Revue des sciences de l'eau
Journal of Water Science

Estimation non paramétrique des quantiles de crue par la méthode des noyaux
Nonparametric estimation of quantiles by the kernel method

D. Faucher, P. F. Rasmussen and B. Bobée

Volume 15, Number 2, 2002

URI: https://id.erudit.org/iderudit/705467ar
DOI: https://doi.org/10.7202/705467ar

See table of contents

Publisher(s)

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

ISSN

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

Explore this journal

Cite this article

Faucher, D., Rasmussen, P. F. & Bobée, B. (2002). Estimation non paramétrique des quantiles de crue par la méthode des noyaux. *Revue des sciences de l'eau / Journal of Water Science*, *15*(2), 515–541. https://doi.org/10.7202/705467ar

Article abstract

Traditional flood frequency analysis involves the fitting of a statistical distribution to observed annual peak flows. The choice of statistical distribution is crucial, since it can have significant impact on design flow estimates. Unfortunately, it is often difficult to determine in an objective way which distribution is the most appropriate.

To avoid the inherent arbitrariness associated with the choice of distribution in parametric frequency analysis, one can employ a method based on nonparametric density estimation. Although potentially subject to larger standard error of quantile estimates, the use of nonparametric densities eliminates the need for selecting a particular distribution and the potential bias associated with a wrong choice.

The kernel method is a conceptually simple approach, similar in nature to a smoothed histogram. The critical parameter in kernel estimation is the smoothing parameter that determines the degree of smoothing. Methods for estimating the smoothing parameter have already been compared in a number of statistical papers. The novelty of our work is the particular emphasis on quantile estimation, in particular the estimation of quantiles outside the range of observed data. The flood estimation problem is unique in this sense and has been the motivating factor for this study.

Seven methods for estimating the smoothing parameter are compared in the paper. All methods are based on some goodness-of-fit measures. More specifically, we considered the least-squares cross-validation method, the maximum likelihood cross-validation method, Adamowski's (1985) method, a plug-in method developed by Altman and Leger (1995) and modified by the authors (Faucher et al., 2001), Breiman's goodness-of-fit criterion method (Breiman, 1977), the variable-kernel maximum likelihood method, and the variable-kernel least-squares cross-validation method.

The estimation methods can be classified according to whether they are based on fixed or variable kernels, and whether they are based on the goodness-of-fit of the density function or cumulative distribution function.

The quality of the different estimation methods was explored in a Monte Carlo study. Hundred (100) samples of sizes 10, 20, 50, and 100 were simulated from an LP3 distribution. The nonparametric estimation methods were then applied to each of the simulated samples, and quantiles with return period 10, 20, 50, 100, 200, and 1000 were estimated. Bias and root-mean square error of quantile estimates were the key figures used to compare methods. The results of the study can be summarized as follows :

1. Comparison of kernels. The literature reports that the kernel choice is relatively unimportant compared to the choice of the smoothing parameter. To determine whether this assertion also holds in the case of the estimation of large quantiles outside the range of data, we compared six different kernel candidates. We found no major differences between the biweight, the Normal, the Epanechnikov, and the EV1 kernels. However, the rectangular and the Cauchy kernel should be avoided.

2. Comparison of sample size. The quality of estimates, whether parametric or nonparametric, deteriorates as sample size decreases. To examine the degree of sensitivity to sample size, we compared estimates of the 200-year event obtained by assuming a GEV distribution and a nonparametric density estimated by maximum likelihood cross-validation. The main conclusion is that the root mean square error for the parametric model (GEV) is more sensitive to sample size than the nonparametric model.

3. Comparison of estimators of the smoothing parameter. Among the methods considered in the study, the plug-in method, developed by Altman and Leger (1995) and modified by the authors (Faucher et al. 2001), turned out to perform the best along with the least-squares cross-validation method which had a similar performance. Adamowski's method had to be excluded, because it consistently failed to converge. The methods based on variable kernels generally did not perform as well as the fixed kernel methods.

4. Comparison of density-based and cumulative distribution-based methods. The only cumulative distribution-based method considered in the comparison study was the plug-in method. Adamowski's method is also based on the cumulative distribution function, but was rejected for the reasons mentioned above. Although the plug-in method did well in the comparison, it is not clear whether this can be attributed to the fact that it is based on estimation of the cumulative distribution function. However, one could hypothesize that when the objective is to estimate quantiles, a method that emphasizes the cumulative distribution function rather than the density should have certain advantages.

5. Comparison of parametric and nonparametric methods. Nonparametric methods were compared with conventional parametric methods. The LP3, the 2-parameter lognormal, and the GEV distributions were used to fit the simulated samples. It was found that nonparametric methods perform quite similarly to the parametric methods. This is a significant result, because data were generated from an LP3 distribution so one would intuitively expect the LP3 model to be superior which however was not the case. In actual applications, flood distributions are often irregular and in such cases nonparametric methods would likely be superior to parametric methods.

REVUE DES SCIENCES DE L'EAU, Rev. Sci. Eau 15/2(2002) 515-541

Estimation non paramétrique des quantiles de crue par la méthode des noyaux

Nonparametric estimation of quantiles by the kernel method

D. FAUCHER [1], P.F. RASMUSSEN [2 *], B. BOBÉE [3]

Reçu le 29 mai 2000, accepté le 25 février 2002**.

SUMMARY

Traditional flood frequency analysis involves the fitting of a statistical distribution to observed annual peak flows. The choice of statistical distribution is crucial, since it can have significant impact on design flow estimates. Unfortunately, it is often difficult to determine in an objective way which distribution is the most appropriate.

To avoid the inherent arbitrariness associated with the choice of distribution in parametric frequency analysis, one can employ a method based on nonparametric density estimation. Although potentially subject to larger standard error of quantile estimates, the use of nonparametric densities eliminates the need for selecting a particular distribution and the potential bias associated with a wrong choice.

The kernel method is a conceptually simple approach, similar in nature to a smoothed histogram. The critical parameter in kernel estimation is the smoothing parameter that determines the degree of smoothing. Methods for estimating the smoothing parameter have already been compared in a number of statistical papers. The novelty of our work is the particular emphasis on quantile estimation, in particular the estimation of quantiles outside the range of observed data. The flood estimation problem is unique in this sense and has been the motivating factor for this study.

1. Chaire industrielle en hydrologie statistique, CRSNG/Hydro-Québec, INRS-Eau, Université du Québec, Ste-Foy (Qc), Canada.
2. Département de génie civil, Université du Manitoba. Winnipeg (Manitoba), R3T 5V6 Canada.
3. Chaire industrielle en hydrologie statistique, CRSNG /Hydro-Québec, INRS-Eau, Université du Québec, Ste-Foy (Qc), Canada.

* *Correspondance. E-mail : rasmusse@cc.umanitoba.ca.*

** Les commentaires seront reçus jusqu'au 30 avril 2003.

Seven methods for estimating the smoothing parameter are compared in the paper. All methods are based on some goodness-of-fit measures. More specifically, we considered the least-squares cross-validation method, the maximum likelihood cross-validation method, ADAMOWSKI's (1985) method, a plug-in method developed by ALTMAN and LEGER (1995) and modified by the authors (FAUCHER *et al.*, 2001), BREIMAN's goodness-of-fit criterion method (BREIMAN, 1977), the variable-kernel maximum likelihood method, and the variable-kernel least-squares cross-validation method.

The estimation methods can be classified according to whether they are based on fixed or variable kernels, and whether they are based on the goodness-of-fit of the density function or cumulative distribution function.

The quality of the different estimation methods was explored in a Monte Carlo study. Hundred (100) samples of sizes 10, 20, 50, and 100 were simulated from an LP3 distribution. The nonparametric estimation methods were then applied to each of the simulated samples, and quantiles with return period 10, 20, 50, 100, 200, and 1000 were estimated. Bias and root-mean square error of quantile estimates were the key figures used to compare methods. The results of the study can be summarized as follows:

1. Comparison of kernels. The literature reports that the kernel choice is relatively unimportant compared to the choice of the smoothing parameter. To determine whether this assertion also holds in the case of the estimation of large quantiles outside the range of data, we compared six different kernel candidates. We found no major differences between the biweight, the Normal, the Epanechnikov, and the EV1 kernels. However, the rectangular and the Cauchy kernel should be avoided.

2. Comparison of sample size. The quality of estimates, whether parametric or nonparametric, deteriorates as sample size decreases. To examine the degree of sensitivity to sample size, we compared estimates of the 200 year event obtained by assuming a GEV distribution and a nonparametric density estimated by maximum likelihood cross-validation. The main conclusion is that the root mean square error for the parametric model (GEV) is more sensitive to sample size than the nonparametric model.

3. Comparison of estimators of the smoothing parameter. Among the methods considered in the study, the plug-in method, developed by ALTMAN and LEGER (1995) and modified by the authors (FAUCHER *et al.*, 2001), turned out to perform the best along with the least-squares cross-validation method which had a similar performance. ADAMOWSKI's method had to be excluded, because it consistently failed to converge. The methods based on variable kernels generally did not perform as well as the fixed kernel methods.

4. Comparison of density-based and cumulative distribution-based methods. The only cumulative distribution-based method considered in the comparison study was the plug-in method. ADAMOWSKI's method is also based on the cumulative distribution function, but was rejected for the reasons mentioned above. Although the plug-in method did well in the comparison, it is not clear whether this can be attributed to the fact that it is based on estimation of the cumulative distribution function. However, one could hypothesize that when the objective is to estimate quantiles, a method that emphasizes the cumulative distribution function rather than the density should have certain advantages.

5. Comparison of parametric and nonparametric methods. Nonparametric methods were compared with conventional parametric methods. The LP3, the 2 parameter lognormal, and the GEV distributions were used to fit the simulated samples. It was found that nonparametric methods perform quite similarly to the parametric methods. This is a significant result, because data were generated from an LP3 distribution so one would intuitively expect the

LP3 model to be superior which however was not the case. In actual applications, flood distributions are often irregular and in such cases nonparametric methods would likely be superior to parametric methods.

Key-words: *floods, extreme values, frequency analysis, kernel method, smoothing parameter, comparison.*

RÉSUMÉ

La détermination du débit de crue d'une période de retour donnée nécessite l'estimation de la distribution des crues annuelles. L'utilisation des distributions non paramétriques — comme alternative aux lois statistiques — est examinée dans cet ouvrage. Le principal défi dans l'estimation par la méthode des noyaux réside dans le calcul du paramètre qui détermine le degré de lissage de la densité non paramétrique. Nous avons comparé plusieurs méthodes et avons retenu la méthode *plug-in* et la méthode des moindres carrés avec validation croisée comme les plus prometteuses.

Plusieurs conclusions intéressantes ont été tirées de cette étude. Entre autres, pour l'estimation des quantiles de crue, il semble préférable de considérer des estimateurs basés directement sur la fonction de distribution plutôt que sur la fonction de densité. Une comparaison de la méthode *plug-in* à l'ajustement de trois lois statistiques a permis de conclure que la méthode des noyaux représente une alternative intéressante aux méthodes paramétriques traditionnelles.

Mots clés : *crue, valeurs extrêmes, analyses statistiques, méthode des noyaux, paramètre de lissage, étude de comparaison.*

1 - INTRODUCTION

L'étude des probabilités d'occurrence des crues extrêmes s'effectue généralement à l'aide de méthodes dites paramétriques qui consistent à ajuster des distributions statistiques aux séries de débit maximum annuel. Comme on ne dispose généralement que de peu d'information concernant la distribution de la population, le choix de la loi pour l'ajustement peut être relativement subjectif. L'incertitude liée à la connaissance de la distribution d'une population est donc une source d'erreur qui s'ajoute à l'erreur liée à l'échantillonnage. Pour cette raison, les méthodes non paramétriques représentent un intérêt particulier en hydrologie. Le présent travail porte sur la méthode des noyaux (*kernel method*), méthode non paramétrique qui a gagné au cours des années de plus en plus de popularité en hydrologie et dans d'autres domaines.

Le principe de base de la méthode des noyaux s'apparente relativement bien à la notion d'histogramme couramment utilisée pour l'analyse exploratoire d'un échantillon. L'histogramme donne une idée de la forme de la distribution empirique d'un échantillon en calculant la proportion d'observations se trouvant dans chacun des intervalles de largeur h. Le choix de la largeur h des fenêtres de l'histogramme est déterminant mais quelque peu subjective.

Le concept de fenêtre de l'histogramme est aussi présent dans la méthode des noyaux. On utilise alors le terme « paramètre de lissage » pour désigner la

fenêtre h. Comme dans le cas de l'histogramme, l'estimation de la fonction de probabilité par la méthode des noyaux est principalement conditionnée par le paramètre de lissage. L'importance du paramètre de lissage dans l'estimation par la méthode des noyaux a eu pour effet de concentrer les recherches sur des techniques de calcul du paramètre de lissage conduisant ainsi à une variété importante de méthodes. L'objectif principal de cette étude est d'identifier les techniques d'estimation du paramètre de lissage les plus prometteuses pour l'estimation des quantiles de crue que l'on tentera d'atteindre par le biais d'une comparaison des méthodes les plus fréquemment utilisées en hydrologie.

On doit noter que l'estimation des quantiles de crue pose des problèmes particuliers qui distinguent cette problématique du problème plus général qu'est l'estimation de percentiles. Plus spécifiquement, l'estimation des quantiles de crue implique souvent une ou plusieurs des caractéristiques suivantes :

- séries d'observations courtes (typiquement 20-50 données) ;
- extrapolation à l'extérieur de la zone des données (période de retour > 50 ans) ;
- présence de valeurs singulières (« crue du siècle » dans un petit échantillon) ;
- précision des quantiles estimés importante au niveau économique ;
- besoin de procédures relativement uniformes.

L'utilisation des lois statistiques est la technique la plus commune en hydrologie. Il est donc impératif de comparer la méthode des noyaux aux méthodes d'ajustement de distributions statistiques paramétriques afin d'identifier les avantages de cette méthode. Il est aussi impératif d'en connaître les limites afin d'en éviter une utilisation abusive.

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 » pour désigner la fonction que l'on utilise dans les méthodes non paramétriques. En hydrologie, c'est YAKOWITZ (1983) et ADAMOWSKI et FELUCH (1983) qui ont introduit indépendamment la méthode des noyaux lors d'une conférence de l'AGU à l'automne 1983.

Dans la méthode des noyaux, une fonction $K(x)$ est associée à chaque observation de l'échantillon d'intérêt. La seule véritable restriction concernant le noyau K est que son intégration sur tout le domaine de définition de x doit être égale à un. On rencontre parfois d'autres restrictions théoriques qui sont appliquées à K, comme la symétrie ou la positivité sur tout le domaine de définition du noyau (ADAMOWSKI, 1989). Toutefois, ces restrictions sont surtout introduites afin de simplifier les développements théoriques. L'estimation non paramétrique de la fonction de densité d'un échantillon $\hat{f}_K$ peut se voir comme le cumul des fonctions K de chaque observation sur tout le domaine :

$$\hat{f}_K(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \quad \text{pour} \quad \int K(x)dx = 1 \tag{1}$$

où n est la taille de l'échantillon $\{x_1, x_2, \ldots, x_n\}$, K est la fonction noyau et h est le paramètre de lissage. La fonction de distribution non paramétrique est obtenue en intégrant l'équation (1) :

$$\hat{F}_K(x) = \frac{1}{n}\sum_{i=1}^{n} K_i\left(\frac{x - x_i}{h}\right) \quad \text{où} \quad K_i(U) = \int_{-\infty}^{u} K(w)dw \tag{2}$$

La fonction de distribution non paramétrique (2) peut être utilisée pour estimer les percentiles correspondant à une probabilité au dépassement. Par exemple, dans un contexte hydrologique, on peut estimer le quantile de crue x_T de période de retour de T années en inversant la fonction de distribution de la façon suivante :

$$\hat{x}_T = \hat{F}_K^{-1}\left(1 - \frac{1}{T}\right) \tag{3}$$

où $\hat{F}_K^{-1}$ représente la fonction inverse de $\hat{F}_K$. Dans la plupart des cas, l'expression (3) n'est pas explicite et $\hat{x}_T$ doit être obtenue par une méthode numérique.

De manière générale, le choix du noyau K a relativement peu d'importance par rapport au choix du paramètre de lissage h, qui détermine l'étendue du noyau de chaque côté de l'observation. En appliquant certaines restrictions pour la commodité des développements théoriques, EPANECHNIKOV (1969) a suggéré un noyau qui est optimal au sens de l'IMSE (*integrated mean square error*), c'est-à-dire le noyau qui minimise l'erreur quadratique moyenne intégrée (une définition plus précise de l'IMSE sera donnée plus loin). RAO (1983) quant à lui est arrivé à la conclusion que le choix d'un noyau autre que le noyau optimal ne menait qu'à une faible perte de précision. LALL *et al.* (1993) considèrent que le choix du noyau a une certaine importance, mais que son influence sur l'ensemble de l'estimation est relativement faible. Il est toutefois important d'apporter une certaine précision dans ces conclusions. Sur l'ensemble de l'estimation d'une distribution de probabilité, le type de noyau n'a probablement que peu d'influence, mais dans le cas de l'estimation d'un percentile élevé, l'impact du noyau peut être relativement important. Lorsque l'on extrapole la fonction de distribution, les caractéristiques des extrémités du noyau conditionnent l'estimation et le degré d'extrapolation. Ainsi, lorsque l'on s'intéresse aux événements extrêmes, on croit que le choix du noyau peut être important.

Le paramètre de lissage h est un facteur important dans l'estimation par la méthode des noyaux. Il représente en quelque sorte une fenêtre qui, centrée sur chaque observation, détermine le degré de lissage de l'estimation d'une fonction de densité. Un faible paramètre de lissage implique un faible degré de lissage et résulte en une fonction de densité irrégulière. À l'opposé, une large valeur de h conduit à une estimation lisse. La *figure 1* illustre le rôle du paramètre de lissage dans l'estimation de la fonction de densité. Pour cet exemple, des données ont été simulées à partir d'une distribution normale. On y remarque clairement que le degré de lissage s'accroît avec la largeur du paramètre de lissage. Cette figure illustre l'importance du paramètre de lissage

dans l'estimation de la fonction de densité. Si h est trop faible (*figure 1a*), chacune des observations influence considérablement la forme de la densité. À l'opposé, si h est trop élevé, un surlissage risque de camoufler les particularités de la véritable fonction de densité, comme l'asymétrie ou la multimodalité. Dans le cas de la fonction de distribution, on peut montrer que :

$$\lim_{h\to 0}\hat{F}_K(x_j)=\frac{j-0.5}{n} \quad \text{et} \quad \lim_{h\to\infty}\hat{F}_K(x_j)=\frac{1}{2} \tag{4}$$

où j est le rang de l'observation x_j dans l'échantillon rangé en ordre croissant. Lorsque $h \to 0$, la distribution non paramétrique tend vers la formule de probabilité empirique de Hazen, alors que lorsque $h \to \infty$, la fonction de distribution correspond à une loi uniforme. En raison de l'importance de ce paramètre, la plupart des travaux effectués sur la méthode des noyaux ont porté sur l'estimation du paramètre de lissage. Ainsi, il existe plusieurs techniques d'estimation de h dont certaines des plus connues seront discutées dans ce qui suit.

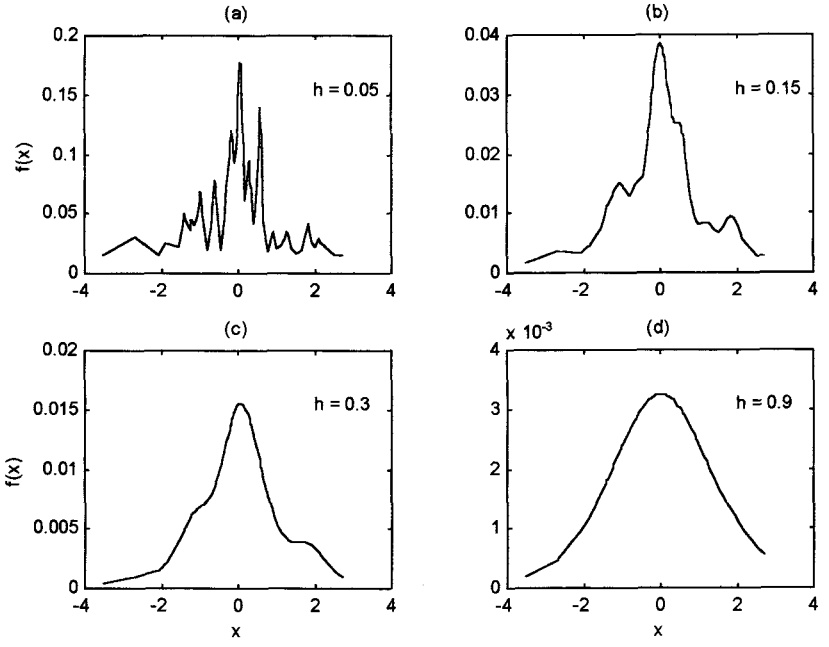

Figure 1 Effet du paramètre de lissage sur l'estimation de la fonction de densité.

Sensitivity of the kernel density to bandwidth.

3 – ESTIMATION DU PARAMÈTRE DE LISSAGE

Dans cette section, les principales méthodes de calcul du paramètre de lissage propice à l'utilisation en hydrologie sont présentées. Ces méthodes peuvent être classées selon deux critères. Le premier critère dépend du type de fonction considérée : la fonction de densité ou la fonction de distribution. Le

second critère est relatif au type de fenêtre considéré : fixe ou variable. Ces deux critères de distinction seront discutés dans ce qui suit et chacune des méthodes considérées dans cette étude sera décrite par la suite.

La qualité de l'estimation d'une fonction g en un point x peut être mesurée par la différence entre la valeur réelle et la valeur estimée de la fonction, c'est-à-dire $g(x) - \hat{g}(x)$. En pratique, on peut considérer l'erreur quadratique moyenne $MSE_g(x)$ ou l'erreur quadratique moyenne intégrée $IMSE_g$ pour évaluer la qualité de l'estimation. Ainsi, l'erreur quadratique moyenne $MSE_g(x)$ pour l'estimation d'une fonction g en x est définie par l'espérance du carré de la différence entre la fonction théorique $g(x)$ et la fonction estimée $\hat{g}(x)$:

$$MSE_g(x) = E\left[\left(g(x) - \hat{g}(x)\right)^2\right] \tag{5}$$

La fonction d'erreur quadratique intégrée $IMSE_g$ est définie quant à elle par l'intégration de la fonction $MSE_g(x)$ sur tout le domaine de x :

$$IMSE_g = \int_{-\infty}^{\infty} E\left[\left(g(x) - \hat{g}(x)\right)^2\right] dx \tag{6}$$

La plupart des méthodes de calcul du paramètre de lissage reposent sur le même principe général : on minimise une estimation de l'erreur quadratique moyenne ou de l'erreur quadratique moyenne intégrée. La fonction g considérée dans (5) et (6) peut être la fonction de densité f ou la fonction de distribution F. Dans la majorité des méthodes d'estimation du paramètre de lissage, la fonction d'erreur est calculée à partir de la fonction de densité. Les estimations obtenues en minimisant la fonction d'erreur quadratique moyenne pour f diffèrent de celles obtenues pour F (FAUCHER, 1999). Les méthodes basées sur la fonction de distribution présentent un intérêt particulier pour l'analyse fréquentielle de crue puisque dans ce cas le calcul du paramètre de lissage est issu de l'erreur de la fonction de distribution qui est à la base du calcul des percentiles. Ainsi, dans cette étude, on distingue les méthodes d'estimation basées sur la fonction de densité de celles basées sur la fonction de distribution.

Les méthodes les plus fréquemment utilisées pour le calcul du paramètre de lissage sont des méthodes à fenêtre fixe, c'est-à-dire que le même paramètre de lissage est utilisé pour chacune des observations de l'échantillon. D'autres méthodes permettent d'adapter le paramètre de lissage selon que l'on se trouve dans une zone à faible ou à forte densité de données. Dans les zones où il y a peu d'observations, généralement les débits faibles ou élevés, le paramètre de lissage est plus large afin de considérer un plus grand nombre de données pour l'estimation locale. À l'inverse, dans les zones où les données sont fortement concentrées, la fenêtre est moins large puisque le nombre d'observations se trouvant dans le voisinage est assez élevé pour effectuer une estimation fiable.

Au *tableau 1*, ces méthodes sont regroupées selon les deux critères décrits précédemment. Les méthodes qui ont été retenues pour l'étude de comparaison sont présentées dans ce qui suit.

Tableau 1 Classification des méthodes de calcul du paramètre de lissage.

MC-VC : Moindres carrés avec validation croisée ; MV-VC : Maximum de vraisemblance avec validation croisée ; CA : Critère d'ADAMOWSKI ; *plug-in* ; CQA : Critère de la qualité de l'adéquation ; MVFV : Maximum de vraisemblance à fenêtre variable ; MCFV-VC : Moindres carrés à fenêtre variable avec validation croisée.

Table 1 *Classification of bandwidth estimation techniques.*

MC-VC: Least squares cross-validation; MV-VC: Maximum likelihood cross-validation; CA: Adamowski criterion; plug-in; CQA: Goodness-of-fit criterion; MVFV: Maximum likelihood variable kernel; MCFV-VC: Least squares cross-validation variable kernel.

Méthode	Fenêtre fixe	Fenêtre variable
f	MC-VC MV-VC	CQA MVFV MCFV-VC
F	CA *plug-in*	–

3.1 Moindres carrés avec validation croisée (MC-VC)

La méthode des moindres carrés avec validation croisée est probablement la méthode la plus connue et la plus utilisée pour l'estimation du paramètre de lissage de la méthode des noyaux. Cette méthode, proposée par RUDEMO (1982) et BOWMAN (1984), consiste à estimer le paramètre de lissage optimal en minimisant l'erreur quadratique moyenne intégrée $IMSE_f$ de la fonction de densité f :

$$IMSE_f = \int \left[\hat{f}_K(x) - f(x)\right]^2 dx \tag{7}$$

Comme la densité théorique f est inconnue, l'*IMSE* doit être estimée. RUDEMO (1982) et BOWMAN (1984) proposent d'utiliser le concept de validation croisée pour estimer l'expression (7). Pour la fonction de densité, le concept de validation croisée consiste à supprimer l'observation i de l'échantillon pour l'estimation de la fonction de densité $f(x_i)$ à ce point. Cet estimateur est dénoté par $\hat{f}_{-i}(x_i)$ et s'exprime de la manière suivante :

$$\hat{f}_{-i}(x_i) = \frac{1}{h(n-1)} \sum_{\substack{j=1 \\ j \neq i}}^{n} K\left(\frac{x_i - x_j}{h}\right) \tag{8}$$

En utilisant (8), RUDEMO (1982) et BOWMAN (1984) ont proposé l'estimateur de l'erreur quadratique moyenne intégrée $IMSE_f$, à une constante additive près, suivant :

$$MCVC_f(h) = \frac{1}{n^2 h} \sum_{i=1}^{n} \sum_{j=1}^{n} K^{(2)}\left(\frac{x_i - x_j}{h}\right) - \frac{2}{nh(n-1)} \sum_{i=1}^{n} \sum_{\substack{j \neq i \\ j=1}}^{n} K\left(\frac{x_i - x_j}{h}\right) \tag{9}$$

où $K^{(2)}$ représente la convolution du noyau K avec lui-même. Ainsi, l'estimation du paramètre de lissage optimal par la méthode des moindres carrés avec validation croisée est obtenue en minimisant l'expression (9).

3.2 Maximum de vraisemblance avec validation croisée (MV-VC)

Les paramètres d'un modèle statistiques peuvent être estimés par la méthode du maximum de vraisemblance. Dans cette méthode, on retient les paramètres qui maximisent la fonction de vraisemblance ou, plus souvent, le logarithme de la fonction de vraisemblance. Par exemple, pour l'estimation d'une loi statistique à partir d'un échantillon, l'estimateur du maximum de vraisemblance serait :

$$\hat{\theta} = \arg \max_{\theta} \ell(\theta|\mathbf{x}) \tag{10}$$

où θ est un vecteur de paramètres, $\mathbf{x}$ est un vecteur d'observations, et

$$\ell(\theta|\mathbf{x}) = \log \prod_{i=1}^{n} f(x_i|\theta) = \sum_{i=1}^{n} \log f(x_i|\theta) \tag{11}$$

Dans le contexte non paramétrique, la forme de f dépend des observations et l'équation (10) n'est pas directement applicable. HABBEMA *et al.* (1974) et DUIN (1976) ont donc suggéré de considérer la méthode de validation croisée pour estimer la fonction de vraisemblance. Supposons que l'on dispose d'une nouvelle donnée y indépendante des données déjà disponibles $\mathbf{x}$. En considérant $\mathbf{x}$ fixe, la fonction log-vraisemblance de h se note de la façon suivante :

$$\ell(h|y) = \log\left[\hat{f}(y|h,\mathbf{x})\right] \tag{12}$$

où $\hat{f}$ est une densité de noyau basée sur les observations $\mathbf{x}$. Comme on ne dispose généralement pas d'un échantillon supplémentaire, la technique de validation croisée est nécessaire. On remplace donc (12) par :

$$\ell(h|x_i) = \log\left[\hat{f}_{-i}(x_i|h)\right] \tag{13}$$

Ceci peut être fait pour chacune des observations. En prenant la moyenne et en utilisant (8), on peut obtenir — après quelques manipulations mathématiques — l'expression suivante pour la log-vraisemblance avec validation croisée :

$$MVVC_f(h) = \frac{1}{n}\sum_{i=1}^{n} \log\left[\sum_{j\neq i} K\left(\frac{x_i - x_j}{h}\right)\right] + \log\left[\frac{1}{h(n-1)}\right] \tag{14}$$

L'estimation du paramètre de lissage optimal par la méthode du maximum de vraisemblance avec validation croisée est obtenue en maximisant cette expression.

3.3 Critère d'ADAMOWSKI (CA)

La méthode d'ADAMOWSKI (1985) s'appuie sur la fonction de distribution pour estimer le paramètre de lissage. Elle consiste à minimiser la différence entre l'estimation de la fonction de distribution par la méthode des noyaux et une estimation de la fonction de distribution empirique. On doit d'abord classer les observations en ordre croissant et calculer la valeur de la probabilité empirique p_j associée à chacune des observations. Le paramètre de lissage optimal est celui qui minimise l'erreur quadratique entre l'estimation non paramétrique et les p_j :

$$AC_F(h) = \sum_{j=1}^{n}\left[\hat{F}_K(x_j) - p_j\right]^2 \tag{15}$$

Il existe plusieurs formules de probabilité empirique. Les formules de probabilités empiriques sont souvent utilisées pour visualiser la distribution des observations d'un échantillon. Elles permettent aussi de détecter la présence de valeurs singulières. Ces formules peuvent s'exprimer sous la forme générale suivante :

$$p_j = \frac{j-\alpha}{n+1-2\alpha} \quad \text{avec} \quad 0 \le \alpha \le 1 \tag{16}$$

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. ADAMOWSKI (1981) propose de considérer un coefficient $\alpha = 0{,}25$ dans l'expression (16) pour le calcul des probabilités empiriques. Ainsi, l'estimation du paramètre de lissage optimal par la méthode d'ADAMOWSKI est obtenu en minimisant l'expression (15).

3.4 Méthode *plug-in* de la fonction de distribution

Une méthode automatique suggérée par ALTMAN et LÉGER (1995) permet d'estimer le paramètre de lissage en minimisant l'erreur quadratique moyenne intégrée de la fonction de distribution $IMSE_F$. Une méthode est dite « automatique » lorsque le paramètre de lissage est calculé directement à partir d'une estimation de l'expression analytique de la valeur optimale du paramètre de lissage. En dérivant la fonction $IMSE_F$ par rapport à h et en égalant à zéro, on peut déterminer la valeur analytique optimale du paramètre de lissage :

$$h_{opt,IMSE_F} = \left[\frac{2\int f(x)^2 dx \int xK(x)K_I(x)dx}{n\int f'(x)^2 f(x)dx \left[\int x^2 K(x)dx\right]^2} \right]^{1/3} \tag{17}$$

où $K_I(x)$ est défini dans (2). ALTMAN et LÉGER (1995) proposent une méthode pour estimer chacun des termes de l'expression (17) impliquant la fonction inconnue f ou sa dérivée. Dénotons par $A(F)$ et $B(F)$ les deux termes inconnus :

$$A(F) = \int f(x)^2 dx \text{ et } B(F) = \int f'(x)^2 f(x)dx \tag{18}$$

Pour estimer le premier terme inconnu $A(F)$, on utilise le concept de validation croisée :

$$\hat{A}(F) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \ne i} \frac{1}{h_A} K_A^{(2)} \left(\frac{x_i - x_j}{h_A} \right) \tag{19}$$

où K_A et h_A sont le noyau et le paramètre de lissage associés à l'estimation de $A(F)$, et $K_A^{(2)}$ représente la convolution de K_A avec lui-même. Pour l'estimation de $B(F)$, ALTMAN et LÉGER (1995) ont introduit l'estimateur suivant :

$$\hat{B}(F) = \frac{1}{n^3 h_B^4} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} K'_B \left(\frac{x_i - x_j}{h_B} \right) K'_B \left(\frac{x_i - x_k}{h_B} \right) \tag{20}$$

où K_B et h_B sont le noyau et le paramètre de lissage associés à l'estimation de $B(F)$ et $K'_B(t)$ est la dérivée du noyau $K_B(t)$ par rapport à t. En introduisant les estimateurs (19) et (20) dans l'expression (17), on obtient l'estimation du paramètre $h_{opt,\, IMSE_F}$ suivante :

$$\hat{h}_{opt,IMSE_F} = \left[\frac{2\hat{A}(F)\int xK(x)K_I(x)dx}{n\hat{B}(F)\left[\int x^2K(x)dx\right]^2} \right]^{1/3} \quad (21)$$

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. 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 pilote n'est fonction que de la taille de l'échantillon et doit être modifié pour tenir compte de la variabilité de l'échantillon. Dans le but de proposer un tel paramètre, nous considérons les contraintes suivantes (HALL et MARRON, 1987) :

$$n\alpha \to \infty \quad \text{et} \quad n^{1/4}\,\alpha \to 0 \quad \text{lorsque } n \to \infty \quad (22)$$

En prenant $\alpha = n^p$, on peut montrer que pour satisfaire ces contraintes, p doit être compris entre – 1 et – 0,25. Dans leurs études de simulation, ALTMAN et LÉGER (1995) ont trouvé qu'une valeur de p = – 0,3 donnait les meilleurs résultats (LÉGER, communication personnelle). Toutefois, toute valeur respectant les conditions (22) peut être considérée. Une façon intuitive de considérer la variabilité de l'échantillon est d'introduire l'écart-type dans le calcul du paramètre pilote de la façon suivante :

$$\alpha_v = \hat{\sigma} n^p \quad \text{pour} \quad -1 < p < -0,25 \quad (23)$$

Le paramètre p a été calibré lors d'une étude de simulation effectuée à partir de trois distributions parentes, la LP3, la GEV et un mélange de distributions normales (FAUCHER *et al.*, 2001). On a donc déterminé une expression pour le calcul du paramètre pilote, qui est fonction de la période de retour, de la taille de l'échantillon et de la variance empirique :

$$\alpha_v = \hat{\sigma} n^{a+b\,\log(T)} \quad (24)$$

où a = – 1,0234 et b = 0,0469. Ainsi, l'estimation du paramètre de lissage optimal par la méthode *plug-in* de la fonction de distribution est obtenue à partir des expressions (19-21) en utilisant l'équation (24) pour le calcul du paramètre pilote.

3.5 Critère de la qualité de l'adéquation (CQA)

Le critère de la qualité de l'adéquation (BREIMAN *et al.*, 1977) s'applique en considérant une fenêtre h variable. Ainsi, on calcule un paramètre de lissage pour chacune des observations de l'échantillon. Dans la méthode développée par BREIMAN *et al.* (1977), la valeur du paramètre de lissage associé à une observation donnée dépend de la distance de cette observation à son k^e voisin le plus proche. Ils proposent donc d'utiliser un paramètre de lissage variable h_i que l'on estime par $a_k d_{ki}$, où a_k est un paramètre 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. L'estimation de la fonction de densité par la méthode des noyaux à fenêtre variable s'exprime alors de la manière suivante :

$$\hat{f}_K(x) = \sum_{i=1}^{n} \frac{1}{n a_k d_{ki}} K\left(\frac{x - x_i}{a_k d_{ki}}\right) \tag{25}$$

Au lieu d'avoir à estimer un seul paramètre h, on doit optimiser deux paramètres, k et a_k, qui permettent de calculer le paramètre de lissage pour chacune des observations. Pour ce faire, BREIMAN *et al.* (1977) notent que la quantité :

$$\hat{W}_j = e^{-2n\hat{f}_K(x_j)d_{1j}} \qquad j = 1,\ldots,n \tag{26}$$

où d_{1j} est la distance entre la j^e observation et son premier voisin le plus proche est distribuée approximativement selon une loi uniforme sur [0,1]. Ils proposent ensuite de ranger l'échantillon de W en ordre croissant et d'utiliser l'estimateur :

$$\hat{S} = \sum_{j=1}^{n} \left(\hat{W}_{(j)} - \frac{j}{n}\right)^2 \tag{27}$$

pour évaluer l'adéquation de l'ajustement. Ainsi, l'estimation des paramètres de lissage par la méthode du critère de la qualité de l'adéquation est obtenue en calculant les paramètres k et a_k correspondants au minimum de l'expression (27).

3.6 Maximum de vraisemblance à fenêtre variable (MVFV)

Une méthode à fenêtre variable basée sur l'optimisation de la fonction de vraisemblance a été proposée par ADAMOWSKI (1989). Les paramètres à estimer sont les mêmes que dans le cas de la méthode du critère de la qualité de l'ajustement (CQA), c'est-à-dire les paramètres k et a_k ainsi que la fonction de densité définie en (25). Pour le choix de la valeur k, ADAMOWSKI (1989) s'appuie sur la propriété suivante :

$$\frac{a_k \overline{d}_k^2}{\sigma_{d_k}} \approx \text{constante} \tag{28}$$

où $\overline{d}_k$ 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 valeur optimale de k correspond à un coude sur la courbe de $\overline{d}_k$ en fonction de k. Ainsi, on retient la valeur de k suivant un changement brusque de la courbe. L'identification visuelle du coude de la fonction peut parfois être difficile. En pratique, il est possible d'identifier le changement brusque de la courbe en évaluant la dérivée seconde pour chaque valeur de k. On considère alors la valeur k correspondant à la dérivée seconde maximale (FAUCHER, 1999).

Une fois la valeur de k fixée, on peut dériver la valeur de a_k qui maximise la fonction de vraisemblance définie de la manière suivante :

$$L\left[a_k \middle| x_1, x_2, \ldots, x_n\right] = \sum_{i=1}^{n} \log\left[\hat{f}_K\left(x_i \middle| a_k\right)\right] \tag{29}$$

où $\hat{f}_K(x_i|a_k)$ est obtenu à partir de (25). Si on dérive cette fonction par rapport à a_k et qu'on l'égale à zéro, il est possible de déterminer le maximum de la fonction. Il s'agit de résoudre :

$$a_k + \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\sum_{j\neq i}^{n}\frac{x_i - x_j}{d_{kj}^2}K'\left(\frac{x_i - x_j}{a_k d_{kj}}\right)}{\sum_{j\neq i}^{n}\frac{1}{d_{kj}}K\left(\frac{x_i - x_j}{a_k d_{kj}}\right)}\right] = 0 \qquad (30)$$

Ainsi, l'estimation des paramètres de lissage par la méthode à fenêtre variable du maximum de vraisemblance est obtenue en déterminant graphiquement le paramètre k et en calculant la valeur de a_k constituant la solution de l'expression (30).

3.7 Adaptation des méthodes à fenêtre fixe

Il est possible d'adapter les méthodes d'estimation à fenêtre fixe pour quelles soient utilisables dans un contexte d'estimation 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 la présente étude, seule la méthode des moindres carrés avec validation croisée pour noyau fixe a été considérée dans le cas à fenêtre variable (MCFV-VC). Dans ce cas, l'estimation de la fonction d'erreur $IMSE_f$ est donnée par :

$$MCVC_{a_k,k} = \frac{1}{n^2 a_k}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{d_{ki}}K^{(2)}\left(\frac{x_i - x_j}{a_k d_{ki}}\right) - \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j\neq i\\ j=1}}^{n}\frac{1}{a_k d_{ki}}K\left(\frac{x_i - x_j}{a_k d_{ki}}\right) \qquad (31)$$

Ainsi, l'estimation des paramètres de lissage par la méthode à fenêtre variable des moindres carrés avec validation croisée est obtenue en minimisant l'expression (31). À noter que 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.

4 – ÉVALUATION ET COMPARAISON

La comparaison des techniques de calcul du paramètre de lissage ainsi que la comparaison de la méthode des noyaux aux méthodes paramétriques ont été effectuées par le biais d'une étude de simulation de Monte Carlo. Pour ce faire, des groupes de 100 échantillons de taille n = 10, 20, 50 et 100 ont été simulés à partir de la distribution log-pearson type 3 (LP3). La comparaison des méthodes est effectuée en estimant des quantiles de crue de période T = 10, 20, 50, 100, 200 et 1 000 ans, pour les 400 échantillons, à partir des sept méthodes non paramétriques décrites précédemment et à partir de l'ajustement des distributions log-pearson type 3 (LP3), loi généralisée des valeurs extrêmes (GEV) et log-normale à deux paramètres (LN2). Les méthodes d'ajustement considérées sont la méthode des moments (MM) pour LP3, la méthode

des moments pondérés (MMP) pour GEV et la méthode du maximum de vraisemblance (MV) pour LN2. Les détails concernant l'étude de simulation sont présentés au *tableau 2*.

Tableau 2 Paramètres de simulation.
Table 2 *Simulation parameters.*

	Données						
	α	λ	m	C_v	C_s	n	T (années)
LP3	119,72	846,88	−1,85	0,2489	0,8417	10, 20, 50, 100	10, 20, 50, 100, 200, 1 000
	Méthodes d'estimation						
Paramétriques	LP3 (MM) GEV (MMP) LN2 (MV)						
M. des noyaux	MC-VC MV-VC AC *Plug-in* CQA MVFV MCFV-VC						

4.1 Comparaison des noyaux

Dans le but d'évaluer l'importance de la fonction noyau dans l'estimation non paramétrique, une comparaison des résultats obtenus à partir de six noyaux différents a été effectuée (biweight, Epanechnikov, normal, EV1, Cauchy et rectangulaire.) Ces noyaux sont présentés au *tableau 3* où ils sont classés en deux catégories. La catégorie des noyaux à support borné concerne les noyaux qui sont non nuls dans un certain domaine borné et nuls pour les valeurs de t à l'extérieur de ce domaine. Les noyaux asymptotiques sont quant à eux définis sur tout le domaine de t. On peut noter que tous ces noyaux sont symétriques sauf le noyau EV1 qui présente une asymétrie positive.

Tableau 3 Présentation des noyaux ($t = h^{-1}(x - x_i)$).
Table 3 *Kernel functions* ($t = h^{-1}(x - x_i)$).

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

Pour cette comparaison, les calculs ont été effectués à l'aide de la méthode MV-VC sur 100 échantillons de taille n = 50. Les résultats sont présentés à la *figure 2*. En examinant la courbe du RMSE (*figure 2a*) et la courbe du biais (*figure 2b*), on identifie clairement quatre noyaux menant à des résultats semblables, les noyaux biweight, Epanechnikov, normal et EV1. Le type de support ne joue pas un rôle important dans l'estimation puisque parmi les quatre

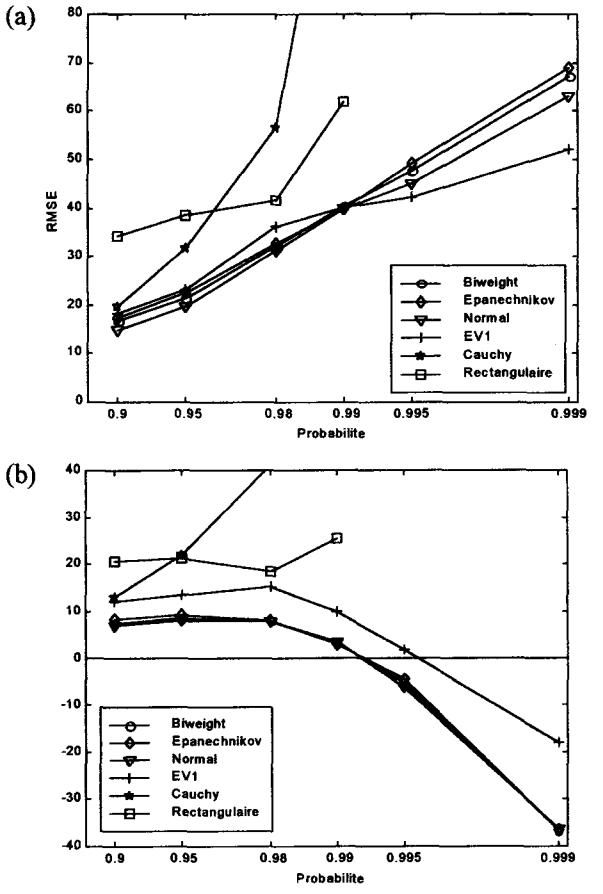

Figure 2 Comparaison de six noyaux : (a) RMSE ; (b) biais.
Comparison of kernels: (a) RMSE; (b) bias.

meilleurs noyaux, deux sont à support fini et deux sont asymptotiques. Le noyau asymétrique EV1 semble préférable pour de très grandes périodes de retour. Pour les autres périodes de retour, il est tout à fait convenable d'utiliser un des quatre noyaux pour l'estimation de quantiles. Les deux autres noyaux, le noyau Cauchy et le noyau rectangulaire, ne peuvent être recommandés sur la base de cette étude. Le noyau Cauchy conduit à une surestimation importante alors que la variance des quantiles estimés à l'aide du noyau rectangulaire devient considérable au-delà de 100 ans. Ces résultats sont aussi valables pour d'autres tailles d'échantillon.

4.2 Importance de la taille de l'échantillon

En statistique appliquée, on est souvent confronté aux problèmes qu'engendrent les échantillons de faible taille. Dans ces conditions, l'efficacité des méthodes non paramétriques peut être mise en doute. Toutefois, les problèmes liés aux échantillons de faible taille sont aussi présents dans l'application des

méthodes paramétriques. On a donc tenté d'évaluer l'importance de la taille d'échantillon pour les deux types de méthodes. Pour ce faire, des groupes de 100 échantillons de taille n = 10, 25, 50 et 100 ont été générés à partir de la LP3. Une loi paramétrique, soit la GEV estimée par la méthode des moments pondérés, ainsi qu'une méthode non paramétrique basée sur l'estimateur MV-VC ont été appliquées en estimant les six quantiles précédents. Les résultats obtenus avec les deux méthodes pour une période de retour de 200 ans sont présentés à la *figure 3*.

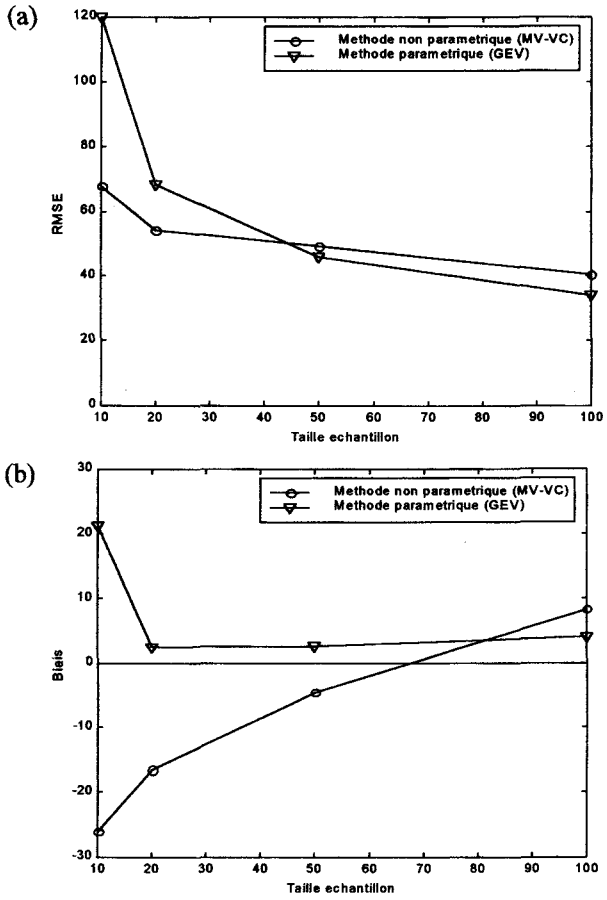

Figure 3 Importance de la taille de l'échantillon sur la qualité de l'estimation pour une période de retour de 200 ans : (a) RMSE ; (b) biais.

Importance of sample size for the accuracy of the 200 year event estimate: (a) RMSE; (b) bias.

La méthode par la GEV semble davantage sensible à la taille d'échantillon que la méthode des noyaux lorsque l'on considère le RMSE. En examinant la *figure 3a*, on remarque que l'écart entre le RMSE pour n = 10 et le RMSE pour n = 100 est plus élevé dans le cas de la GEV que dans le cas de la méthode des noyaux. À la *figure 3b* on remarque que l'accroissement de la taille de

l'échantillon permet de réduire le biais de l'estimation avec la méthode des noyaux. On note ainsi une tendance de la méthode MV-VC à sous-estimer les quantiles pour des échantillons de faible taille et à tendre vers une surestimation en augmentant la taille des échantillons. Pour ce qui est de la méthode paramétrique, on observe un biais relativement faible et constant, sauf pour les échantillons de faible taille (n = 10). Ces observations sont aussi valables pour les autres périodes de retour qui ne sont pas présentées ici. D'autres comparaisons effectuées en considérant les résultats obtenus avec la distribution log-normale à deux paramètres et la distribution log-Pearson type 3 permettent de tirer les mêmes conclusions que dans le cas de la GEV.

En résumé, l'augmentation de la taille de l'échantillon aurait une plus grande incidence sur la valeur du biais dans le cadre de la méthode des noyaux que dans le cas de la méthode paramétrique. En revanche, l'accroissement de n aurait un plus grand impact sur la variance de l'estimation dans le cas de la distribution paramétrique que dans le cas de la méthode des noyaux. Bien sûr, 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 nécessairement avec la taille de l'échantillon.

4.3 Comparaison des méthodes de calcul du paramètre de lissage

La comparaison porte sur six des sept méthodes présentées précédemment. Au cours de nos simulations, nous avons noté que la méthode d'Adamowski sous-estime systématiquement le paramètre de lissage. En fait, il est possible de démontrer (FAUCHER, 1999) que le minimum de la fonction d'ADAMOWSKI est atteint pour un paramètre de lissage qui est inférieur au plus faible des écarts entre les observations prises deux-à-deux. Comme il n'y a en pratique aucun lissage dans ce cas, cette méthode n'est pas prise en compte dans l'étude de comparaison.

La *figure 4* présente les résultats de la comparaison pour des échantillons de taille n = 50. La *figure 4a,* représentant le RMSE, est présentée sous échelle logarithmique. On remarque sur cette figure, que la méthode CQA et la MCFV-VC présentent des estimations de variabilité de relativement grande. Les quatre autres méthodes sont très semblables en terme de RMSE. Sur la *figure 4b* seules les quatre méthodes ayant les RMSE les plus faibles ont été présentées. La méthode *plug-in* est la méthode qui présente le biais le plus faible. La méthode MC-VC présente toutefois des valeurs de biais approximativement de même ordre pour des périodes de retour de T = 10, 20, 50 et 100 ans, une différence significative devient présente à 200 ans et s'accroît jusqu'à 1 000 ans.

Les paramètres locaux estimés par la méthode CQA et la méthode MCFV-VC présentent une variabilité importante par rapport au paramètre global entraînant ainsi une surestimation considérable des quantiles. Il n'est donc pas recommandé d'appliquer ces méthodes pour l'estimation des quantiles de crue.

La seule méthode à fenêtre variable qui semble procurer des résultats raisonnables est la méthode MVFV. Toutefois, cette méthode induit un biais positif peu importe la valeur de T. Si on favorise une approche conservatrice, sachant que la méthode MVFV surestime le débit, on a peut-être avantage à utiliser cette méthode, surtout pour estimer le quantile correspondant à une

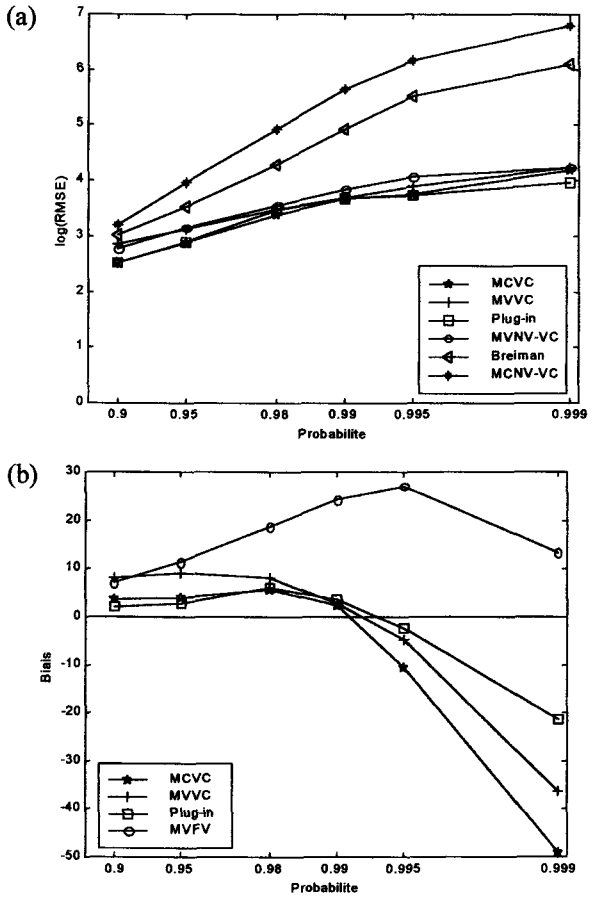

Figure 4 Comparaison des méthodes d'estimation du paramètre de lissage : (a) RMSE ; (b) biais.

Comparision of bandwidth selection methods: (a) RMSE; (b) bias.

période de retour aussi élevée que 1 000 ans. À noter que le problème pratique que cause généralement le choix visuel du paramètre k pour cette méthode peut être évité en évaluant la valeur de la dérivée seconde comme on l'a mentionné précédemment.

L'estimation du paramètre de lissage avec la méthode MV-VC mène à des résultats relativement intéressants, surtout pour des périodes de retour de 100 et de 200 ans. Toutefois, pour des périodes de retour de 10 et 20 ans, le RMSE et le biais sont légèrement plus élevés que ceux obtenus avec la méthode MC-VC et la méthode *plug-in*. De plus, l'estimation par la fonction de vraisemblance est très sensible aux valeurs extrêmes de l'échantillon.

Parmi les deux méthodes les plus efficaces (*plug-in* et MC-VC) pour l'estimation de quantile, on accorde un avantage à la méthode *plug-in* pour l'estimation des quantiles supérieurs (200 et 1 000 ans). En effet, la méthode MC-VC présente un inconvénient majeur qui ne se retrouve pas dans la

méthode *plug-in* : la difficulté d'estimer *h* pour des échantillons formés de données discrètes ou arrondies (SILVERMAN, 1986). L'optimisation de l'expression (9) peut conduire à un paramètre de lissage nul dans de pareils cas (FAUCHER *et al.*, 2001).

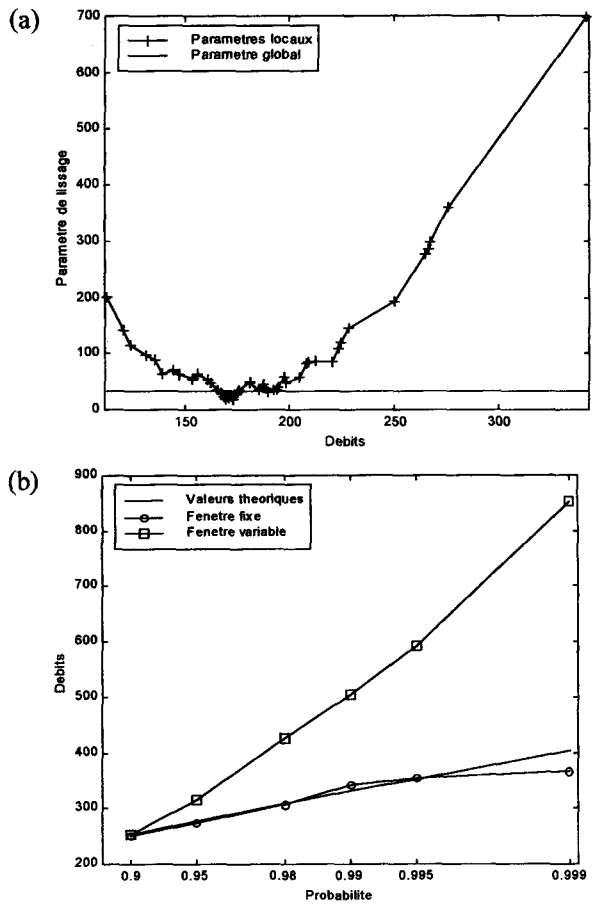

Figure 5 (a) Comparaison des paramètres de lissage variables au paramètre de lissage global calculés avec la méthode des moindres carrés (MC-VC) ; (b) estimation des quantiles de crue.

(a) Comparison of variable (MCFV-VC) and fixed (MC-VC) bandwidth methods; (b) Quantiles estimated by variable and fixed bandwidth methods.

4.4 Méthodes à fenêtre variable

Comme on a pu le constater sur la *figure 4*, deux des trois méthodes à fenêtre variable conduisent à une surestimation importante des quantiles par rapport aux méthodes à fenêtre fixe. Ce problème provient du fait que les paramètres de lissage locaux sont relativement variables à l'intérieur d'un même échantillon et qu'ils accordent un poids considérable aux observations

extrêmes. Un exemple de comparaison des paramètres locaux au paramètre global est présenté à la *figure 5*. La *figure 5a* met en évidence les valeurs élevées des paramètres de lissages associés aux grandes valeurs de l'échantillon. Il en résulte un surlissage de la queue droite de la distribution et de manière générale, la surestimation des quantiles comme on peut le constater sur la *figure 5b*. La méthode à fenêtre fixe quant à elle, permet d'estimer adéquatement les quantiles avec un paramètre global raisonnable.

On a vu que la méthode du maximum de vraisemblance à fenêtre variable procure tout de même des résultats presque comparables à ceux correspondant aux méthodes à fenêtre fixe. L'estimation des valeurs de a_k est relativement plus faible pour cette méthode que dans le cas des deux autres méthodes à fenêtre variable et les distances d_{ki} sont davantage uniformes sur l'échantillon. La *figure 6* montre la distribution des coefficients de variation ($\sigma_{d_{ki}}/\mu_{d_{ki}}$) pour les 100 échantillons du groupe de taille n = 50 pour chacune des trois méthodes à fenêtre variable. On remarque que les coefficients de variation de la méthode du maximum de vraisemblance sont significativement plus faibles que ceux obtenus avec les deux autres méthodes. Par conséquent, la faible variabilité des paramètres locaux fait en sorte que l'on se rapproche d'un paramètre global, d'où la similarité des résultats avec ceux des méthodes à fenêtre fixe.

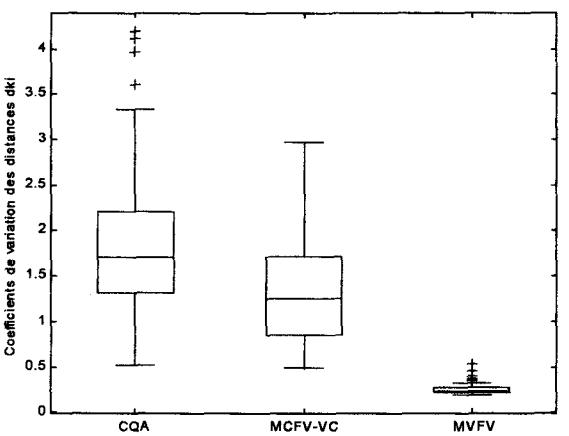

Figure 6 Comparaison des coefficients de variation des distances pour chacune des méthodes à fenêtre variable.

Coefficient of variation of for variable bandwidth selection methods.

4.5 Méthode des noyaux et méthodes paramétriques

La méthode des noyaux a été comparée à l'ajustement de trois lois statistiques, la LP3, la GEV et la LN2. Pour la comparaison, nous avons retenu la méthode d'estimation du paramètre de lissage qui nous semble la plus efficace, soit la méthode *plug-in*. Cette comparaison est illustrée à la *figure 7*. Sur cette figure, seuls les résultats obtenus pour le groupe d'échantillons de taille n = 50 sont présentés. D'abord, on remarque que le RMSE est minimal pour la loi LN2 et qu'il est relativement comparable pour les autres méthodes (*figure 7a*).

Pour ce qui est du biais, on remarque sur la *figure 7b* qu'il est minimal avec la loi GEV. Le biais de la méthode *plug-in* est comparable à celui de la LP3 pour des périodes de retour de 10 à 100 ans. Toutefois, à partir d'une période de retour de 200 ans, le biais de la méthode *plug-in* tend à augmenter plus rapidement que pour les méthodes paramétriques. À T = 200 et 1 000 ans, il n'est tout de même pas surprenant de voir la méthode des noyaux sous-estimer les quantiles.

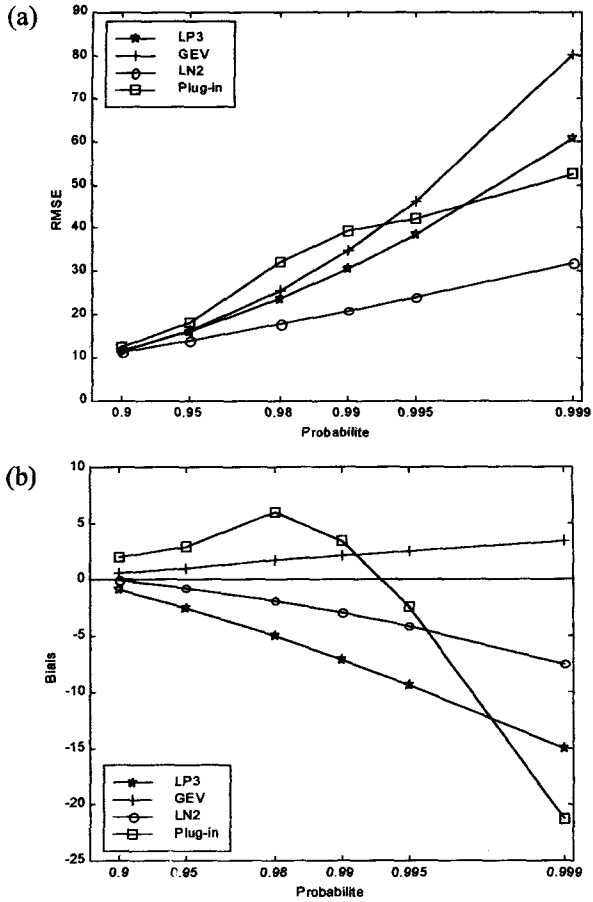

Figure 7 Comparaison de la méthode *plug-in* à l'ajustement la LP3, la GEV et la LN2 pour n = 50 : (a) RMSE ; (b) biais.

Comparison of quantile estimates obtained by the plug-in method and the LP3, GEV and LN2 distributions (n = 50): (a) RMSE; (b) bias.

Les distributions à deux paramètres sont généralement caractérisées par une faible variance des quantiles estimés. En revanche, les lois de ce type sont moins flexibles que les lois à plus de deux paramètres, provoquant souvent un biais important dans l'estimation des quantiles élevés. Ainsi, on observe un RMSE relativement faible pour la LN2 par rapport à la GEV, la LP3 et à la méthode *plug-in* dans la présente étude de simulation. Apparemment, la LN2

s'ajuste bien aux données simulées à partir de la LP3 puisque même le biais est relativement faible. Toutefois, la LN2 ne s'ajuste pas aussi bien à tous les échantillons. Un autre ensemble de paramètres pour la loi parente pourrait mener à des résultats différents avec la LN2. Dans ce sens, les méthodes non paramétriques ont l'avantage d'être beaucoup plus flexibles que les méthodes paramétriques et d'être applicables pour tous les types de données.

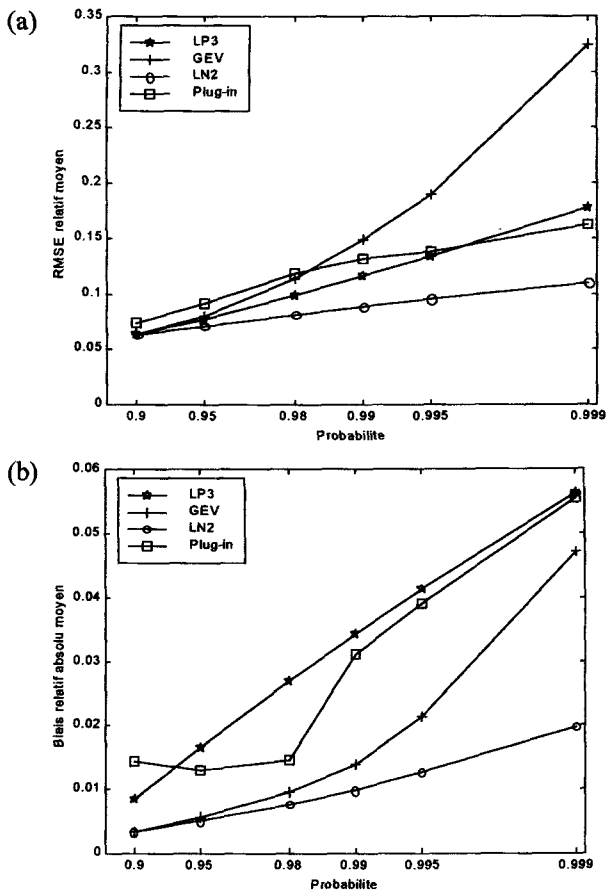

Figure 8 Comparaison de la méthode *plug-in* à l'ajustement de la LP3, la GEV et la LN2 pour n = 10, 25, 50 et 100 : (a) RMSE relatif moyen ; (b) biais relatif absolu moyen.

Comparison of quantile estimates obtained by the plug-in method and the LP3, GEV and LN2 distributions for n = 10, 25, 50 and 100: (a) RMSE; (b) bias.

Il a été discuté précédemment que la taille d'échantillon pouvait jouer un rôle important sur la précision de l'estimation, que ce soit avec la méthode des noyaux ou avec les méthodes paramétriques. Le même type de comparaison que celle présentée à la *figure 7* a donc été effectuée pour des tailles d'échantillon de n = 10, 25 et 100. Les résultats pour n = 10, 25, 50 et 100 ont été com-

binés et présentés sous forme de RMSE relatif moyen et de biais relatif absolu moyen. Ainsi, les deux statistiques ont été relativisées par rapport aux valeurs théoriques des quantiles et on a considéré la moyenne des résultats obtenus pour les quatre tailles d'échantillon. Les résultats — illustrés sur la *figure 8a* — sont sensiblement les mêmes que ceux de la *figure 7a*. Sur la *figure 8b*, on remarque que le biais minimal est atteint avec la LN2 et que le biais de la méthode *plug-in* est relativement comparable à celui de la LP3. En examinant individuellement les figures de biais pour chacune des tailles d'échantillon (non présentées), on a remarqué que l'amplitude du biais moyen de la LP3 provient des échantillons de faible taille. Pour la LP3, le biais diminue de façon importante avec l'augmentation de la taille de l'échantillon, étant relativement élevé pour $n = 10$ et 25, mais faible pour $n = 50$ et presque nul pour $n = 100$. Le biais de la LN2 demeure assez constant avec l'augmentation de la taille de l'échantillon alors que le biais des autres méthodes diminue lorsque l'on accroît n. À noter que la LN2 est la seule des méthodes qui sous-estime systématiquement les quantiles pour les échantillons de grande taille ($n = 100$). En somme, en terme de RMSE, la méthode *plug-in* est comparable à la LP3 et à la LN2 et plus efficace que la GEV. Le biais de la méthode *plug-in* peut paraître élevé par rapport à la GEV ou à la LN2, mais le maximum atteint pour $T = 1\,000$ ans est relativement faible puisqu'il est inférieur à 6 %. Ainsi, les résultats obtenus avec la méthode des noyaux sont relativement comparables à ce qui a été obtenu par les méthodes paramétriques classiques.

4.6 Application aux données de la Rivière rouge

Une comparaison rigoureuse de l'efficacité de différentes méthodes d'estimation nécessite généralement la simulation des données artificielles afin de pouvoir calculer les caractéristiques d'échantillonnage des estimateurs. Cependant, il est aussi d'intérêt d'appliquer les méthodes à des données réelles afin d'examiner leur comportement dans un contexte plus pratique. Nous avons donc examiné les débits maximums annuels de la Rivière rouge, observés à Emerson près de la frontière séparant le Dakota du Nord et le Manitoba. Les observations disponibles débutent en 1913. Nous avons considéré la période d'observation 1913-1997, 1997 étant l'année où la « crue du siècle » a eu lieu dans le bassin. Nous avons entre autre voulu comparer le comportement des méthodes paramétriques et non paramétriques en présence d'une valeur singulière.

Les distributions paramétriques présentées au *tableau 2* ainsi que la méthode non paramétrique *plug-in* ont été appliquées à ces données. Le résultat des ajustements est présenté aux *figures 9a* et *9b* où l'échantillon est considéré avec et sans l'année 1997, respectivement. Les lois LP3 et GEV donnent sensiblement les même résultats pour les périodes de retour inférieures à 100 ans. La distribution non paramétrique est très similaire aux deux lois à trois paramètres pour les périodes de retour entre 1 et 25 ans. Après 25 ans, on voit l'influence des deux ou trois valeurs extrêmes sur la distribution non paramétrique. Celle-ci a tendance à suivre davantage les observations que les lois paramétriques. On peut aussi noter que la loi log normale est assez similaire au modèle non paramétrique pour les grandes périodes de retour. En revanche, en raison de son manque de flexibilité, la log normale ne s'ajuste pas bien aux données associées aux périodes de retour de 10 à 25 ans.

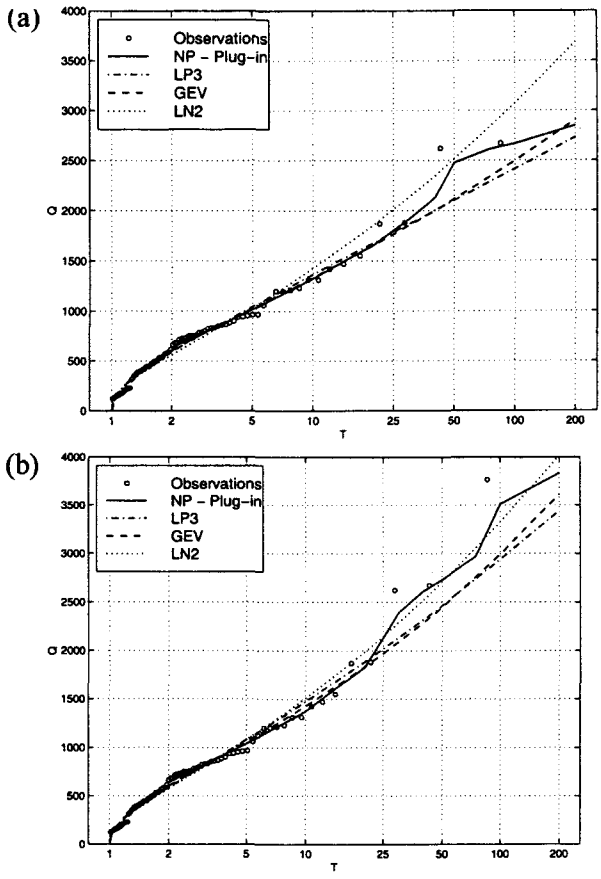

Figure 9 Ajustements des lois paramétriques et distributions non paramétriques aux crues maximums annuelles de la Rivière rouge à Emerson : (a) 1913-96 ; (b) 1913-97. 1997 était l'année de la « crue du siècle ».

Fitting of parametric and non parametric distributions to annual peak flows of the Red River at Emerson: (a) 1913-96 ; (b) 1913-97. 1997 was the year when the "flood of the century" occurred in the Red River basin.

En raison de l'ignorance de la véritable distribution de crues annuelles, il est difficile de déterminer laquelle des méthodes donne les meilleurs résultats. Les graphiques donnent l'impression que les lois à trois paramètres sous-estiment la fréquence de crues extrêmes. En revanche, on pourrait argumenter que la distribution non paramétrique suit trop les grandes valeurs de l'échantillon et conduit à une surestimation de la fréquence de crue. La vérité se retrouve probablement quelques part entre les deux. On doit insister sur le fait que l'utilisation des distributions non paramétriques ne résout pas le problème lié au manque d'information sur le comportement statistique des événements extrêmes, puisque ce problème est inhérent à l'analyse fréquentielle des crues.

5- CONCLUSIONS

Les principales conclusions de cette étude sont :

1. Même dans un contexte d'extrapolation, l'influence du type de noyau est bien moindre que le choix du paramètre de lissage. Toutefois, certains noyaux ne peuvent être recommandés pour l'extrapolation des quantiles de crue. Le noyau Cauchy, un noyau à queues lourdes, conduit à une surestimation considérable, alors que le degré d'extrapolation du noyau rectangulaire est relativement limité. L'utilisation d'un noyau à asymétrie positive permet d'améliorer l'estimation pour les quantiles d'ordre supérieur. On note d'ailleurs qu'il existe des méthodes permettant l'estimation simultanée de l'ordre d'un noyau et du paramètre de lissage, par exemple VIEU (1999).

2. Comme c'est le cas avec les méthodes paramétriques, la qualité des estimations non paramétriques décroît lorsque la taille d'échantillon décroît. Cependant, les méthodes non paramétriques sont plus robustes en terme de RMSE (racine carrée de l'erreur quadratique moyenne) pour les petites tailles d'échantillon, c'est-à-dire que le RMSE des quantiles estimés décroît moins rapidement que celui de méthodes paramétriques.

3. L'étude de sept méthodes d'estimation du paramètre de lissage conduit à des résultats à la fois intéressants et décevants. D'abord, la méthode d'ADAMOWSKI qui a été fréquemment citée en hydrologie se révèle être considérablement biaisée. En fait, il n'y aucune convergence du paramètre de lissage vers une valeur suffisamment grande pour qu'il y ait véritablement un lissage. Ainsi, cette méthode doit être abandonnée. La méthode CQA et l'adaptation à fenêtre variable de la méthode des moindres carrés, sont des méthodes pour lesquelles l'estimation est considérablement variable et biaisée. En revanche, la méthode *plug-in*, récemment introduite pour l'estimation des quantiles de crue (FAUCHER *et al.*, 2001) s'avère être relativement efficace. Elle demeure préférable à la méthode MC-VC du fait qu'elle permet d'éviter le problème rencontré pour les échantillons formés de données discrètes ou arrondies.

4. Les résultats décevants obtenus avec les méthodes à fenêtre variable nous indiquent qu'il est préférable de considérer un paramètre de lissage global plutôt que de conditionner la valeur de celui-ci sur la distance aux autres observations. Cette affirmation a été confirmée par le fait que les meilleurs résultats ont été obtenus avec la méthode conduisant à des paramètres locaux peu variables. Les méthodes à fenêtre variable considérées dans cette étude sont des méthodes qui permettent de faire varier le paramètre de lissage en se basant sur la distance au k^e voisin le plus proche. Une autre façon de procéder serait de considérer un paramètre de lissage qui soit variable selon la densité locale (SIMONOV, 1998). Ainsi, dans l'expression (25), au lieu de considérer $h(x_i) = a_k d_{ki}$ on pourrait plutôt considérer $h(x_i) = h_l \tilde{f}(x_i)^{-1/2}$. Le paramètre h_l est le paramètre de lissage qui doit être optimisé et la densité $\tilde{f}(x_i)$ est une densité « pilote » qui peut être estimée à partir d'une méthode à noyau fixe par exemple. Comme le paramètre de lissage est inversement proportionnel à la densité locale, l'amplitude de $h(x_i)$ dépend véritablement de la densité locale des observations.

5. Il est difficile d'évaluer de manière générale la performance des méthodes d'estimation du paramètre de lissage sur la fonction de densité par

rapport à celles considérant la fonction de distribution. À un niveau théorique, il est convenable de supposer que les estimateurs de la fonction de densité ne conduisent pas à un paramètre de lissage optimal lorsqu'on se trouve dans un contexte d'estimation de quantile de crue (LALL *et al.*, 1993). En revanche, il n'a pas été possible en pratique de confirmer cette affirmation. Dans la présente étude, les méthodes d'estimation de F n'étaient représentées que par la méthode *plug-in*, la méthode d'Adamowski ayant été rejetée, alors que les cinq autres méthodes étaient basées sur l'estimation de la fonction de densité f. Malgré le fait que la méthode *plug-in* s'est avérée être plus efficace que les méthodes basées sur f, il est difficile de généraliser cette conclusion.

6. La méthode des noyaux constitue une alternative intéressante aux méthodes paramétriques traditionnelles pour l'estimation de quantiles de crue. Elle est relativement simple d'utilisation puisqu'il ne s'agit que de sélectionner la fonction noyau et de calculer le paramètre de lissage qui détermine le degré d'influence des observations pour l'estimation. La méthode des noyaux permet d'éviter de faire le choix souvent subjectif de la distribution de la population étudiée. Les données synthétiques qui ont servi aux diverses comparaisons effectuées dans cette étude proviennent d'une population distribuée selon la loi LP3. Ainsi, l'estimation est effectuée dans des conditions qui sont plutôt défavorables aux méthodes paramétriques. Les résultats obtenus avec la méthode des noyaux, en particulier pour les méthodes *plug-in* et MC-VC, sont relativement comparables à ceux obtenus avec les méthodes paramétriques. En effectuant le même type de comparaison sur des données provenant d'une distribution multimodale par exemple ou bien du mélange de plusieurs distributions, il est probable que la méthode des noyaux aurait gagné de la précision par rapport aux trois distributions considérées.

RÉFÉRENCES BIBLIOGRAPHIQUES

ADAMOWSKI K., 1989. A Monte Carlo comparison of parametric and nonparametric estimation of flood frequencies. *J. Hydrol.*, 108, 295-308.

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

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

ADAMOWSKI K., FELUCH W., 1983. Application of pattern analysis to flood frequency determination. *American Geophysical Union, EOS Transactions,* 64 (45), 1035-1047.

ALTMAN N., LÉGER C., 1995. Bandwidth selection for kernel distribution function estimates. *J. Statist. Plann. Inference,* 46, 195-214.

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

BREIMAN L., MEISEL W., PURCELL E., 1977. Variable kernel estimates of multivariate densities. *Technometrics,* 19 (2), 135-144.

CACOULLOS T., 1966. Estimation of a multivariate density. *Annals of the Instiute of Statistical Mathematics,* 18, 178-189.

DUIN R.P.W., 1976. On the choice of smoothing parameters for Parzen estimators of probability density functions. *IEEE Trans. Comput,* C-25, 1175-1179.

EPANECHNIKOV V.A., 1969. Nonparametric estimation of a multidimensional probability density. *Theory of Probability and its Applications,* 14, 153-158.

FAN J., HALL P., MARTIN M.A., PATIL P., 1996. On local smoothing of nonparametric curve estimator. *J. Am. Stat. Assoc.,* 91 (433), 258-266.

FAUCHER D., 1999. Estimation non paramétrique des quantiles de crue par la méthode des noyaux. Mémoire de Maîtrise INRS-Eau, Université du Québec, 164 p.

FAUCHER D., RASMUSSEN P.F., BOBÉE B., 2001. A distribution function based bandwidth selection method for kernel quantile estimation. *J. Hydrol.,* 250, 1-11.

HABBEMA J.D.F., HERMANS J., BROEK V.D., 1974. A stepwise discrimination program using density estimation, *In:* G. Bruckman, *Compstat 1974* , Physica Verlag, Viennes, 100-110.

HALL P., MARRON J.S., 1987. Estimation of integrated squared density derivatives. *Stat. Probab. Lett.,* 6, 109-115.

LALL U., MOON Y.I., BOSWORTH K., 1993. Kernel flood frequency estimators: bandwidth selection and kernel choice. *Water Resour. Res.,* 29 (4), 1003-1015.

MOON Y.-I., LALL U., BOSWORTH K., 1993. A comparison of tail probability estimators for flood frequency analysis. *J. Hydrol.,* 151, 343-363.

RAO P.B.L., 1983. *Nonparametric Functionnal Estimation.* Academic Press, NewYork.

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.

SILVERMAN B.W., 1986. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, New York, 175 p.

SIMONOV J.S., 1998. *Smoothing Methods in Statistics*. Springer Verlag, 2e édition, 338 p.

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

VIEU P., 1999. Multiple kernel procedure: An asymptotic support. *Scand. J. Statist.*, 26, 61-72.

YAKOWITZ S.J., 1983. Some "model-free" techniques for flood frequency analysis. *American Geophysical Union, EOS Transactions,* 64 (45), 706.