La figure 5.6 illustre bien l'importance des valeurs extrêmes. Sur la figure 5.6 (a), on a la fonction objectif (*MVVC*) pour un échantillon complet simulé de taille $n = 50$. Dans ce cas, le maximum de la fonction de vraisemblance se trouve aux environs de h=76. On remarque que la fonction objectif est nulle pour une valeur de h inférieure à 70. La figure 5.6 (b) représente la fonction de vraisemblance calculée à partir du même échantillon, en enlevant toutefois l'observation singulière (donnée encerclée sur la figure 5.6 (a)). En enlevant une seule valeur, la fonction de vraisemblance est complètement différente. La valeur de h optimale se trouve plutôt aux environs de 30 et la fonction de vraisemblance est nulle pour $h < 20$.



(a) (b)

Figure 5.6 : Effet des valeurs extrêmes sur le calcul du paramètre de lissage optimal avec la méthode du maximum de vraisemblance (*MVVC*), pour n=50.

Comme on l'a vu précédemment lors de la comparaison du noyau normal et du noyau d'Epanechnikov, le fait qu'un noyau soit borné ou non, ne semble pas avoir une très grande influence pour l'estimation dans un contexte d'extrapolation, en utilisant une des deux méthodes considérant la validation croisée. Par contre, on a vu que la méthode des noyaux est relativement sensible aux valeurs singulières particulièrement lorsque l'on considère un noyau à support fini. Le paramètre de lissage est très variable par rapport à la distance de la plus grande valeur aux autres valeurs de l'échantillon, ce qui peut considérablement influencer l'estimation et limiter l'extrapolation.

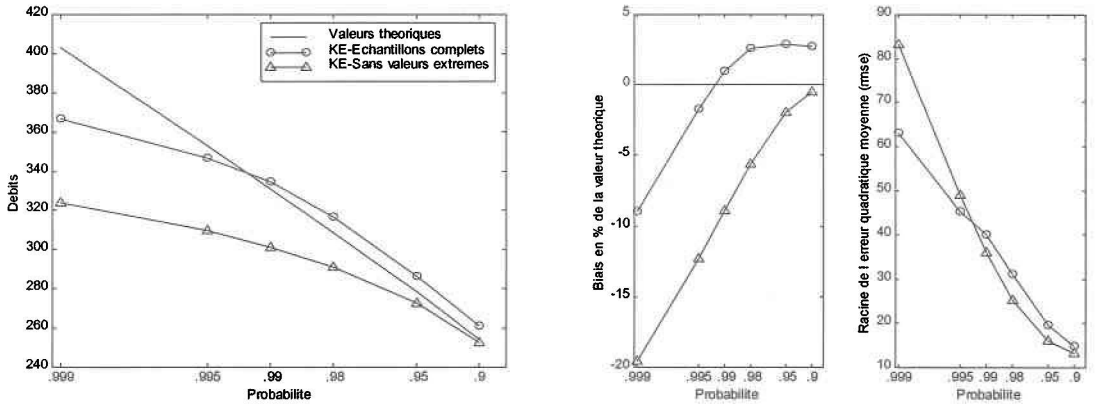


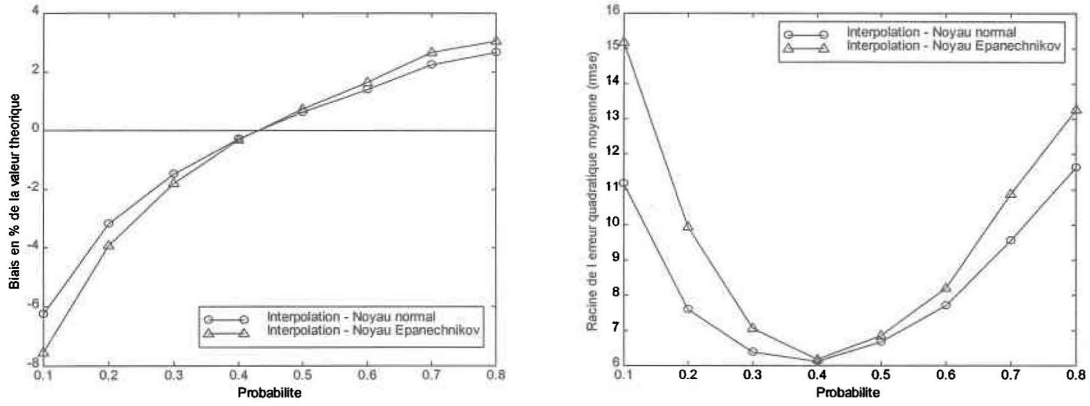
Figure 5.7 : Influence des valeurs extrêmes sur l'estimation selon le type de support du noyau considéré, en utilisant la méthode *MVVC* avec *n*=50.

La figure 5.7 montre que les valeurs extrêmes de l'échantillon sont d'une relative importance pour l'extrapolation. Pour les échantillons que l'on dénotent par « échantillons sans valeurs extrêmes », la plus grande valeur de chaque échantillon a été retirée. D'abord, on remarque que l'estimation est relativement plus faible lorsque l'on considère les échantillons sans leur valeur extrême que pour les échantillons complets. On a une sous-estimation importante pour les échantillons sans valeurs extrêmes. Avec les échantillons complets, on surestime la valeur des quantiles correspondant à des périodes de retour de 10, 20, 50 et 100 ans alors que lorsque l'on supprime les valeurs extrêmes, on sous-estime. Le

fait de ne pas considérer les valeurs extrêmes a pour conséquence d'augmenter le biais pour les grandes périodes de retour. Par contre, la variance est plus faible que dans le cas des échantillons complets sauf pour des périodes de retour de 200 et 1000 ans.

5.3.1.3 Conclusion sur le domaine de définition

Finalement, le type de support ne semble pas avoir une très grande influence sur l'estimation. Il faut toutefois tenir compte du fait que lorsque l'on considère un noyau à support fini, les valeurs extrêmes jouent un rôle plus important pour l'estimation des quantiles d'ordre supérieur que pour un noyau à support non-fini. En accordant un paramètre de lissage relativement élevé, on augmente le degré d'extrapolation, mais on risque de causer un problème pour l'interpolation. L'estimation résultant d'un paramètre de lissage large dépend d'un grand nombre d'observations, l'effet des valeurs voisines de l'estimation étant ainsi atténué par les observations plus éloignées.



(a) (b)

Figure 5.8 : Comparaison des noyaux selon le type de support dans un contexte d'interpolation, en utilisant la méthode *MVVC* avec n=50.

La figure 5.8 présente le résultat de l'estimation de quantiles d'ordre inférieurs (interpolation) en considérant le noyau normal et le noyau d'Epanechnikov. À noter que l'échelle logarithmique n'a pas été considérée dans ce cas-ci contrairement aux figures

relatives à l'extrapolation. Le biais est toujours supérieur pour le noyau borné, sauf pour une probabilité d'environ 0.45 où les deux courbes se croisent. On observe la même chose pour la variance (*RMSE*). Les différences entre les deux courbes sont plus grandes que dans le cas de l'extrapolation, ce qui illustre bien le fait que les valeurs extrêmes ont une plus grande importance lorsque l'on considère un noyau borné.

5.3.1.4 Cas du noyau rectangulaire

Le cas du noyau rectangulaire est un peu particulier par rapport aux autres noyaux à support fini. Comme il est constant sur tout son domaine de définition, la précision que l'on peut atteindre est relativement limitée. La fonction de densité estimée que l'on obtient en considérant ce noyau est relativement irrégulière. Par conséquent, il en est de même pour la fonction de répartition (figure 5.9).

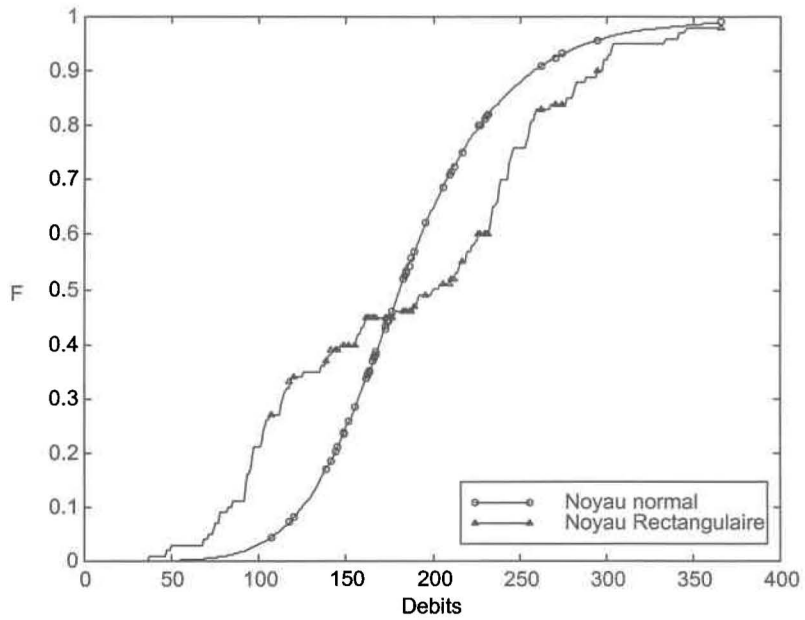


Figure 5.9 : Irrégularité de la fonction de distribution estimée en considérant le noyau rectangulaire. Échantillon de taille *n*=50 en utilisant la méthode *MVVC*.

On observe sur la figure que la fonction de répartition correspondant au noyau rectangulaire est beaucoup moins lisse que celle obtenue avec le noyau normal. En étant construite sous forme de paliers, la fonction de répartition estimée à l'aide du noyau rectangulaire fait en sorte qu'à une valeur de probabilité donnée, corresponde une gamme de débits. Ainsi, l'estimation d'un quantile pour une période de retour donnée correspond à un intervalle de débits, ce qui affecte considérablement la précision de l'estimation par rapport aux autres noyaux. Cet exemple n'est pas effectué dans le but de comparer le noyau rectangulaire au noyau normal, puisqu'il n'est basé que sur un seul échantillon, mais bien pour illustrer le caractère irrégulier de l'estimation dans le cas d'un noyau uniforme sur tout son domaine.

5.3.2 Symétrie

Lorsque Epanechnikov (1969) a obtenu le noyau optimal au sens de l'*IMSE*, il avait établi que le noyau K, utilisé dans l'estimation non paramétrique, devait être symétrique. Toutefois, on s'aperçoit que la propriété de symétrie est surtout utile pour simplifier les calculs menant au noyau Epanechnikov (Silverman ; 1986). Par conséquent, des noyaux asymétriques peuvent également être considérés avec la méthode des noyaux.

5.3.2.1 Asymétrie du noyau EV1

Le noyau EV1 est asymétrique et de façon dont il est utilisé habituellement, fait qu'il est centré sur le mode de la distribution ; c'est à dire que le maximum de la fonction $K(t)$ est atteint pour $t = 0$. On pourrait croire que le fait qu'il soit asymétrique puisse causer un certain biais dans l'estimation, puisque l'on n'a pas la même probabilité de chaque côté du mode (3.30). Des comparaisons ont donc été effectuées dans le but de déterminer à quel endroit on doit centrer le noyau pour minimiser le biais. Pour ce faire, trois statistiques ont été considérés, la moyenne $\bar{x}$ (0.5772 dans le cas d'une EV1 standardisée), la médiane x_{50} et la valeur correspondant au mode ($t = 0$) (figure 5.10 (d)). Les données utilisées pour cette étude sont les 100 réplications de taille 25 simulées à partir de la loi LP3. Le

paramètre de lissage a été estimé par la méthode du maximum de vraisemblance avec validation croisée.

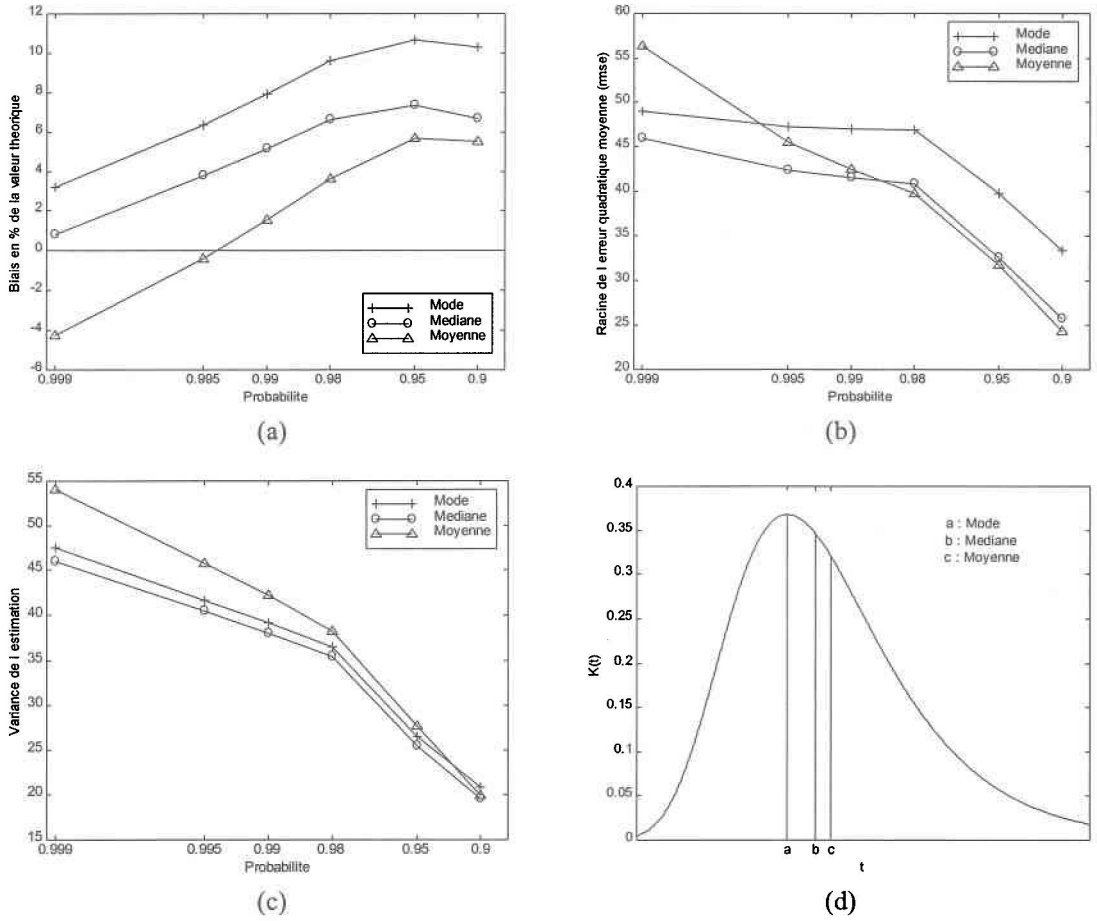


Figure 5.10 : Comparaison des différents types d'asymétrie pour le noyau EV1.

Sur la figure 5.10 (a), on remarque que le biais est minimum lorsque l'on considère la moyenne sauf pour une période de retour de 1000 ans où il semble préférable de centrer le noyau sur la médiane. Pour ce qui est du *RMSE* (figure 5.10 (b)), pour les quatre premières périodes de retour (10, 20, 50 et 100 ans), le fait de centrer sur la médiane ou la moyenne procure des résultats similaires. Pour les quantiles 200 ans et 1000 ans, le *RMSE* est minimum pour la médiane. Le noyau EV1 étant habituellement centré sur le mode, on

remarque toutefois sur les deux figures, 5.10 (a) et (b), qu'il ne semble pas être adéquat pour l'estimation de quantiles de crue.

On doit souligner que l'une des qualités que l'on souhaite retrouver avec les noyaux asymétriques, et entre autres le noyau EV1, c'est leur capacité à effectuer une bonne estimation dans les queues puisqu'ils permettent d'ajouter un poids supplémentaire aux dernières observations. On a vu sur la figure 5.10, que le fait de centrer le noyau sur la médiane ou la moyenne permettait d'améliorer l'estimation. Sur la figure 5.10 (c), on remarque que lorsque l'on centre le noyau sur la médiane, la variance de l'estimation est minimisée par rapport aux deux autres statistiques. Le noyau EV1 a donc été décalé sur la médiane.

5.3.2.2 Remarques sur la symétrie

Parmi les noyaux présentés au tableau 3.1, seul le noyau EV1 est asymétrique autour de 0. Comme on peut le constater à la figure 3.2, l'extrémité droite du noyau EV1 est relativement plus lourde que celle de gauche, ce qui fait en sorte que l'estimation de l'extrémité de droite dépend de presque tout l'échantillon, alors que celle de gauche ne dépend que d'une certaine partie des données de l'échantillon. De cette façon, un poids supérieur est associé aux dernières observations de l'échantillon pour l'estimation et la queue asymptotique de droite permet d'augmenter le degré d'extrapolation sans pour autant nécessiter un paramètre de lissage relativement grand. On s'attend donc à ce que le noyau EV1 soit plus efficace pour l'extrapolation que les autres noyaux.

La figure 5.11 présente le résultat de la comparaison d'un noyau symétrique avec le noyau EV1. Le noyau normal est le noyau symétrique qui a été considéré pour cette comparaison, les résultats obtenus avec ce noyau étant comparables avec ceux obtenus en considérant le noyau Epanechnikov. La méthode du maximum de vraisemblance avec validation croisée pour $n = 50$ a été considérée. On remarque que le noyau asymétrique conduit à un biais et

une variance inférieurs à ceux obtenus avec le noyau symétrique pour les quantiles élevés, c'est à dire pour $T = 200$ et $T = 1000$. La différence entre les deux courbes du biais pour une période de retour de 1000 ans atteint même 4%. Pour les autres périodes de retour, le noyau symétrique procure une meilleure estimation.

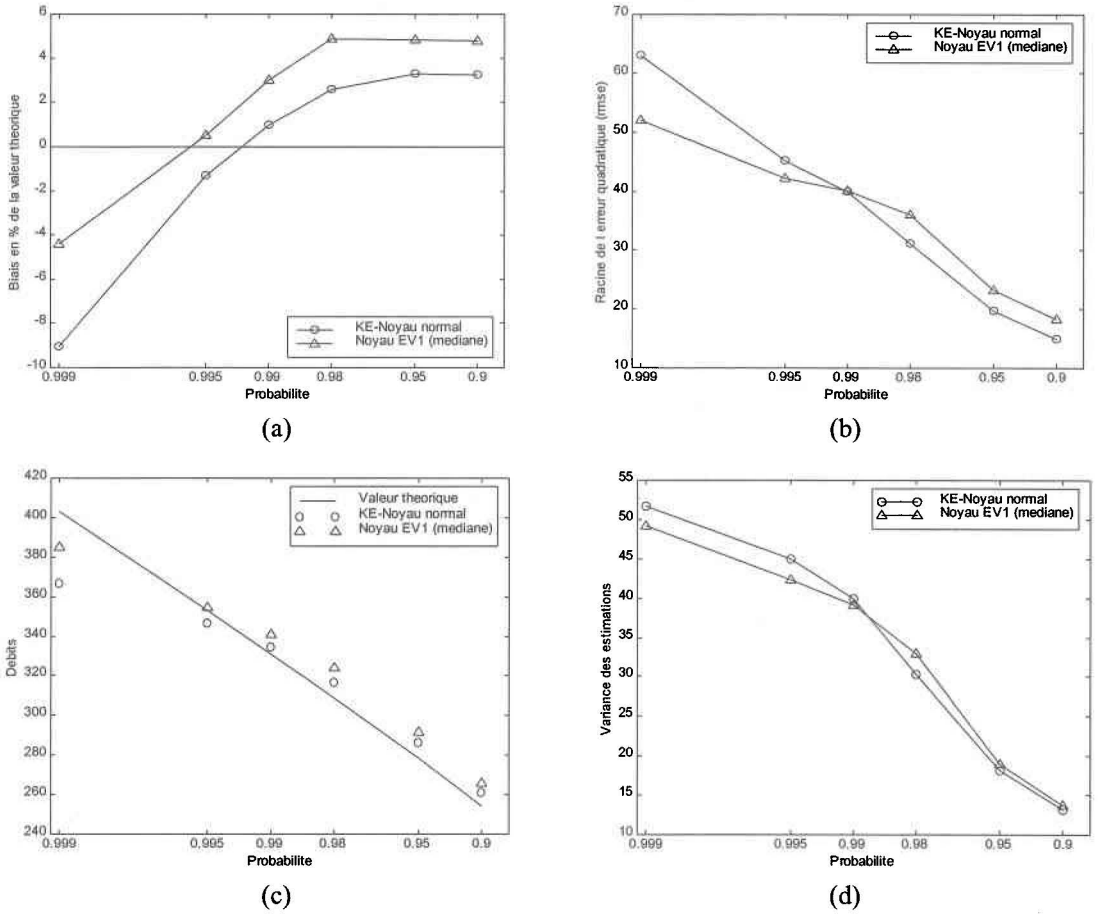


Figure 5.11 : Comparaison de noyaux symétrique et asymétrique, en utilisant la méthode *MVVC* avec *n*=50.

Il est relativement difficile de tirer des conclusions générales quant à la symétrie que le noyau doit posséder pour l'estimation de quantiles avec la méthode des noyaux. Toutefois, il apparaît évident que l'utilisation d'un noyau asymétrique n'augmente pas le biais dans les extrémités, puisqu'il permet de le réduire dans le cas d'un noyau à asymétrie positive comme c'est le cas pour le noyau EV1.

5.3.3 Conclusions sur l'utilisation des noyaux

Les résultats obtenus pour les diverses simulations effectuées dans le cadre de cette étude permettent de conclure que, contrairement à ce qui a généralement été postulé dans la littérature, le choix du noyau à considérer pour l'estimation non paramétrique de grands quantiles peut être d'une relative importance. Une certaine gamme de noyaux procurent une estimation acceptable, c'est le cas du noyau normal, du noyau Epanechnikov et du noyau biweight. Par contre, le noyau Cauchy a les queues trop lourdes et il provoque une surestimation des quantiles estimés en extrapolant. Le noyau rectangulaire, quant à lui, limite le degré d'extrapolation et semble être trop simpliste pour être efficace, même pour l'interpolation étant donné qu'il est constant sur tout son domaine de variation. Le seul noyau asymétrique considéré, le noyau EV1, permet de réduire le biais et la variance pour des quantiles élevés. Toutefois, pour les quantiles de plus faible période de retour (10, 20, 50 et 100 ans), il ne donne pas de meilleurs résultats que les trois noyaux symétriques (normal, Epanechnikov et biweight). On a vu à la section 1.3.1.1 que les noyaux à support fini (Epanechnikov et biweight) conduisent à des paramètres de lissage relativement élevés et qu'ils sont sensibles aux valeurs extrêmes. Cette sensibilité peut jouer un rôle important dans l'estimation lorsque l'on considère la méthode des noyaux pour les données d'un seul échantillon.

Bien sûr, le choix du noyau n'est pas aussi critique que le calcul du paramètre de lissage, mais il demeure important de choisir un noyau qui n'augmente pas inutilement le biais et la variance de l'estimation induites par l'estimation de h.

5.4 Taille d'échantillon

L'efficacité des méthodes non paramétriques lorsque l'on a des échantillons de faible taille peut être mise en doute, puisque celles-ci ont la propriété de très bien s'ajuster aux données

de l'échantillon. En ne disposant que d'un faible nombre d'observations, les chances que l'échantillon ne soit pas représentatif de la population sont grandes. Par contre, avec les méthodes paramétriques, on peut souvent avoir une idée du type de distribution de la population et on peut ainsi imposer un modèle aux données de l'échantillon, ce qui réduit le risque d'effectuer une mauvaise estimation. Pour cette raison, quatre tailles d'échantillon ont été considérées dans la présente étude, n = 10, 25, 50 et 100. L'objet de cette section est de tenter de mettre en évidence les différences existant pour l'estimation selon les tailles d'échantillon considérées de façon à évaluer l'importance de la taille d'échantillon dans l'estimation non paramétrique.

La figure 5.12 présente les résultats obtenus pour les différentes tailles d'échantillon. Comme pour les autres comparaisons concernant les noyaux, la méthode du maximum de vraisemblance avec validation croisée a été considérée pour estimer les paramètres de lissage de chacun des groupes d'échantillons avec un noyau normal. Évidemment, le biais et la variance sont inversement proportionnels à la taille de l'échantillon. Toutefois, on remarque sur la figure de gauche, que pour des périodes de retour de 50 et 100 ans, le biais inférieur est atteint pour des tailles d'échantillons faibles. En ce qui a trait aux autres périodes de retour, le biais est minimum pour une taille d'échantillon de 100 et atteint sa valeur maximale pour n = 10. Il est donc relativement difficile, à partir du biais, de tirer des conclusions générales concernant la taille d'échantillon. Cependant, on constate que le biais est minimisé pour de très grandes et de faibles périodes de retour, en maximisant la taille de l'échantillon.

Les résultats sont plus éloquents en ce qui a trait au *RMSE*. Sur la figure 5.12 (b), on remarque que la variance est directement inversement proportionnelle à la taille de l'échantillon, le groupe d'échantillons de taille 100 étant celui qui minimise le *RMSE* et ce, pour tout les quantiles considérés.

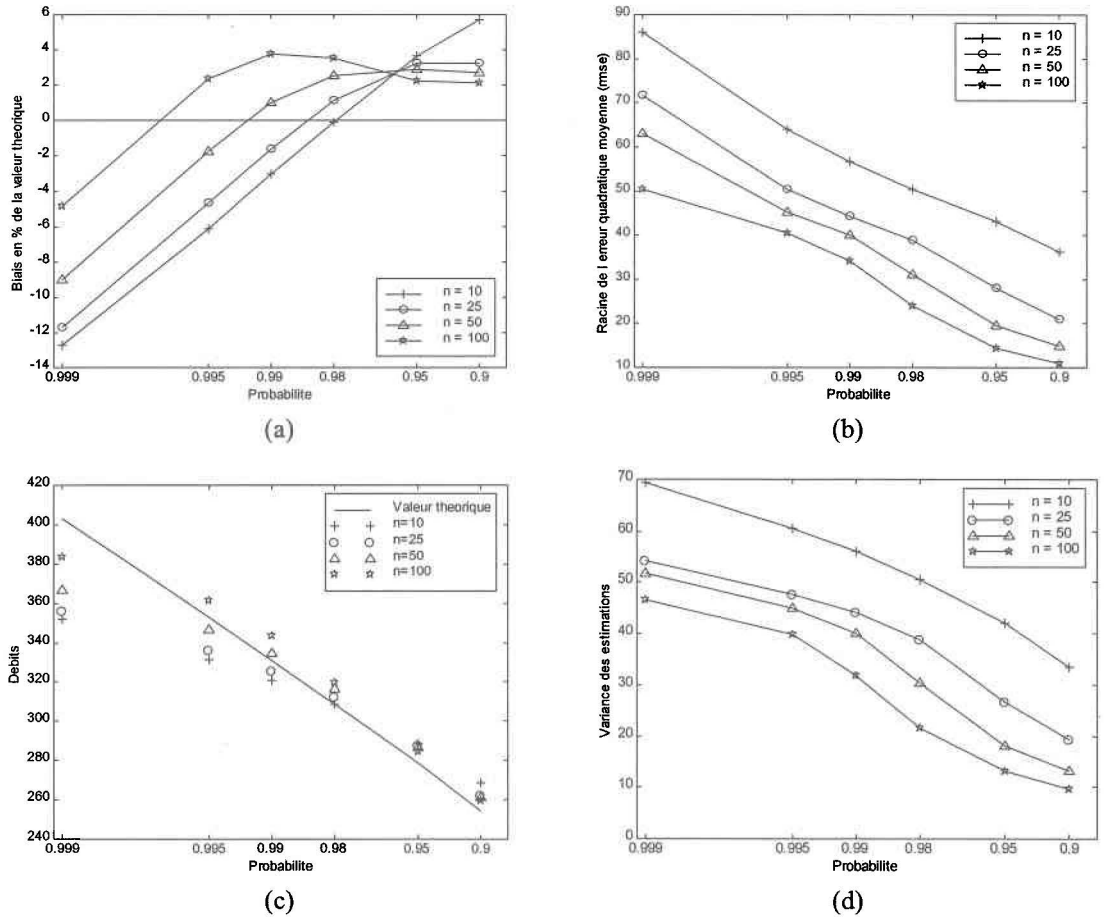


Figure 5.12 : Comparaison des résultats obtenus pour les diverses tailles d'échantillon, en utilisant la méthode *MVVC*.

Évidemment, comme la taille de l'échantillon permet de réduire le *RMSE*, il demeure préférable que *n* soit le plus élevé possible. Afin de pouvoir doser l'importance de la taille d'échantillon, une comparaison a été effectuée avec une méthode paramétrique. La figure 5.13 met en évidence le rôle de la taille de l'échantillon sur la valeur du biais et du *RMSE* pour la méthode des noyaux et pour une méthode paramétrique. La méthode non paramétrique considérée est la méthode du maximum de vraisemblance avec validation croisée (*MVVC*) avec un noyau normal, tandis que la loi généralisée des valeurs extrêmes (GEV) avec la méthode des moments pondérés a été considérée pour représenter la méthode paramétrique utilisée dans la comparaison. Les courbes de variation du biais ont été

obtenues en soustrayant du biais obtenu pour le groupe d'échantillons de taille n=10 le biais obtenu pour les échantillons de taille 100. Les biais sont représentés en pourcentage de la valeur théorique x_{T_m}. Les valeurs sur la figure 5.13 (a) représentent donc la valeur de réduction du biais en pourcentage de la valeur théorique, lorsque l'on augmente la taille des échantillons de 10 à 100. Les courbes de variance (figure 5.13 (b)) ont été obtenues quant à elles en soustrayant du *RMSE* pour n = 10 le *RMSE* correspondant pour n = 100 et en divisant par la valeur obtenue pour n = 10. Ce pourcentage relatif représente le taux de réduction de la variance que l'on obtient en augmentant la taille des échantillons de 10 à 100.

En résumé, voici la façon dont le biais relatif et la variance relative ont été calculés pour les deux méthodes (*MVVC* et GEV) :

$$\left\{B_{T_m}\right\}_r = \frac{B_{T_m}\{n=10\} - B_{T_m}\{n=100\}}{x_{T_m}} \qquad (5.11a)$$

$$\left\{RMSE_{T_m}\right\}_r = \frac{RMSE_{T_m}\{n=10\} - RMSE_{T_m}\{n=100\}}{RMSE_{T_m}\{n=10\}} \qquad (5.12b)$$

où B_{T_m}, $RMSE_{T_m}$ et x_{T_m} sont respectivement le biais, la racine de l'erreur quadratique moyenne et le quantile théorique pour une période de retour T_m.

Ces résultats montrent que la taille de l'échantillon semble avoir une importance plus grande pour l'ajustement de la loi GEV que pour la méthode des noyaux. Sur la 5.13 (a), la réduction du biais est supérieure pour la GEV, sauf pour l'estimation des deux quantiles inférieurs. Par contre, si on examine la figure D2 en annexe, on remarque que pour les deux autres distributions statistiques considérées dans cette étude, en l'occurrence la LP3 et la LN2, la réduction du biais est inférieure à celle obtenue avec la méthode des noyaux. Sur le

la figure 5.13 (b), il est clair que l'augmentation de la taille de l'échantillon favorise la réduction de la variance de la méthode paramétrique par rapport à la méthode non paramétrique. En résumé, l'augmentation de la taille de l'échantillon serait un facteur plus déterminant pour la réduction du biais dans le cadre de la méthode des noyaux que dans le cas de la majorité des méthodes paramétriques (deux méthodes sur trois). Par contre, l'accroissement de *n* aurait un plus grand impact sur la variance de l'estimation dans le cas des trois distributions que dans le cas de la méthode des noyaux. La variation du biais et du *RMSE* selon la taille de l'échantillon est présentée en annexe D3 et D4 pour chacune des périodes de retour.

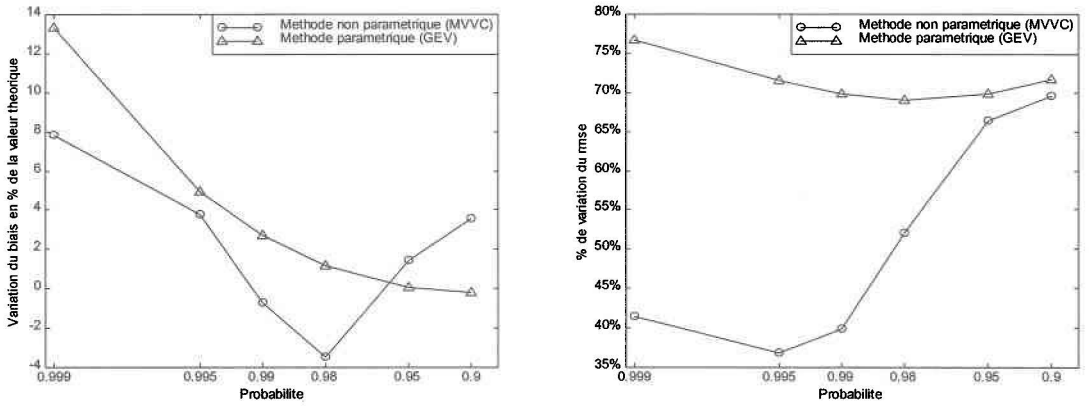


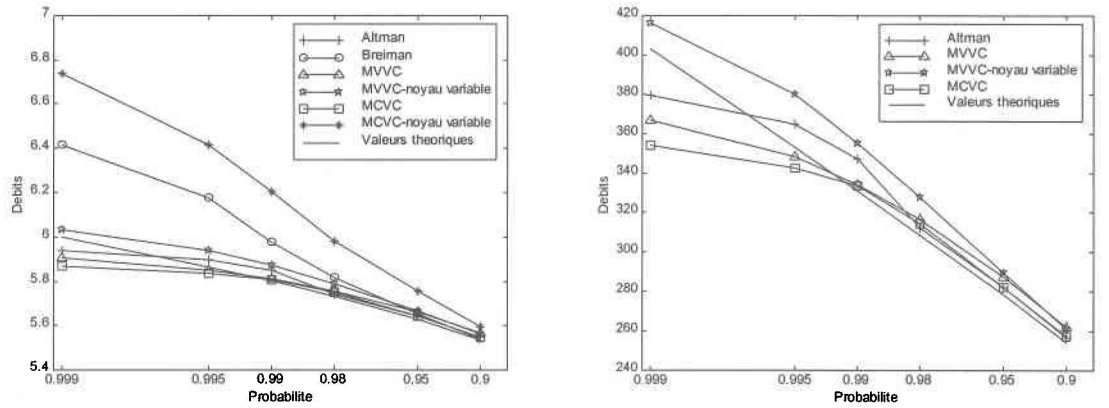
Figure 5.13 : Impact de la taille de l'échantillon sur le biais et le *RMSE* (différences des résultats obtenus pour *n*=10 et ceux obtenus pour *n*=100).

Bien sûr, même si selon les résultats obtenus, l'augmentation de la taille des échantillons a pour effet d'augmenter légèrement le biais pour l'estimation de certains quantiles (50 et 100 ans) avec la méthode des noyaux, il est important de considérer la taille de l'échantillon comme un facteur déterminant dans l'estimation. Le degré de représentativité de la population croit avec la taille de l'échantillon.

5.5 Comparaisons des méthodes de calcul de *h*

La présente section est consacrée à la comparaison des différentes méthodes de calcul du paramètre de lissage présentée au chapitre 4. Comme on l'a mentionné à la section 4.4.1, le critère d'Adamowski a été exclu de l'étude de comparaison. De plus, la méthode *plug-in* de Gasser *et al.* est traitée séparément pour une raison qui sera donnée plus loin.

Les résultats présentés dans cette section sont tous issus du groupe de 100 échantillons de taille 50, en considérant un noyau d'Epanechnikov. Les résultats correspondants aux autres tailles d'échantillon et les autres noyaux sont inclus en annexe D.



(a) (b)

Figure 5.14 : Estimation des quantiles pour les différentes méthodes d'estimation du paramètre de lissage (noyau d'Epanechnikov, $n = 50$).

Sur la figure 5.14 on retrouve l'estimation des six quantiles, c'est à dire la moyenne des estimations pour les 100 échantillons du groupe, pour chacune des méthodes d'estimation de *h*. On remarque sur la figure 5.14 (a) que deux méthodes (*MCVC* à fenêtre variable et Breiman) en particulier conduisent à une surestimation relativement importante par rapport aux autres. Seules les quatre autres méthodes sont présentées sur la figure 5.14 (b), de façon à mieux visualiser les résultats. Une échelle logarithmique a été utilisée sur la figure 5.14

(a), afin de compresser l'ordonnée et ainsi visualiser les résultats des six méthodes sur une même figure. La variance de l'estimation est présentée en annexe D5.

La méthode de Breiman *et al.* et la méthode des moindres carrés à fenêtre variable sont les méthodes qui surestiment trop pour être considérées dans un contexte d'extrapolation. Parmi les quatre autres méthodes, ce sont la méthode de Altman et la méthode des moindres carrés qui conduisent à la meilleure estimation pour les trois quantiles inférieurs. Pour les périodes de retour de 100 et de 200 ans, la méthode du maximum de vraisemblance et des moindres carrés , toutes deux avec validation croisée, semblent être les méthodes les plus efficaces. Par contre, au-delà de 200 ans, c'est la méthode de Altman qui est supérieure en sous-estimant d'environ 24 m^3/sec pour T =1000 et la méthode du maximum de vraisemblance à fenêtre variable qui surestime le quantile 1000 ans de moins de 14 m^3/sec.

Quant au biais et à la variance, ils sont présentés à la figure 5.15. Les conclusions que l'on peut tirer à partir de la figure 5.15 (a) sont sensiblement les mêmes que dans le cas de l'estimation (figure 5.14). On remarque toutefois que les valeurs de biais pour l'ensemble des méthodes demeurent relativement faibles, ne dépassant pas 12% de la valeur théorique, même pour une période de retour aussi grande que 1000 ans. Il faut être conscient que la méthode des noyaux, en étant basée uniquement sur les observations elles-mêmes, devrait être quelque peu limitée pour l'extrapolation à un degré supérieur. Il est donc assez surprenant d'obtenir de pareils résultats pour une période de retour de 1000 ans.

Sur la figure 5.15 (b), on observe que les deux méthodes qui minimisent le *RMSE* sont la méthode de Altman et la méthode des moindres carrés avec validation croisée. À noter que la différence de la variance de l'estimation du quantile de période de retour de 1000 ans pour la méthode de Altman et les trois autres méthodes est relativement importante.

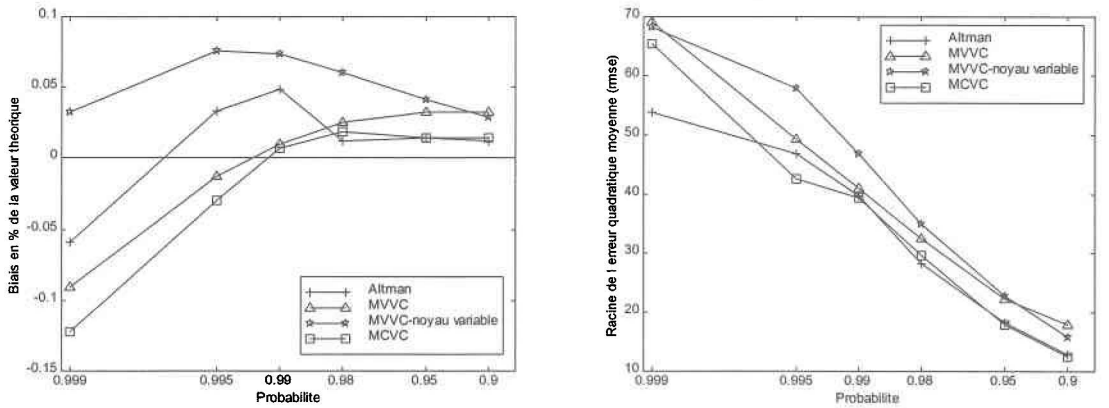


Figure 5.15 : Biais et *RMSE* pour les différentes méthodes d'estimation du paramètre de lissage (noyau d'Epanechnikov, n = 50).

Dans les six sections suivantes, les résultats obtenus pour chacune des méthodes sont brièvement discutés séparément de façon à dégager et à analyser les performances et faiblesses. Une section supplémentaire permettra de commenter les résultats obtenus avec la méthode *plug-in* de Gasser et Lall.

5.5.1 Méthode de Altman et Léger

D'après la revue de littérature effectuée pour ce travail, il semble que ce soit la première fois que la méthode de Altman et Léger est appliquée dans un contexte hydrologique. De plus, la façon dont elle a été présentée par Altman et Léger n'est pas adaptée pour être applicable à des données de débits, comme on l'a vu à la section 4.4.2.

Les résultats obtenus avec cette méthode sont intéressants, puisqu'elle permet d'obtenir une estimation raisonnable jusqu'à une période de retour de 1000 ans. Elle est une des méthodes avec laquelle on obtient la meilleure estimation pour le groupe d'échantillon de taille 50 et elle est la méthode avec laquelle on obtient la plus faible variance. Elle est relativement efficace pour des périodes de retour de 10, 20, 50 et 1000 ans. Un problème semble toutefois se produire pour l'estimation des quantiles de période de retour de 100 et 200 ans,

étant donné la discontinuité de la courbe du biais. Il est possible que ce problème provienne du fait que les minimums des courbes de la figure 4.2 à la section 4.4.2 ne sont pas clairement définis. L'analyse de sensibilité du paramètre p a été effectuée sur 25 échantillons de taille n = 50. Il aurait peut-être été préférable de considérer un plus grand nombre de réplications.

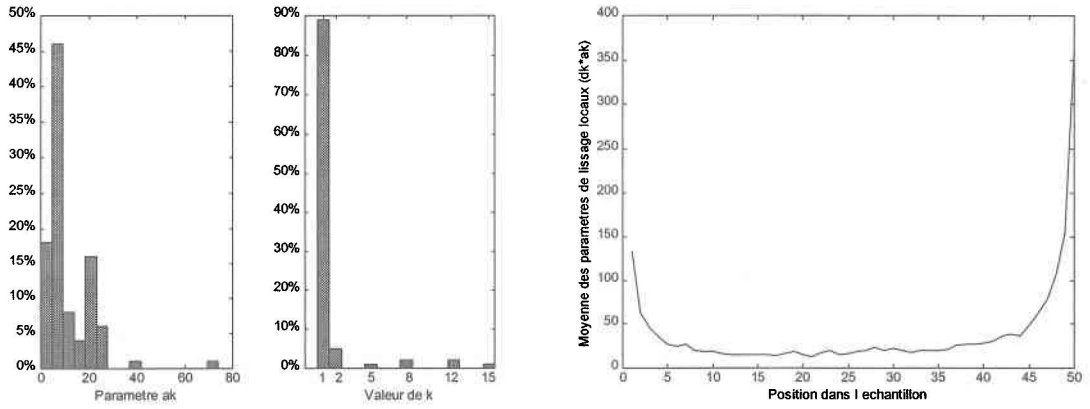
Malgré l'irrégularité dans la courbe du biais pour les quantiles de période de retour de 100 et de 200 ans, la méthode de Altman et Léger procure une bonne estimation des quantiles et mérite d'être considérée pour l'analyse de fréquence de crue. Elle possède l'avantage d'être complètement automatique, puisque h est estimé directement, sans avoir à effectuer une minimisation. De plus, elle est obtenue à partir de la fonction de répartition au lieu de la fonction de densité, ce qui semble être davantage acceptable dans un contexte d'estimation de quantiles (cf. section 4.2).

5.5.2 Méthode de Breiman *et al.*

Comme toutes les méthodes à fenêtre variable considérées dans cette étude, la méthode de Breiman *et al.* conduit à des résultats relativement décevants. Avec cette méthode on surestime les quantiles pour toutes les périodes de retour considérées. La différence entre la valeur du quantile théorique et l'estimation croît avec la période de retour, de sorte que l'on obtient un biais positif maximal pour T = 1000 ans. On remarque que la méthode de Breiman et les deux autres méthodes à fenêtre variable (*MVVC* et *MCVC*) surestiment le quantile 1000 ans alors que les autres méthodes ont tendance à le sous-estimer.

La figure 5.16 présente la répartition des paramètres a_k, des valeurs du voisin le plus proche à considérer dans l'estimation ainsi que les valeurs moyennes des paramètres locaux pour le groupe d'échantillon de taille 50 en utilisant le noyau Epanechnikov. La figure 5.16 (a) permet d'établir que les paramètres a_k sont relativement faibles puisqu'un peu moins de 20% des valeurs de a_k du groupe d'échantillon sont inférieures à 5 et environ 65% sont

inférieures à 10. Sur la figure 5.16 (b), on remarque que le voisin le plus proche à considérer est dans 90% des cas le premier (k=1). La distance au k^e voisin le plus proche est la variable qui a le plus de poids dans l'estimation des paramètres locaux. Effectivement, la valeur du paramètre a_k ne compte en moyenne que pour moins de 20% du produit $a_k d_{ki}$ et par conséquent, 80% de la valeur du paramètre de lissage local provient de la distance au k^e voisin le plus proche.



(a) (b)

Figure 5.16 : Répartition des valeurs des paramètres a_k, des valeurs du voisin le plus proche k à considérer ainsi que la valeur des paramètres de lissage locaux pour la méthode de Breiman appliquée au groupe d'échantillons de taille n = 50.

Sur la figure 5.16 (c), la moyenne des paramètres de lissage locaux est présentée selon la position de chacune des observations dans l'échantillon. Par exemple pour l'observation #3, si on considère les données rangées en ordre croissant, le produit $a_k d_{ki}$ est d'environ 45 m^3/sec en moyenne pour l'ensemble des échantillons. On remarque donc sur cette figure que le paramètre attribué aux valeurs aux extrémités, c'est à dire les faibles et les forts débits, est considérablement plus élevé que pour les autres observations de l'échantillon.

5.5.3 Méthode du maximum de vraisemblance avec validation croisée

Cette méthode basée sur la fonction de densité permet d'obtenir une estimation relativement précise des quantiles de période de retour de 100 et de 200 ans. Mais pour une période de retour de 1000 ans, elle sous-estime considérablement le quantile (figure 5.14 (b)). Pour les quantiles de 10, 20 et 50 ans, elle conduit à une estimation raisonnable, mais elle surestime toutefois davantage que les autres méthodes. Pour une méthode dérivée à partir de la fonction de densité, elle semble tout à fait convenable, puisqu'elle conduit à un biais inférieur que celui de la méthode de Altman et Léger pour les quantiles 100 et 200 ans, cette dernière étant basée sur la fonction de répartition.

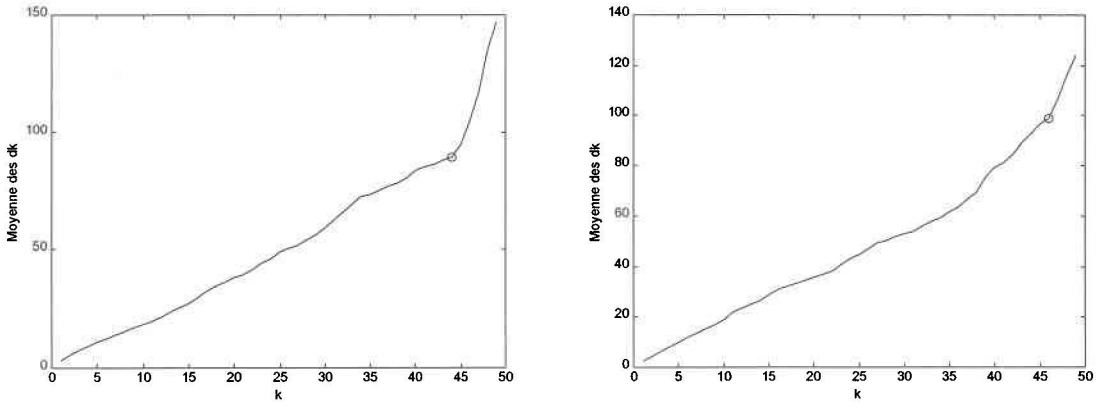
Comme il a été mentionné précédemment, cette méthode lorsqu'elle est utilisée avec un noyau à support fini, peut conduire à des valeurs de paramètres de lissage élevées. Lorsque l'échantillon d'intérêt comporte des valeurs quelque peu extrêmes, pour de faibles valeurs de h, la fonction de vraisemblance est nulle à cause de la validation croisée (section 5.3.1.1). Pour que la fonction de vraisemblance soit non-nulle, le paramètre de lissage doit être relativement élevé, de sorte qu'il y ait toujours, pour chacunes des observations, au moins une autre observation située à l'intérieur de la fenêtre h. Il est donc important de bien étudier la dispersion de l'échantillon à partir duquel on veut effectuer une estimation puisque la présence de valeurs extrêmes peut induire une surestimation des quantiles. Par ailleurs, l'étude de simulation a permis d'observer que lorsque l'échantillon contient au moins deux valeurs identiques, la méthode du maximum de vraisemblance avec validation croisée conduit à une estimation nulle du paramètre de lissage optimal.

5.5.4 Méthode du maximum de vraisemblance à fenêtre variable

Cette méthode de calcul des paramètres de lissage provoque une surestimation de tous les quantiles considérés (figure 5.14 (b)). Elle conduit tout de même à des résultats relativement plus précis que les deux autres méthodes à fenêtre variable considérées (figure

5.14 (a)). Pour le groupe d'échantillons de taille 50, elle procure la meilleure estimation pour le quantile de période de retour de 1000 ans et ce, à moins de 14 m^3/sec près, correspondant à une erreur relative d'environ 3%. Dans les cas où l'analyse de fréquence de crue sert au dimensionnement d'un ouvrage, il est plus sécuritaire de surestimer le débit de période de retour T servant au dimensionnement que de le sous-estimer. Pour ce qui est de la variance, elle est relativement élevée par rapport aux autres méthodes, mais elle est tout de même à peu près du même ordre que celle obtenue avec son équivalent pour fenêtre fixe.

Le choix de la valeur de k, le voisin le plus proche servant au calcul du paramètre de lissage, demeure relativement subjectif. Comme on a vu à la section 4.6.2, la valeur de k à considérer correspond à la cassure dans la courbe de $\overline{d}_k$ versus k. Cette procédure pour trouver la valeur de k a été jugée optimale par le biais d'études empiriques (Breiman *et al.* ; 1977). Certaines simulations effectuées dans le cadre de ce travail ont permis de remarquer que la cassure dans la courbe n'est pas toujours très apparente et qu'il n'est pas très sûr de ne se fier qu'à l'évaluation visuelle de la courbe.



(a) (b)

Figure 5.17 : Illustration de la difficulté à évaluer visuellement l'emplacement de la cassure sur le graphique de $\overline{d}_k$ versus k pour la méthode du maximum de vraisemblance à fenêtre variable.

La figure 5.17 permet de visualiser à quel point il peut être parfois difficile d'identifier la cassure sur la courbe de $\overline{d}_k$ versus k. Les deux exemples qui constituent cette figure proviennent des échantillons #19 et #5 respectivement. Les cercles sur les courbes sont les valeurs obtenues à l'aide de la méthode de la dérivée seconde. Par conséquent, la procédure qui a été préconisée dans cette étude à la section 4.6.2 pour automatiser le choix de k à partir du calcul de la dérivée seconde, permet d'éviter quelque peu la subjectivité du choix visuel.

5.5.5 Méthode des moindres carrés avec validation croisée

Avec la méthode de Altman et Léger, la méthode des moindres carrées avec validation croisée est celle qui procure la meilleure estimation des quantiles en général. Pour des périodes de retour de 10, 20, 50 et 100 ans elle permet d'estimer les quantiles correspondants avec un biais de moins de 2%. Par contre, pour des périodes de retour de 200 et 1000 ans, elle est moins efficace que la méthode du maximum de vraisemblance avec validation croisée et la méthode de Altman. Elle procure alors une estimation bien inférieure à la valeur du quantile théorique, ce qui n'est pas très sécuritaire. La variance obtenue avec cette méthode est relativement faible, elle est comparable à celle obtenue avec la méthode de Altman et Léger.

Un inconvénient majeur de cette méthode provient du fait qu'elle est peu efficace lorsqu'utilisée pour des échantillons formés de données discrètes ou arrondies. L'optimisation de l'expression (4.22) peut conduire à un paramètre de lissage nul dans de pareils cas. Par ailleurs, il est aussi intéressant de mentionner que la méthode des moindres carrés permet d'obtenir de bons résultats pour l'estimation des quantiles d'étiage (Adamowski; 1996).

5.5.6 Méthode des moindres carrés à fenêtre variable

Cette version adaptée pour fenêtre variable de la méthode des moindres carrés conduit à la pire des estimations parmi les méthodes considérées dans cette étude (figure 5.14 (a)). La méthode provoque une surestimation relativement importante, un biais élevé et une variance nettement supérieure à celle obtenue avec les autres méthodes. En considérant un paramètre de lissage variable pour chacune des observations pour le calcul de la fonction de densité non paramétrique servant dans l'estimation de l'*ISE*, le minimum de la fonction (4.22) est généralement atteint pour une valeur élevée du paramètre de lissage, c'est à dire des valeurs des paramètres locaux $a_k d_{ki}$ très grandes. Les paramètres de lissage locaux en plus d'être relativement variables d'un échantillon à l'autre, le sont à l'intérieur d'un même échantillon. Comme pour la méthode de Breiman *et al.* discutée précédemment, une fenêtre nettement plus grande est accordée aux valeurs extrêmes de l'échantillon, les faibles et les grands débits.

Avec de tels résultats, il est clair que cette méthode n'est pas adéquate pour l'estimation de quantiles de crue. De surcroît, elle n'est pas nécessairement très pratique d'utilisation, puisque l'on doit effectuer une optimisation à deux paramètres et que la fonction objectif n'est pas vraiment sensible à la variation de ces paramètres. Par contre, l'estimation des quantiles est quant à elle, relativement sensible aux paramètres a_k et k.

Pour ce qui est de la procédure adaptative proposée par Abramson (1982), une étude effectuée par Sharma *et al.* (1998) a démontré qu'elle ne permettait d'améliorer les résultats lorsque considérée avec la méthode des moindres carrés ou avec la méthode du maximum de vraisemblance (avec validation croisée). Il est toutefois important de noter que la comparaison de Sharma *et al.* (1998) a été effectuée sur la base de l'estimation de la fonction de densité et non sur l'estimation de quantiles. Il ne semble donc pas vraiment avantageux de considérer une estimation locale du paramètre de lissage puisqu'elle ne conduit pas vraiment à de meilleurs résultats qu'une estimation globale de h.

5.5.7 Méthode *plug-in* de Gasser *et al.*

A priori, la méthode *plug-in* de Gasser *et al.* constituait une technique intéressante pour estimer le paramètre de lissage puisqu'elle permettait d'effectuer l'estimation directement à partir de la fonction quantile plutôt qu'à partir de la fonction de densité ou de la fonction de répartition comme c'est le cas des autres méthodes généralement considérées dans la littérature. De plus, l'utilisation de noyaux limites semblait représenter un moyen de réduire le biais dans les extrémités de la distribution, c'est à dire pour les quantiles extrapolés. Mais, les résultats obtenus par le biais de cette procédure sont décevants, puisqu'elle ne permet pas d'obtenir une estimation raisonnable du paramètre de lissage. Les paramètres optimaux obtenus pour les divers échantillons de l'étude de simulation sont tous trop faibles pour qu'il ne soit possible d'extrapoler au delà des observations.

Comme son nom l'indique, la méthode *plug-in* est une méthode automatique qui permet d'estimer le paramètre de lissage sans avoir à optimiser une certaine fonction objectif, mais plutôt en effectuant 11 estimations récursives de la valeur de h théorique qui minimise l'erreur quadratique moyenne de la fonction quantile $MSE_{x(p)}$. De cette façon, la valeur finale de h obtenue après la 11^{e} itération n'est pas nécessairement une valeur raisonnable pour l'estimation de la fonction quantile, puisqu'elle n'est le résultat d'un calcul itératif et non de l'ajustement de la fonction estimée sous la forme de minimisation d'une certaine fonction d'erreur. Cette approche mériterait d'être étudiée de manière plus approfondie dans une étude future.

5.6 Méthode à fenêtre fixe et à fenêtre variable

Dans cette section, une brève comparaison des méthodes à fenêtre fixe et à fenêtre variable est effectuée. Dans un premier temps, la méthode des moindres carrés avec validation

croisée à fenêtre fixe est comparée à son équivalent pour fenêtre variable. Par la suite, des conclusions générales sont tirées sur les deux types de méthodes.

5.6.1 Méthode des moindres carrés à fenêtre fixe et à fenêtre variable

À la section 4.6.3, on avait supposé que les méthodes à fenêtre fixe de calcul du paramètre de lissage adaptées pour la méthode à fenêtre variable pouvaient permettre d'améliorer l'estimation. Les résultats obtenus en adaptant la méthode des moindres carrés avec validation croisée démontrent qu'il n'est pas nécessairement avantageux d'utiliser des méthodes à fenêtre fixe dans un contexte à fenêtre variable. Les remarques faites dans cette section peuvent aussi permettre de comprendre la différence des résultats obtenus avec les méthodes à fenêtre fixe et à fenêtre variable en général.

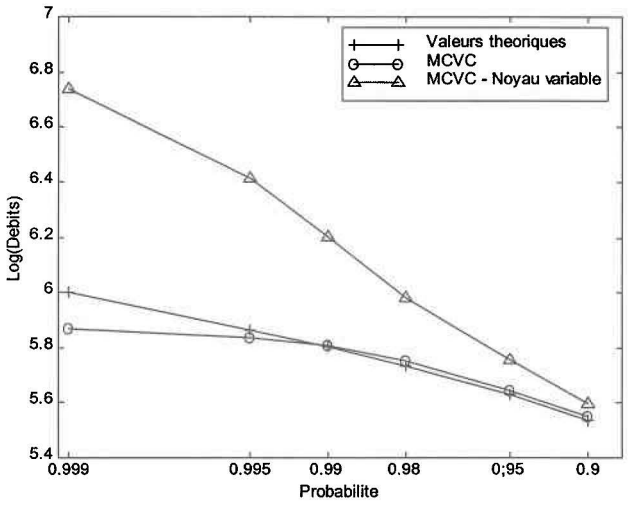


Figure 5.18 : Comparaison des résultats obtenus avec la méthode des moindres carrés avec validation croisée pour fenêtre fixe et fenêtre variable, en considérant le noyau Epanechnikov pour n=50.

La figure 5.18 montre que la méthode des moindres carrés avec validation croisée pour fenêtre variable conduit à une surestimation des quantiles d'ordre supérieur. Malgré le fait

que les données sont présentées avec une échelle logarithmique, la différence entre les quantiles théoriques et les quantiles estimés à partir de la méthode *MCVC* pour fenêtre variable est relativement grande, ce qui illustre à quel point la surestimation est importante.

On a vu précédemment que pour la méthode à fenêtre variable, les paramètres de lissage locaux étaient beaucoup trop variables à l'intérieur d'un échantillon, accordant ainsi un poids considérable aux observations extrêmes. La figure 5.19 donne un exemple de la répartition des paramètres locaux pour deux échantillons sélectionnés au sein du groupe d'échantillons de taille $n = 50$. La figure 5.19 (a) montre un exemple d'échantillon pour lequel les paramètres locaux sont relativement élevés, surtout pour les faibles et les forts débits. Un poids trop grand est ainsi accordé aux valeurs extrêmes par rapport aux autres. Il en résulte la surestimation des quantiles comme on peut le constater sur la figure 5.19 (c). Par contre, la méthode à fenêtre fixe, avec un paramètre global raisonnable, a permis d'obtenir une estimation des quantiles relativement bonne. Sur la figure 5.19 (b), on remarque aussi que les paramètres locaux sont plus importants pour les débits élevés que pour le reste de l'échantillon. Ces valeurs représentent tout de même une fenêtre de taille raisonnable. On remarque ensuite sur la figure 5.19 (d) que l'estimation en est grandement améliorée, les résultats étant comparables à ceux obtenus avec la méthode à fenêtre fixe. Cet exemple illustre bien à quel point la valeur des paramètres de lissage locaux est variable et comment cette variabilité influe sur la qualité de l'estimation.

Il apparaît maintenant évident que la méthode des moindres carrés utilisée pour l'estimation de h dans un contexte de fenêtre fixe, n'est pas applicable dans le cas à fenêtre variable. Il n'est cependant pas possible de généraliser cette conclusion à toutes les autres méthodes à fenêtre fixe qui pourraient être adaptées pour être utilisables dans un contexte à fenêtre variable, puisque l'expérience a été effectuée que pour une seule méthode. Mais il y a tout lieu de croire que la variabilité des paramètres locaux puisse être responsable des piètres résultats obtenus avec les méthodes à fenêtre variable.

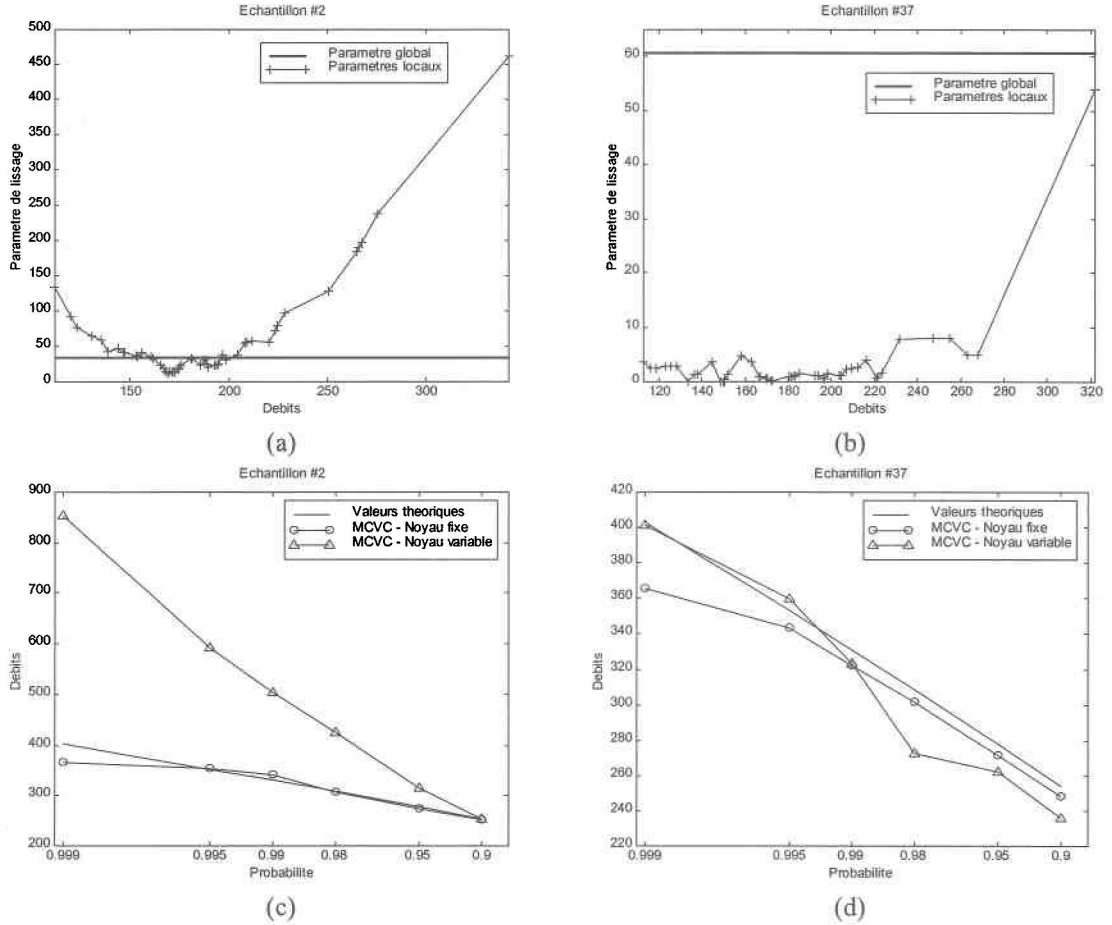


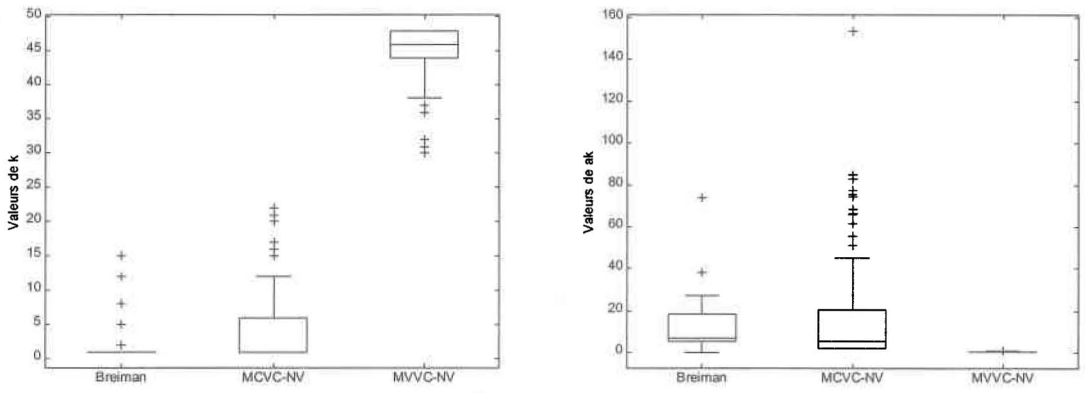
Figure 5.19 : Exemple provenant de deux échantillons de taille 50, illustrant l'influence des paramètres de lissage sur l'estimation pour les méthodes des moindres carrés à fenêtre fixe et à fenêtre variable.

5.6.2 Remarques générales sur les deux types de méthodes

Lors de la comparaison des méthodes de calcul du paramètre de lissage, effectuée à la section 5.5, on avait déjà remarqué que les méthodes à fenêtre fixe procuraient de meilleurs résultats que les méthodes à fenêtre variable, que ce soit par rapport à la qualité de l'estimation, au biais ou au *RMSE*. Seule la méthode du maximum de vraisemblance a conduit à des résultats acceptables dans le cas des méthodes à fenêtre variable.

Une hypothèse qui pourrait expliquer la surestimation obtenue avec les méthodes à fenêtre variable, provient du fait que les débits les plus élevés sont généralement relativement éloignés des autres observations. Ainsi, la distance à leur k^e voisin le plus proche est assez élevée par rapport à ce que l'on observe pour les observations situées dans la partie centrale de l'échantillon, qui sont relativement rapprochées les unes des autres. Lorsque l'on désire extrapoler la valeur de la fonction de répartition pour un débit critique x_c, les dernières observations de l'échantillon sont d'une grande importance pour l'estimation. Le poids d'une observation dans l'estimation de x_c est inversement proportionnel à la distance la séparant de ce débit critique. Par conséquent, la distance séparant les dernières observations et le débit critique étant relativement faible par rapport aux paramètres de lissage locaux (t près de zéro), la faible valeur du noyau intégré K_I résulte en une faible estimation de la fonction de répartition à ce point. Ainsi, la fonction de répartition estimée se rapproche relativement moins rapidement de la valeur maximale de $F = 1$ avec l'augmentation du débit, que dans le cas des méthodes où le paramètre de lissage est fixe. Pour cette raison, les quantiles estimés avec les méthodes à fenêtre variable sont plus élevés que ceux estimés avec les méthodes à fenêtre fixe.

Parmi les méthodes à fenêtre variable, la méthode du maximum de vraisemblance procure des résultats presque comparables aux méthodes à fenêtre fixe (figure 5.14). La surestimation est beaucoup moins importante que dans le cas des autres méthodes à fenêtre variable. Les paramètres locaux obtenus à partir de cette méthode sont plus élevés que les paramètres obtenus à l'aide des autres méthodes. La figure 5.20 présente une comparaison des paramètres optimaux obtenus à l'aide des trois méthodes à fenêtre variables pour le groupe d'échantillons de taille $n = 50$. On remarque (figure 5.20 (a)) que les voisins les plus proches considérés sont beaucoup plus éloignés dans le cas de la méthode du maximum de vraisemblance que pour la méthode de Breiman et la méthode des moindres carrés. Par conséquent, la valeur du paramètre a_k est beaucoup plus faible pour la méthode du maximum de vraisemblance ($a_k < 1$), étant donné les distance d_{ki} élevées (figure 5.20 (b)).



(a) (b)

Figure 5.20 : Comparaison des valeurs *k* (voisin le plus proche) et des valeurs du paramètre a_k obtenus à partir des trois méthodes à fenêtre variable.

Finalement, la figure 5.21 montre la répartition des coefficients de variation ($\sigma_{d_{ki}} / \mu_{d_{ki}}$) pour les 100 échantillons du groupe de taille $n = 50$ et ce, pour chacune des trois méthodes à fenêtre variable. On remarque que les coefficients de variation obtenus pour la méthode du maximum de vraisemblance sont significativement plus faibles que ceux obtenus avec les deux autres méthodes. La répartition des paramètres de lissage locaux sur l'échantillon est donc moins variable pour la méthode du maximum de vraisemblance, les distances au k^e voisin le plus proche étant semblables pour tout l'échantillon puisque la valeur de k est élevée.

Comme la méthode du maximum de vraisemblance procure une estimation relativement plus précise que les deux autres méthodes, il est convenable de conclure qu'il est préférable que la répartition des paramètres de lissage sur l'échantillon soit relativement uniforme.

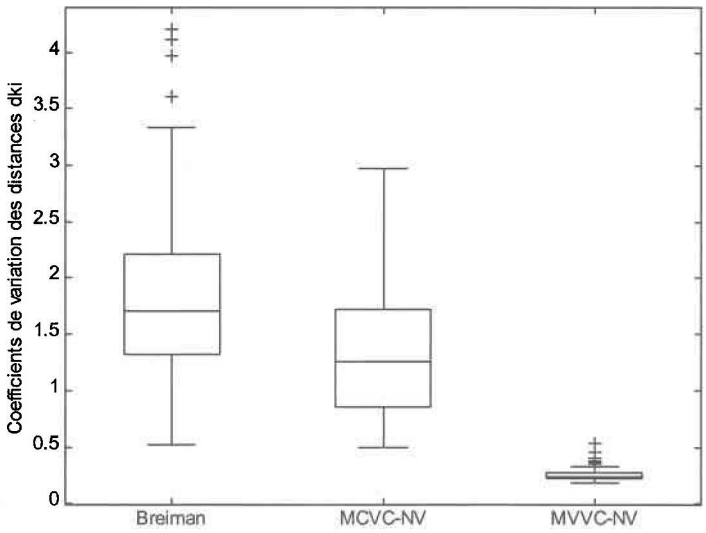


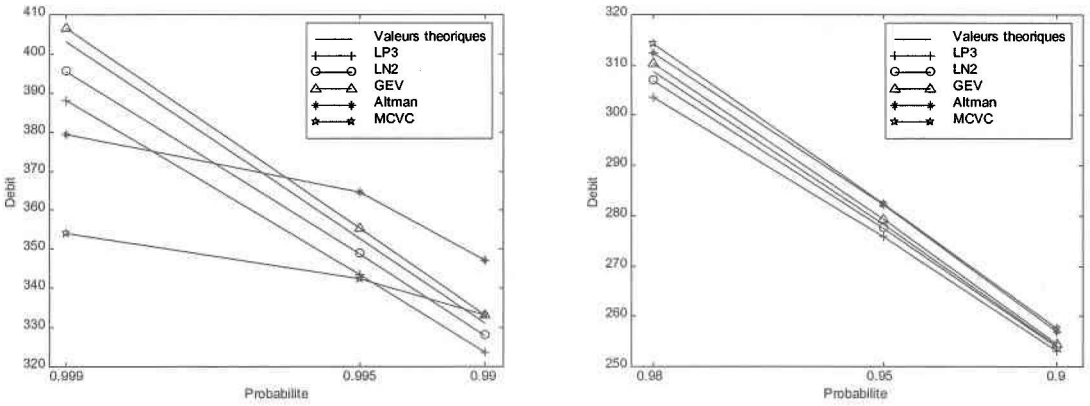
Figure 5.21 : Comparaison de la répartition sur les 100 échantillons des coefficients de variation des distances d_{ki} pour chacune des méthodes à fenêtre variable.

5.7 Méthode des noyaux et méthodes paramétriques

La comparaison de la méthode des noyaux avec les méthodes paramétriques a été effectuée en considérant les deux méthodes non paramétriques qui ont été les plus efficaces dans notre étude de simulation (section 5.5), en l'occurrence la méthode de Altman et Léger et la méthode des moindres carrés avec validation croisée. Les résultats obtenus à l'aide de ces deux méthodes sont comparés aux résultats des ajustements des distributions LP3, LN2 et GEV. A priori, puisque les données sont simulées à partir de la LP3, on s'attendrait à ce que la précision obtenue avec l'ajustement des distributions statistiques soit supérieure à celle obtenue avec les méthodes non paramétriques.

La figure 5.22 présente en deux parties la comparaison de l'estimation des quantiles avec les deux méthodes non paramétriques et les trois distributions paramétriques pour le groupe d'échantillon de taille n = 50. Comme il fallait s'y attendre, les méthodes paramétriques procurent une meilleure estimation que les méthodes non paramétriques, sauf pour une

période de retour de 100 ans où la méthode des moindres carrés et la GEV procurent l'estimation la plus précise. Pour les faibles périodes de retour (figure 5.22 (b)), la qualité de l'estimation est comparable pour toutes les méthodes. Par contre, pour des périodes de retour de 200 et de 1000 ans, les méthodes paramétriques procurent une estimation plus fiable. Il est tout de même surprenant de remarquer que l'ajustement de la LP3 conduit à des estimations moins précises que les deux autres lois statistiques.



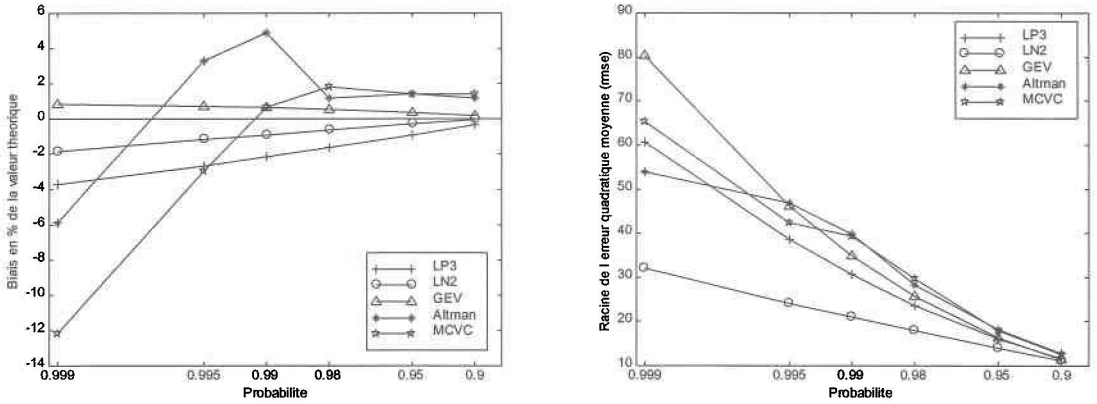
(a) (b)

Figure 5.22 : Comparaison de l'estimation pour les trois distributions paramétriques, la méthode de Altman et la méthode des moindres carrés (noyau Epanechnikov et n = 50). (a) T = 100, 200 et 1000 ans ; (b) T = 10, 20 et 50 ans.

À la figure 5.23 on remarque que le biais est minimisé avec la loi GEV et que les méthodes paramétriques conduisent toutes à une estimation où le biais est plus faible que celui obtenu à l'aide de la méthode des noyaux. Quant à la variance, elle est minimale pour la LN2 et relativement comparable pour les autres méthodes.

À la section 5.4 on avait remarqué que la taille de l'échantillon semblait avoir une plus grande importance sur la qualité de l'estimation dans le cas des méthodes paramétriques que dans le cas de la méthode des noyaux. On avait alors comparé l'ajustement de la GEV à la méthode du maximum de vraisemblance avec validation croisée pour des échantillons de taille 10 et de taille 100. La figure 5.24 permet de confirmer que l'estimation non

paramétrique est moins influencée par la faible taille d'échantillon que l'ajustement de la GEV et de la LP3. Par exemple, le biais de l'estimation provenant de la GEV et de la LP3 est plus élevé que celui obtenu avec la méthode non paramétrique de Altman et Léger pour une taille d'échantillon de $n = 10$ alors que c'est le contraire pour $n = 50$. De même, la variance de l'estimation provenant de la GEV est supérieure à celle des quatre autres méthodes qui elles, sont relativement comparables.



(a) (b)

Figure 5.23 : Comparaison du biais et du *RMSE* pour les trois distributions, la méthode de Altman et Léger et la méthode des moindres carrés ($n = 50$).

Étant donné que les données ont été simulées à partir d'une distribution comme la LP3, il est convenable que les distributions statistiques traditionnelles procurent une meilleure estimation. Mais dans la nature, les observations ne sont généralement pas distribuées parfaitement selon une de ces lois. Il est fort probable que l'on puisse retrouver un échantillon qui provient d'une distribution multimodale ou d'un mélange de distributions. Dans de pareils cas, l'ajustement paramétrique est relativement moins efficace puisqu'il est impossible de connaître le type de distribution selon laquelle la population est distribuée. Les méthodes non paramétriques constituent alors une alternative intéressante. Dans le cadre de ce travail, il n'a pas été convenu d'effectuer des comparaisons à partir d'échantillons provenant de telles populations. En simulant des données à partir d'un mélange de distributions, il est probable que la méthode des noyaux conduirait à de

meilleurs résultats puisqu'elle a permis d'obtenir des résultats comparables à l'ajustement de la LP3, la LN2 et la GEV dans le cas de données provenant de ce type de distribution.

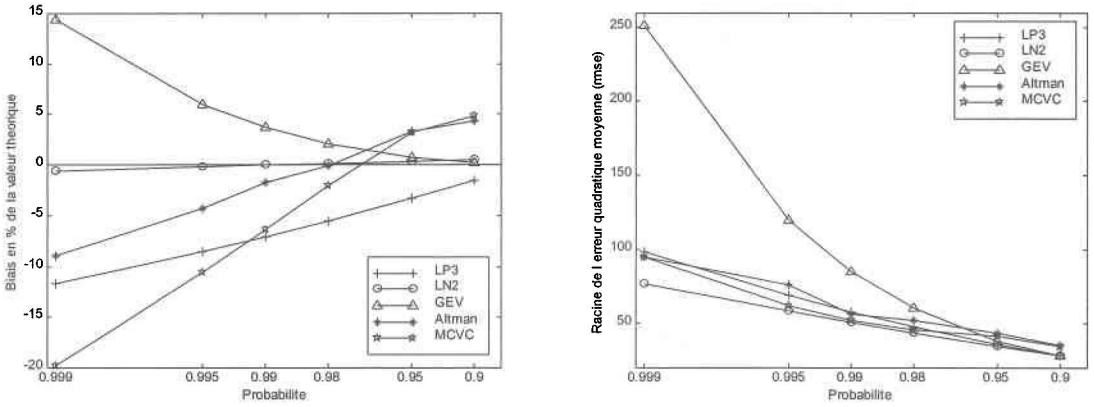


Figure 5.24 : Comparaison du biais et du *RMSE* pour les trois distributions, la méthode de Altman et Léger et la méthode des moindres carrés (n = 10).

6. CONCLUSION

L'étude des probabilités d'occurrence des crues extrêmes s'effectue généralement à l'aide de méthodes paramétriques. En ajustant une certaine distribution statistique aux données d'un échantillon, on peut interpoler ou extrapoler sur la courbe, les valeurs de débits correspondant aux probabilités d'intérêt. Le problème avec les méthodes paramétriques, est que l'on doit faire l'hypothèse d'une distribution. Si on connaît avec certitude la forme de la distribution théorique, alors les méthodes d'ajustement paramétrique sont relativement efficaces, puisque l'on peut estimer les paramètres de la distribution et en déduire les quantiles d'intérêt. Toutefois, il est rare que les données d'un cours d'eau soient distribuées exactement selon une des distributions disponible en analyse de fréquence de crue, elles peuvent aussi provenir d'un mélange de distributions ; il est d'autant plus difficile de connaître la distribution exacte qui caractérise la population. Pour ces raisons, les méthodes non paramétriques sont d'un intérêt grandissant. Elles possèdent la qualité de ne pas nécessiter d'hypothèses sur la distribution de la population. Ces méthodes sont beaucoup plus flexibles que les méthodes paramétriques traditionnelles puisque la densité non paramétrique n'a pas de forme particulière étant donnée qu'elle est déterminée directement à partir des données de l'échantillon. Cette flexibilité peut devenir relativement importante dans les extrémités de la distribution où il est parfois difficile d'ajuster une loi de probabilité, celle-ci étant essentiellement ajustée sur la partie principale de la distribution. Toutefois, le degré d'extrapolation peut être limité étant donné le caractère empirique des méthodes non paramétriques.

La méthode des noyaux constitue une alternative intéressante aux méthodes paramétriques traditionnelles pour l'estimation de fréquence de crue. Elle est relativement simple d'utilisation puisqu'il ne s'agit que de sélectionner une certaine fonction noyau qui, en servant de poids pour chacune des observations, permet d'estimer la fonction de densité non

paramétrique. On doit aussi calculer un paramètre de lissage qui détermine le degré d'influence des observations pour l'estimation.

La méthode des noyaux a été comparée à certaines méthodes paramétriques d'ajustement de lois, la log-Pearson type 3 (LP3) dont les paramètres ont été estimés avec la méthode des moments classique, la log-normale à deux paramètres (LN2) avec ses paramètres estimés par la méthode du maximum de vraisemblance et la loi généralisée des valeurs extrêmes (GEV) dont les paramètres ont été estimés par la méthode des moments pondérés. Les données synthétiques qui ont servis aux diverses comparaisons effectuées dans cette étude ont été simulées à partir de la loi log-Pearson type 3. Quatre tailles d'échantillons ont été considérées (n = 10, 25, 50 et 100) avec 100 réplications pour chacune d'elles. Comme les données proviennent d'une population distribuée selon la loi LP3, l'estimation est effectuée dans les conditions les plus favorables pour les méthodes paramétriques. Il est donc logique que les méthodes d'ajustement paramétriques procurent une meilleure estimation que les méthodes non paramétriques. Toutefois, certains résultats obtenus avec la méthode des noyaux, sont comparables à ceux obtenus avec les méthodes paramétriques. En effectuant le même type de comparaison sur des données provenant d'une distribution multimodale ou bien du mélange de plusieurs distributions, il est fort probable que la méthode des noyaux aurait gagné de la précision par rapport aux trois distributions considérées.

Selon certains travaux effectués sur la méthode des noyaux, le choix du type de noyau à considérer n'aurait que peu d'importance sur la qualité de l'estimation. Il aurait été convenu d'utiliser le noyau Epanechnikov, celui qui minimise l'erreur quadratique moyenne intégrée (*IMSE*), sans crainte de perte de précision. Toutefois, comme la plupart des travaux sont effectués dans un contexte où la méthode des noyaux est utilisée sous forme de régression non paramétrique (Adamowski et Feluch; 1990, Gingras *et al.*; 1995, Härdle; 1990, Lettenmaier; 1984, Nychka; 1991, Smith; 1991), plus précisément pour l'interpolation, on peut se demander si le type de noyau ne pourrait pas avoir une certaine importance dans un

contexte d'extrapolation. En considérant six noyaux différents, une comparaison du type de noyau a été effectuée sur la base du domaine de définition et de la symétrie. Il apparaît clair que même dans un contexte d'extrapolation, l'influence du type de noyau est bien moindre que le choix du paramètre de lissage. Parmi les six noyaux considérés, les noyaux normal, Epanechnikov, biweigth et EV1 procurent des résultats similaires alors que les noyaux Cauchy et rectangulaire ne permettent pas d'obtenir une estimation convenable des quantiles de crue. Le noyau EV1 est particulièrement efficace pour les grandes périodes de retour.

Quant au paramètre de lissage, il existe plusieurs méthodes permettant d'en estimer la valeur optimale. Quelques-unes des principales méthodes proposées dans la littérature ont été étudiées dans ce travail et comparées entre elles dans le but de retenir celle qui est la plus efficace dans un contexte d'estimation de quantiles de crue. On peut regrouper les méthodes de calcul du paramètre de lissage en trois grandes catégories :

- Les méthodes qui s'appuient sur l'estimation de la fonction de densité ;
- Les méthodes qui s'appuient sur l'estimation de la fonction de répartition ;
- Les méthodes qui calculent le paramètre de lissage directement à partir de l'estimation des quantiles.

Dans cette étude, chacune des catégories est représentée par au moins une méthode. La méthode de Altman et Léger (fonction de répartition) et la méthode des moindres carrés (fonction de densité) sont les méthodes qui ont mené aux estimations les plus fiables.

Il existe entre autres deux façons de considérer la méthode des noyaux. D'abord, on peut considérer le paramètre de lissage comme étant constant pour toutes les observations de l'échantillon d'intérêt. On peut aussi estimer le paramètre de lissage localement, en adaptant la valeur de ce dernier selon la position de l'observation dans l'échantillon. La méthode à

noyau variable fait en sorte que le paramètre de lissage est estimé localement à partir de la distance de chacune des observations à leur k^e voisin le plus proche. Les résultats des comparaisons ont permis de montrer que la méthode à noyau variable n'apportait pas vraiment d'amélioration par rapport à la méthode à noyau fixe pour l'estimation de quantiles de crue. Les trois méthodes à noyau variable étudiées ont conduit à une surestimation des six quantiles considérés (T = 10, 20, 50, 100, 200 et 1000 ans). Parmi ces méthodes seule celle du maximum de vraisemblance a permis d'obtenir une estimation raisonnable.

Ce travail constitue une analyse approfondie des propriétés de la méthode des noyaux. Il a aussi permis d'identifier deux méthodes de calcul du paramètre de lissage qui permettent d'obtenir une bonne estimation des quantiles de crue, la méthode de Altman et Léger et la méthode des moindres carrés avec validation croisée. Dans le cadre de ce travail, la méthode de Altman et Léger a due être adaptée pour être utilisable dans un contexte d'estimation de quantiles de crue. Par ailleurs, les remarques faites sur le critère d'Adamowski ont permis de démontrer que cette méthode conduisait à des valeurs de paramètre de lissage trop faibles pour qu'il y ait lissage. La méthode d'Adamowski, bien que fréquemment citée dans la littérature, devra être abandonnée.

6.1 Études futures

Comme on l'a mentionné précédemment, une comparaison pourrait être effectuée entre la méthode des noyaux et les distributions paramétriques à partir de données synthétiques provenant d'un mélange de distributions ou bien à partir de distributions à plusieurs paramètres. À ce sujet, la distribution de Wakeby pourrait être intéressante puisqu'elle possède cinq paramètres (Houghton; 1978). Mais comme les paramètres de cette distribution sont difficiles à estimer, il serait peut-être préférable de considérer un mélange de lois normale. Le même type de comparaison pourrait aussi être effectuée avec des

données réelles, à partir desquelles une plus grande quantité de données pourrait être générée par la méthode du bootstrap. Il serait aussi intéressant de considérer des modèles semi-paramétriques, où la partie principale de la distribution est ajustée avec une loi statistique et où la méthode des noyaux sert à estimer l'extrémité de droite, à partir du rang où l'ajustement paramétrique commence à être biaisée. Il reste aussi la possibilité de considérer un modèle quadratique pour estimer les extrémités (*quadratic tail model*) (Moon *et al.*; 1993). Il serait aussi intéressant d'étudier les méthodes qui permettent d'incorporer de l'information historique, lesquelles sont discutés, entre autres, par Adamowski et Feluch (1990) et par Guo (1991). Finalement, malgré les résultats décevants obtenus avec la méthode *plug-in* de Gasser *et al.* (1991) on peut toutefois dire que cette approche mérite d'être étudiée plus en profondeur dans une étude future. Beaucoup de travaux ont été effectués sur la méthode des noyaux et il reste encore beaucoup à faire pour en favoriser l'utilisation en hydrologie.

7. RÉFÉRENCES

Abramson, I. S. (1982). On bandwidth variation in kernel estimates - a square root law. Ann. Stat., 10(4): 1217-1223.

Adamowski, K. (1981). Plotting formula for flood frequency. Water Resour. Bull., 17(2): 197-202.

Adamowski, K. (1985). Nonparametric kernel estimation of flood frequencies. Water Resour. Res., 21(11): 1585-1590.

Adamowski, K. (1989). A Monte Carlo comparison of parametric and nonparametric estimation of flood frequencies. J. Hydrol., 108: 295-308.

Adamowski, K. (1996). Nonparametric estimation of low-flow frequencies. J. Hydr. Engrg., ASCE, 122(1): 46-49.

Adamowski, K. et W. Feluch (1983). Application of pattern analysis to flood frequency determination. American Geophysical Union, EOS Transactions, 64(45): 705.

Adamowski, K. et W. Feluch (1990). Nonparametric flood-frequency analysis with historical information. J. Hydr. Engrg., ASCE, 116(8): 1035-1047.

Adamowski, K. et W. Feluch (1991). Application of nonparametric regression to groundwater level prediction. Can. J. Civ. Eng., 18: 600-606.

Altman, N. et C. Léger (1995). Bandwidth selection for kernel distribution function estimation. J. Statist Plann. Inference, 46: 195-214.

Azzalini, A. (1981). A note on the estimation of a distribution function and quantiles by a kernel method. Biometrika, 68(1): 326-328.

Benson, M. A. (1968). Uniform flood frequency estimating methods for federal agencies. Water Resour. Res., 4(5): 891-908.

Bobée, B. (1975). The Log Pearson type 3 distribution and its application in hydrology. Water Resour. Res., 11(5): 681-689.

Bobée, B. et F. Ashkar (1991). *The Gamma Family and Derived Distributions Applied in Hydrology*. Water Resources Publications. Littleton, Colorado, 203 pp.

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. Biometrika, 71(2): 353-360.

Breiman, L., W. Meisel et E. Purcell (1977). Variable kernel estimates of multivariate densities. Technometrics, 19(2): 135-144.

Cacoullos, T. (1966). Estimation of a multivariate density. Annals of the Institute of Statistical Mathematics, 18: 178-189.

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Springer-Verlag, New-York, 843pp.

Dooge, J. C. I. (1986). Looking for hydrologic laws. Water Resour. Res., 22(9): 46-58.

Efron, B. et G. Gong (1983). A leisurely look at the bootstrap, the Jackknife and cross-validation. The American Statistician, 37(1): 36-48.

Epanechnikov, V. A. (1969). Nonparametric estimation of a multidimensional probability density. Theory of Probability and its Applications, 14 : 153-158.

Fan, J., P. Hall, M. A. Martin et P. Patil (1996). On local smoothing of nonparametric curve estimator. J. Am. Stat. Assoc., 91(433): 258-266.

Faucher, D., T. B. M. J. Ouarda et B.Bobée (1997). Revue bibliographique des tests de stationnarité. Chaire industrielle en hydrologie statistique. INRS-Eau rapport de recherche no R-499, 66 pp.

Gasser, T. et H.-G. Müller (1984). Estimating regression fonctions and their derivatives by the kernel method. Scand. J. Statist., 11 : 171-185.

Gasser, T., L. Sroka et C. Jennen-Steinmetz (1986). Residual variance and residual pattern in nonlinear regression. Biometrika, 73(3): 625-633.

Gasser, T., A. Kneip et W. Kohler (1991). A flexible and fast method for automatic smoothing. J. Am. Stat. Assoc., 86(415): 643-652.

Gingras, D., M. Alvo et K. Adamowski (1995). Regional flood relationships by nonparametric regression. Nordic Hydrol., 26: 73-90.

Greenwood, J. A., J. M. Landwehr, N. C. Matalas et J. R. Wallis (1979). Probability weighted moments : Definition and relation to parameters of several distributions expressable in inverse form. Water Resour. Res., 15(5): 1049-1054.

Guo, S. L. (1991). Nonparametric variable kernel estimation with historical floods and paleoflood information. Water Resour. Res., 27(1): 91-98.

Hall, M.J. (1984). *Urban hydrology*. Elsevier, Barking, 299 pp.

Hall, P. et J. S. Marron (1987). Estimation of integrated squared density derivatives. Stat. Probab. Lett., 6 : 109-115.

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press, New-York, 333 pp.

Härdle, W (1991). *Smoothing techniques : with implementation in S*. New York : Springer-Verlag, 271 pp.

Houghton, J. C. (1978). Birth of a parent : The Wakeby distribution for modeling flood flows. Water Resour. Res., 14(6): 1105-1109.

HYFRAN (1998). Logiciel d'analyse fréquentielle hydrologique. Chaire industrielle en hydrologie statistique, INRS-Eau.

Izenman, A. J. (1991). Recent developments in nonparametric density estimation. J. Am. Stat. Assoc., 86(413): 205-224.

Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) of meteorological elements. Quaterly journal of the Royal Meteorological Society, 81: 158-171.

Johnk, M. D. (1964). Erzeugung von betaverteilten and gammaverteilen zufallaszahlen. Metrika, 8(1): 5-15.

Jones, M. C. (1990). The performance of kernel density functions in kernel distribution function estimation. Stat. Probab. Lett., 9: 129-132.

Klemes, V. (1986). Dilettantism in hydrology: transition or destiny? Water Resour. Res., 22(9): 177-188.

Lall, U. (1995). Recent advances in nonparametric function estimation: Hydrologic applications. Reviews of geophysics, supplement, 1093-1102, U.S. National Report to International Union of Geodesy and Geophysics 1991-1994.

Lall, U., Y.-I. Moon et K. Bosworth (1993). Kernel flood frequency estimators: bandwitdh selection and kernel choice. Water Resour. Res., 29(4): 1003-1015.

Lejeune, M. et P. Sarda (1992). Smooth estimators of distribution and density functions. Comput. Statist. Data Anal., 14 : 457-471.

Lettenmaier, D. P. (1984). Limitations on seasonal snowmelt forecast accuracy. J. Water Resour. Plann Manage., 110(3): 255-269.

Marron, J. S. (1988). Automatic smoothing parameter selection : a survey. Empirical Economics, 13: 187-208.

Moon, Y.-I., U. Lall et K. Bosworth (1993). A comparison of tail probability estimators for flood frequency analysis. J. Hydrol., 151: 343-363.

Moon, Y.-I. et U. Lall (1994). Kernel quantile function estimator for flood frequency analysis. Water Resour. Res., 30(11): 3095-3103.

Müller, H.-G. (1991). Smooth optimum kernel estimators near endpoints. Biometrika, 78(3): 521-530.

Nychka, D. (1991). Choosing a range for the amount of smoothing in nonparametric regression. J. Am. Stat. Assoc., 86(415): 653-664.

Ondo, J.C., T.B.M.J Ouarda et B.Bobée (1997). Revue bibliographique des tests de d'homogénéité et d'indépendance. Chaire industrielle en hydrologie statistique. INRS-Eau rapport de recherche no R-500, 78 pp.

Park, B. U. et J. S. Marron (1990). Comparison of data-driven bandwidth selectors. J. Am. Stat. Assoc., 85(409): 66-72.

Perreault, L., B. Bobée et P. Legendre (1994). Rapport général du logiciel *Ajuste II*: Théorie et application. INRS-Eau, rapport de recherche no. R-421, 92 p.

Rao, P.B.L. (1983). *Nonparametric Functionnal Estimation*. Academic Press, New York, NY.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. Ann. Math. Statist., 27: 832-837.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. Scand. J. Statist, 9: 65-78.

Sarda, P. (1993). Smoothing parameter selection for smooth distributions functions. J. Statist Plann. Inference, 35: 65-75.

Sharma, A., U. Lall et D. G. Tarboton (1998). Kernel bandwidth selection for a first order nonparametric streamflow simulation model. Stochastic Hydrol. Hydraul., 12: 33-52.

Sharma, A., D. G. Tarboton et U. Lall (1997). Streamflow simulation: a nonparametric approach. Water Resour. Res., 33(2): 291-308.

Sheather, S. J. et J. S. Marron (1990). Kernel quantile estimator. J. Am. Stat. Assoc., 85(410): 410-416.

Sheather, S. J. et M. C. Jones (1991). A reliable data-based bandwidth selection method for kernel density estimation. J. R. Statist. Soc. B, 53(3): 683-690.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, New York, 175 pp.

Smith, J.A. (1991). Long-range streamflow forcasting using nonparametric regression. Water Res. Bull., 27(1): 39-46.

Staniswalis, J.G. (1989). Local bandwidth selection for kernel estimates. J. Am. Stat. Assoc., 84(405): 284-288.

Stone, C.J. (1984). An asymptotically optimal window selection rule for kernel density estimates. Ann. Stat., 12: 1285-1287.

Whittle, P. (1958). On the smoothing of probability density functions . J. Roy. Statist. Soc. B, 20: 334-343.

Yakowitz, S.J. (1983). Some « model-free » techniques for flood frequency analysis. American Geophysical Union, EOS Transactions, 64(45): 706.

Annexe A

Développements théoriques

A) Développement de la fonction d'erreur quadratique moyenne intégrée.

$$\begin{aligned} IMSE &= \int E\left[\left(\hat{f} - E\hat{f} + E\hat{f} - f\right)^2\right] \\ &= \int E\left[\left(\hat{f} - E\hat{f}\right)^2 + 2\left(\hat{f} - E\hat{f}\right)\left(E\hat{f} - f\right) + \left(E\hat{f} - f\right)^2\right] \\ &= \int E\left[\left(\hat{f} - E\hat{f}\right)^2\right] + 2E\left[\hat{f} - E\hat{f}\right]\left(E\hat{f} - f\right) + \left(E\hat{f} - f\right)^2 \\ &= \int E\left[\left(\hat{f} - E\hat{f}\right)^2\right] + \left(E\hat{f} - f\right)^2 \\ &= \int \mathrm{var}\left(\hat{f}\right) + biais\left(\hat{f}\right)^2 \end{aligned}$$

B) Le critère d'Adamowski mène à des paramètres de lissage très faibles. Il en résulte une estimation de la fonction de répartition qui tend vers une formule de probabilité empirique. La démonstration est effectuée pour le noyau rectangulaire, le noyau Cauchy et le noyau normal. On a la fonction de répartition non paramétrique suivante :

$$\hat{F}\left(x_j\right) = \frac{1}{n}\sum_{i=1}^{n}\left\{K_I\left(\frac{x_j - x_i}{h}\right) I_1\left(\left|\frac{x_j - x_i}{h}\right| \le 1\right) + I_2\left(\frac{x_j - x_i}{h} > 1\right)\right\}$$

où K_I est le noyau intégré, $I_1(A)$ et $I_2(B)$ sont des variables dichotomiques prenant les valeurs 1 et 0 si les énoncés A et B sont respectés ou non. En remplaçant $\left(x_j - x_i\right)/h$ par t pour alléger la notation, on a :

$$\hat{F}\left(x_j\right) = \frac{1}{n}\sum_{i=1}^{n}\left\{K_I(t)\; I_1(|t| \le 1) + I_2(t > 1)\right\}$$

Pour un noyau rectangulaire, on a :

$$\hat{F}\left(x_j\right) = \frac{1}{nh}\sum_{i=1}^{n}\left\{\frac{1}{2}\left(x_j - x_i + h\right) I_1(|t| \le 1) + h\, I_2(t > 1)\right\}$$

Lorsque $h < d_m$ (où d_m représente le plus petit des écarts des observations prise deux-à-deux), seule l'observation x_j elle-même se trouve dans l'intervalle $\left[x_j - h; x_j + h\right]$. Par conséquent, la variable I_1 ne vaut 1 que pour $i = j$. La variable I_2 quant à elle, prend la valeur 1 pour toutes les observations inférieures à x_j. On peut donc simplifier l'expression de la façon suivante :

$$\hat{F}_{h<d_m}\left(x_j\right) = \frac{1}{nh}\left[\frac{h}{2} + h\left(j-1\right)\right]$$

Finalement, on obtient une expression indépendante du paramètre de lissage h :

$$\hat{F}_{h<d_m}\left(x_j\right) = \frac{j-0.5}{n}$$

Pour un noyau Cauchy, on a :

$$\hat{F}\left(x_j\right) = \frac{1}{nh}\sum_{i=1}^{n} h \int_{t}^{\infty} \frac{1}{\pi\left(1+w^2\right)} dw$$

En résolvant l'intégrale, on a :

$$\hat{F}\left(x_j\right) = \frac{1}{nh}\sum_{i=1}^{n} \frac{h}{\pi}\left[-arctg(t) + arctg(\infty)\right]$$

Pour une valeur de h très faible et en développant la somme on obtient finalement :

$$\hat{F}_{h<d_m}\left(x_j\right) = \frac{1}{nh}\left[\frac{h}{\pi}\left(\pi\left(j-1\right) + \frac{\pi}{2}\right)\right]$$

$$\hat{F}_{h<d_m}\left(x_j\right) = \frac{j-0.5}{n}$$

Annexe B

Calcul du paramètre optimal *p* pour la méthode de Altman et Léger

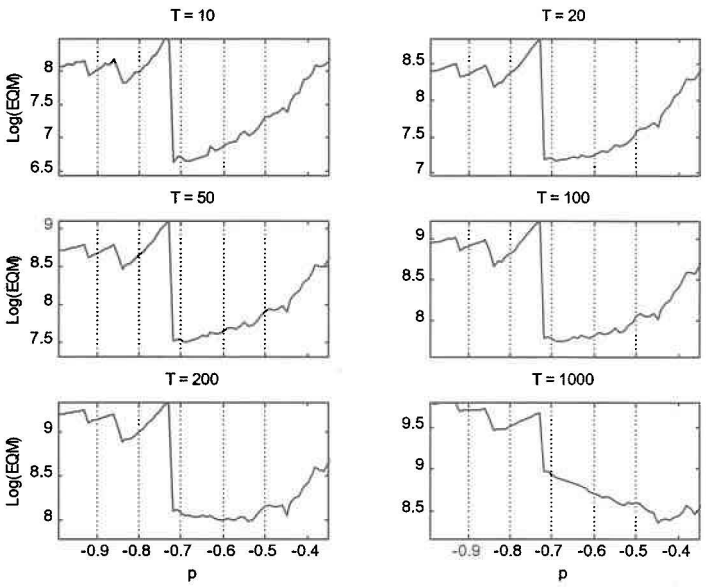


Figure B1: Analyse de sensibilité du paramètre *p* pour l'estimation de quantiles. Résultats provenant de 50 échantillons de taille $n = 10$.

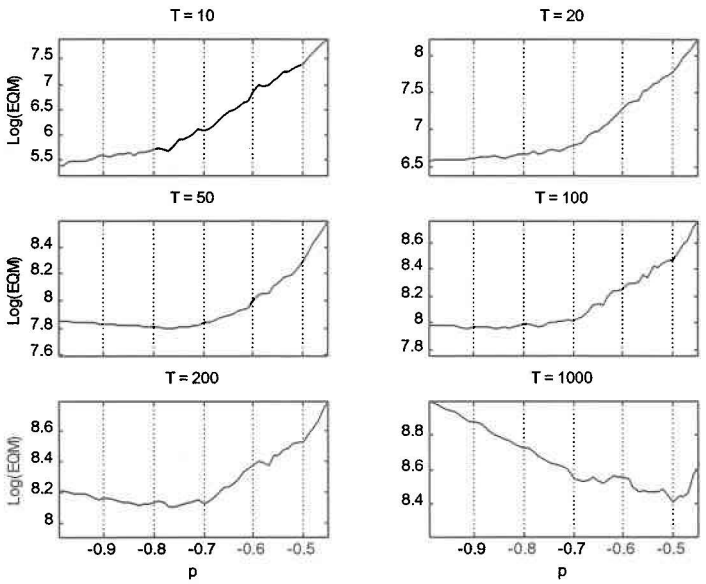


Figure B2: Analyse de sensibilité du paramètre *p* pour l'estimation de quantiles. Résultats provenant de 50 échantillons de taille $n = 25$.

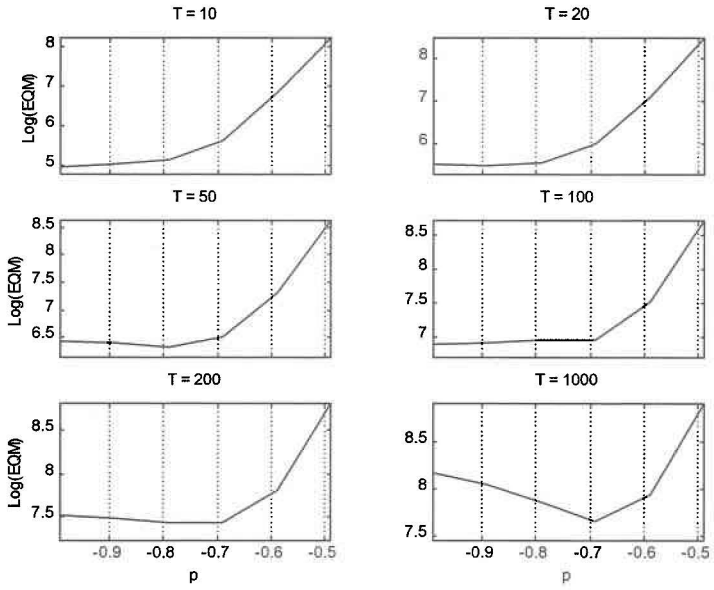
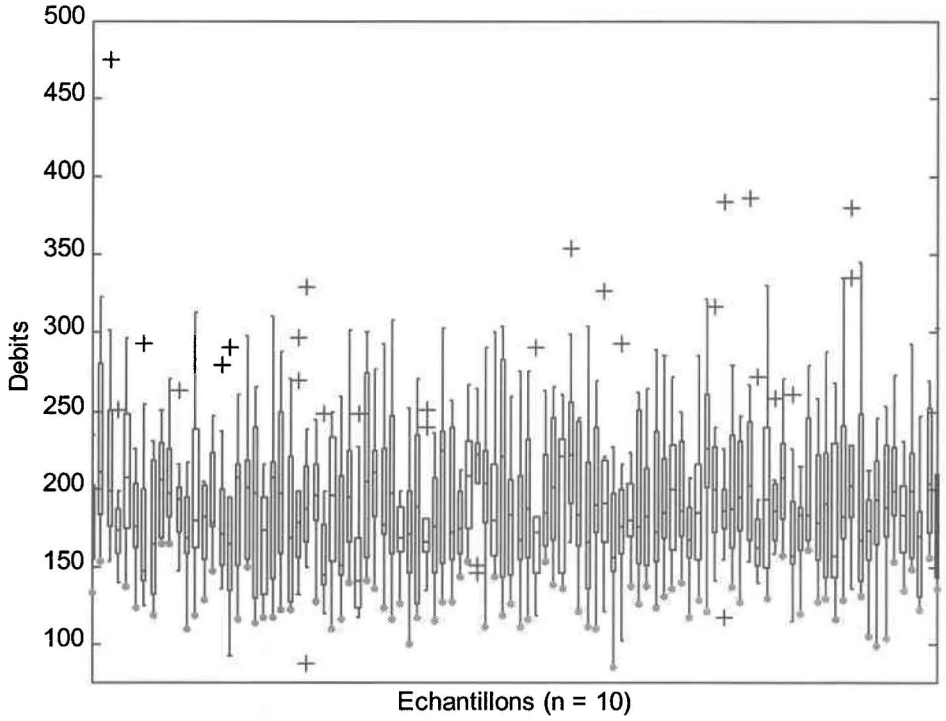
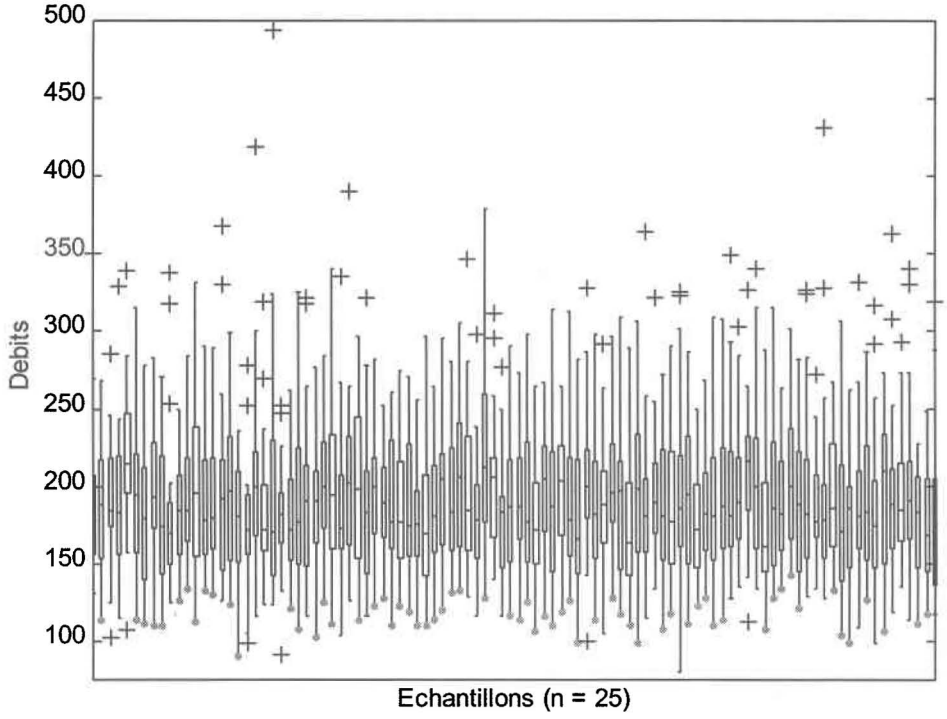


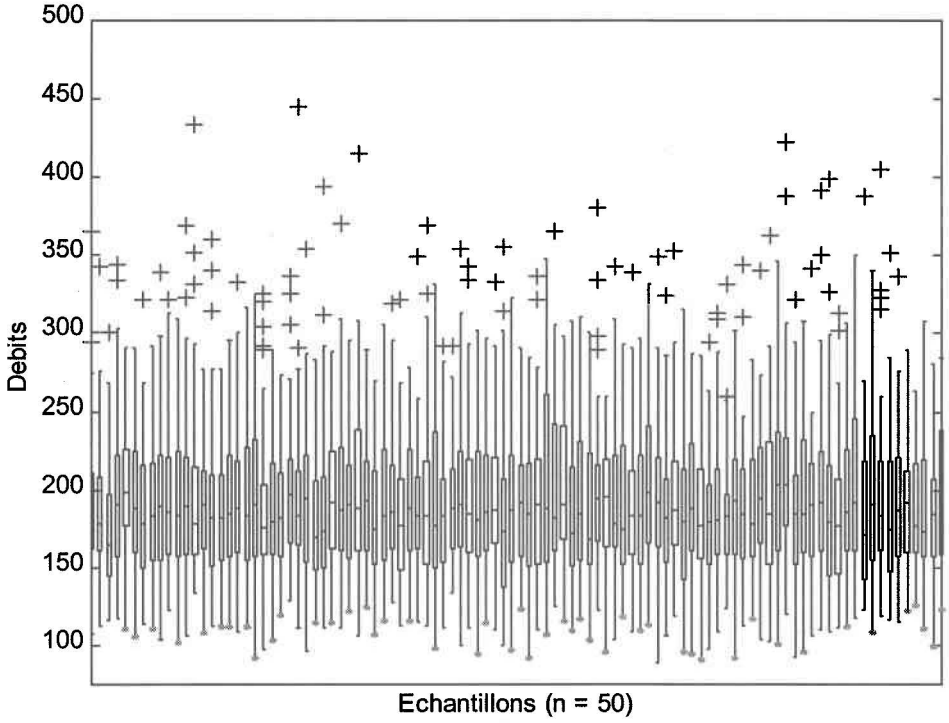
Figure B3: Analyse de sensibilité du paramètre *p* pour l'estimation de quantiles. Résultats provenant de 50 échantillons de taille *n* = 100.

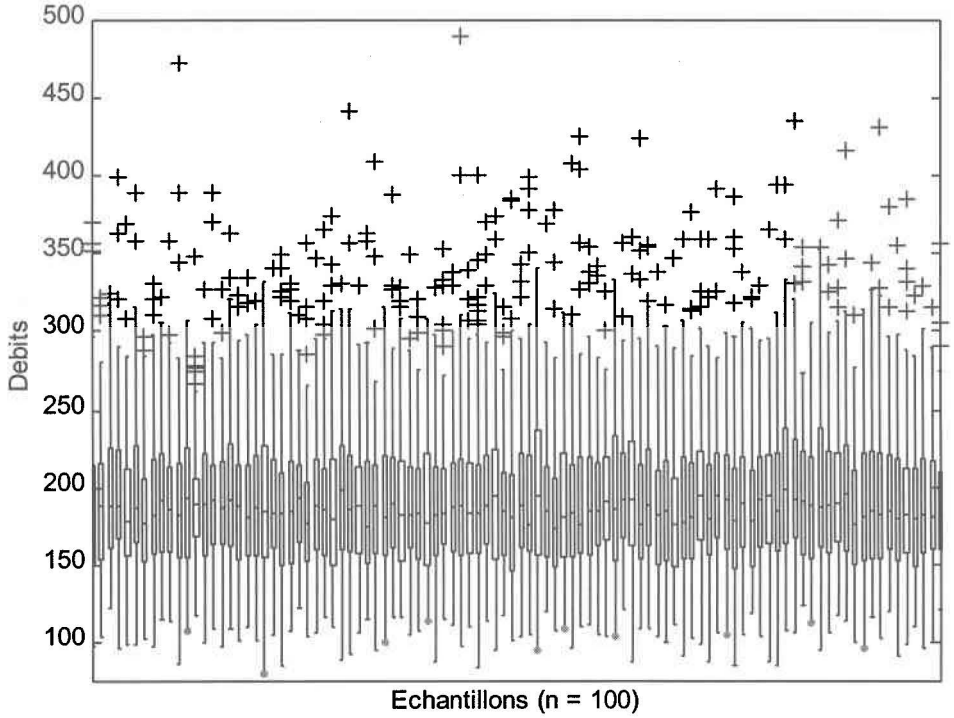
Annexe C

Données synthétiques considérées dans l'étude


500
450
400
350
300
250
200
150
100
Debits
Echantillons (n = 10)


500
450
400
350
300
250
200
150
100
Debits
Echantillons (n = 25)


500
450
400
350
300
250
200
150
100
Debits
Echantillons (n = 50)


500
450
400
350
300
250
200
150
100
Debits
Echantillons (n = 100)

Annexe D

Résultats des différentes comparaisons

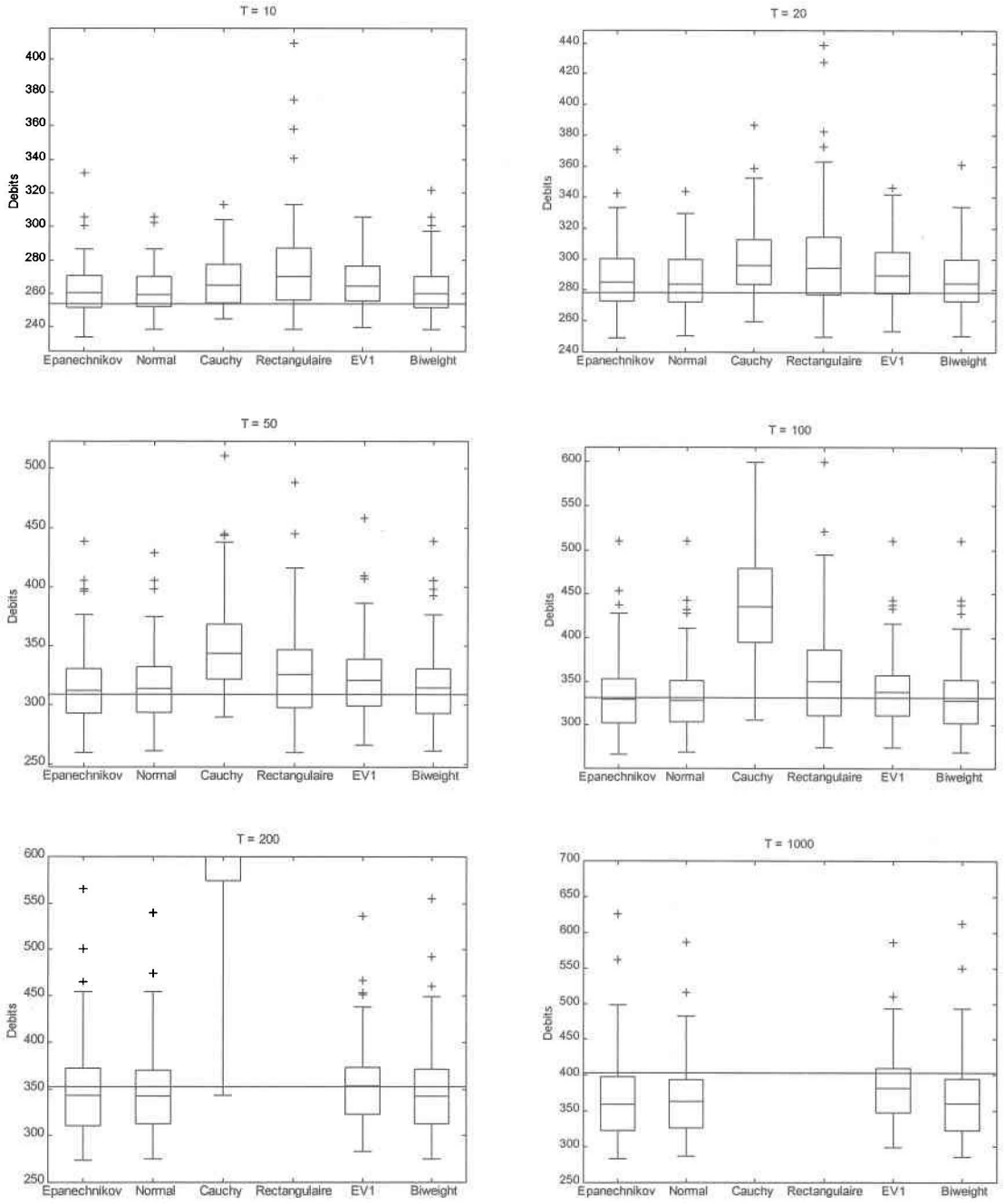


Figure D1 : Estimation des quantiles avec la méthode du maximum de vraisemblance (*MVVC*) pour les différents noyaux avec n=50.

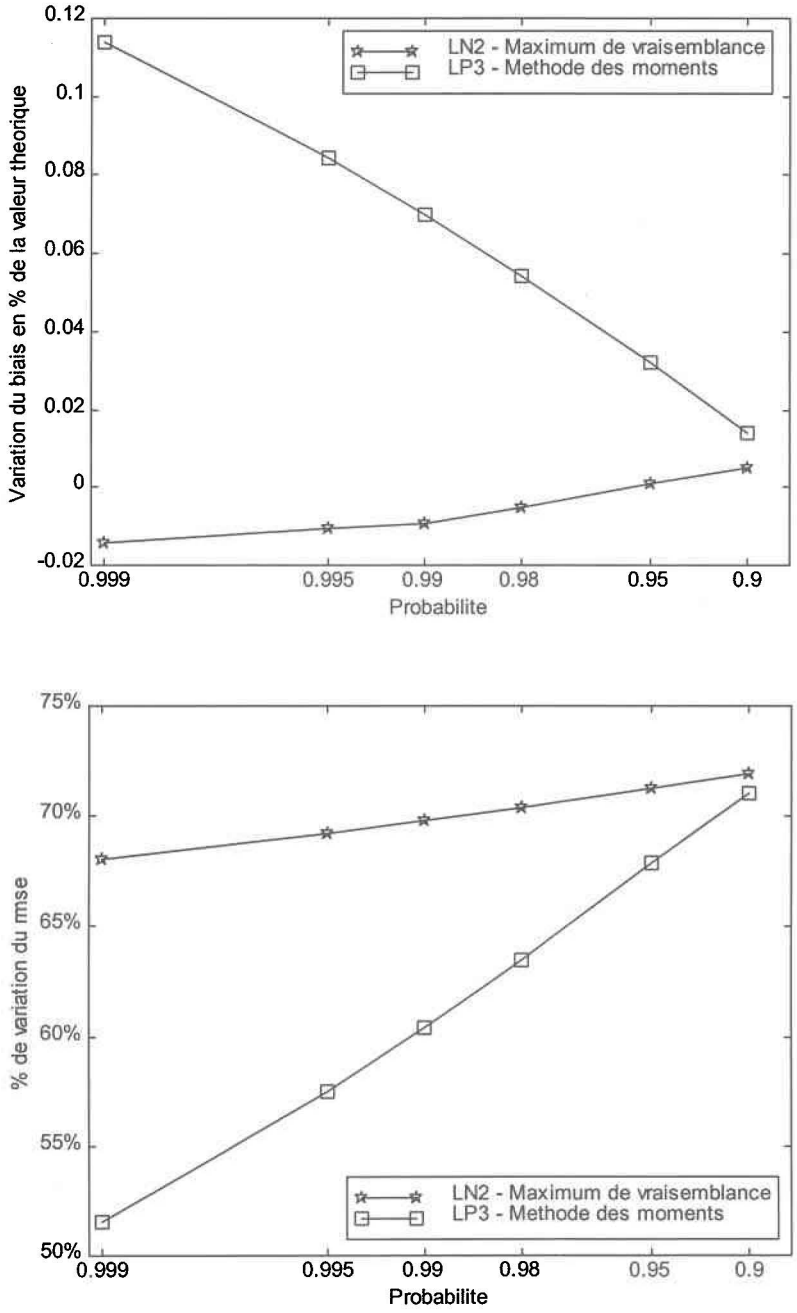


Figure D2 : Impact de la taille de l'échantillon sur le biais et le *RMSE* (différences des résultats obtenus pour n=10 et ceux pour n=100).

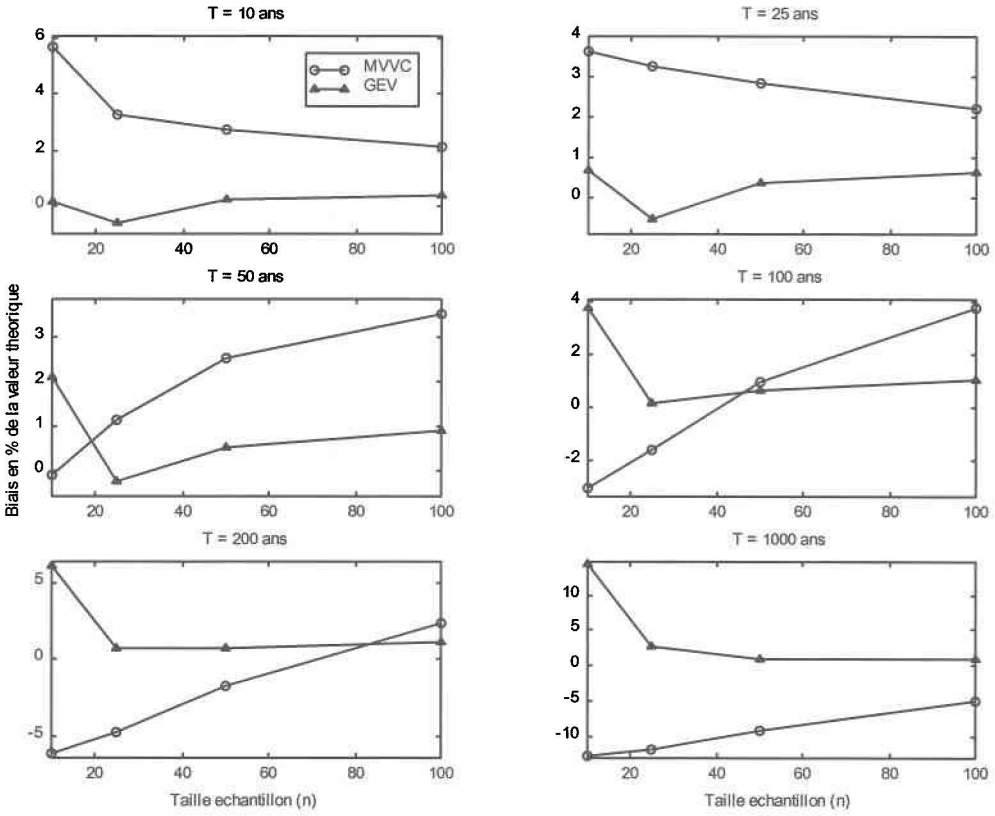


Figure D3 : Variation du biais (exprimé en % de la valeur théorique) selon la taille de l'échantillon pour la méthode *MVVC* (noyau normal) et la GEV.

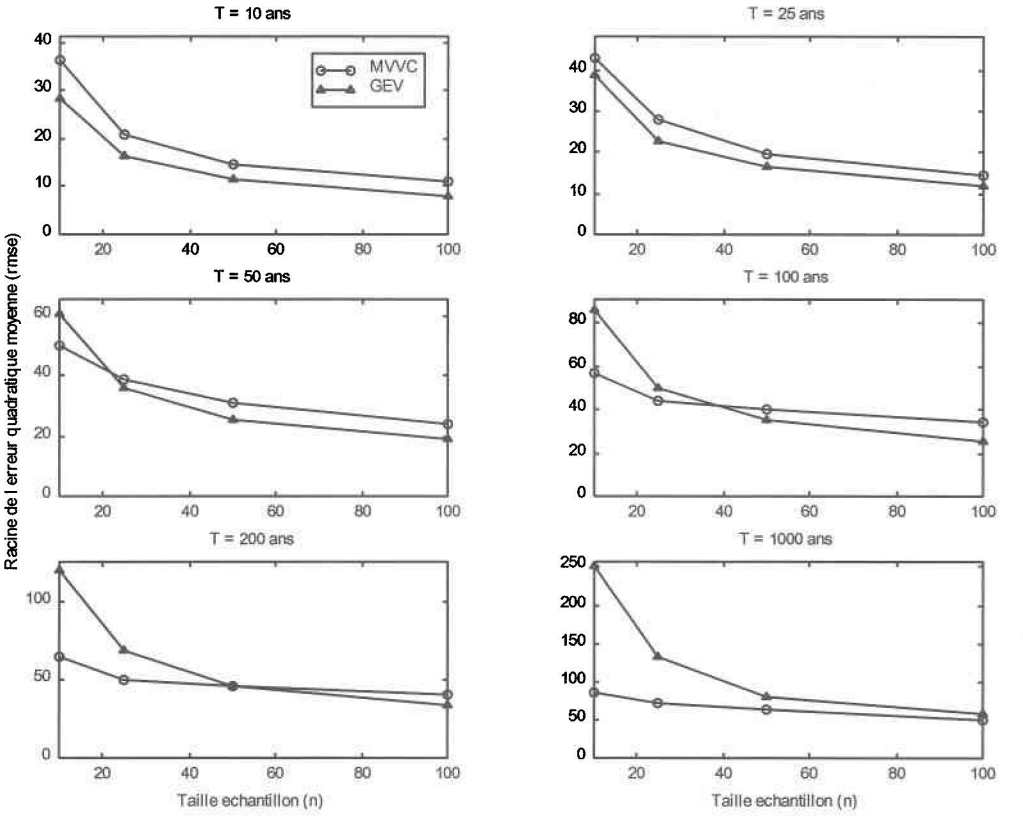


Figure D4 : Variation du *RMSE* selon la taille de l'échantillon pour la méthode *MVVC* (noyau normal) et la GEV.

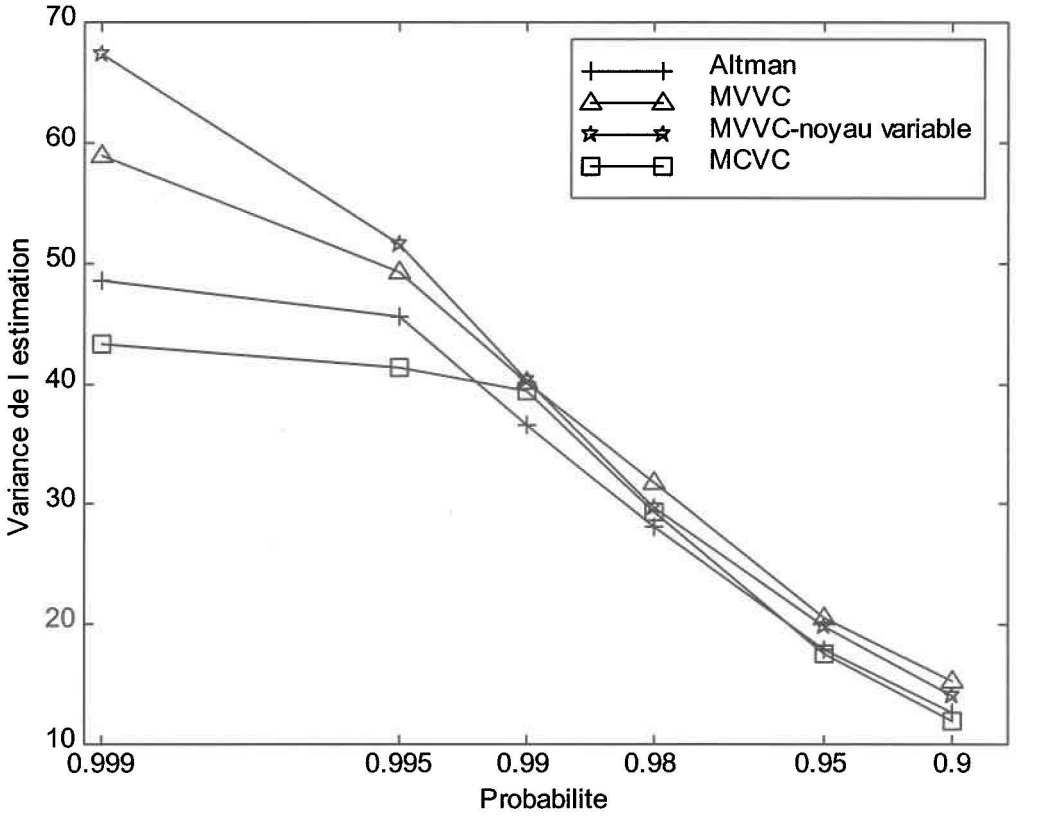


Figure D5 : Variance de l'estimation des quantiles pour les différentes méthodes d'estimation du paramètre de lissage (noyau d'Epanechnikov, $n = 50$).