Université du Québec
INRS-Eau

MODÉLISATION ADDITIVE PAR POLYNÔMES LOCAUX POUR LA RÉGIONALISATION DES QUANTILES DE CRUE : APPROCHE OPTIMALE DE RÉGRESSION PAR RÉGION D'INFLUENCE

Par
Marco Latraverse
M.Sc. Mathématiques

Thèse présentée
pour l'obtention
du grade de Philosophiæ doctorat (Ph.D.)
en Sciences de l'eau

Jury d'évaluation

Président du jury et examinateur externe	Pascal Sarda Laboratoire de statistique et probabilités Université Paul Sabatier
Examinateur externe	Belkacem Abdous Département de mathématiques et d'informatique Université du Québec à Trois-Rivières
Examinateur interne	Taha B.M.J. Ouarda INRS-Eau Université du Québec
Codirecteur de recherche	Peter Funder Rasmussen Department of Civil and Geological Engineering University of Manitoba
Directeur de recherche	Bernard Bobée INRS-Eau Université du Québec

Thèse soutenue le 14 avril 2000

Résumé

En raison des grandes étendues territoriales et du coût associé à l'installation et au maintien de stations de mesures, il arrive fréquemment que les hydrologues doivent produire une estimation des quantiles de crue de période de retour donnée T, notés Q_T, en un site non jaugé où ils ne disposent d'aucune information hydrométrique. Dans cette situation, une approche employée consiste à utiliser des procédures de régionalisation afin de transférer l'information disponible en des sites jaugés vers le site non jaugé où l'on désire produire une estimation. De manière générale, une procédure de régionalisation comporte deux étapes distinctes qui consistent à (1) choisir les sites jaugés à partir desquels s'effectuera le transfert d'information et (2) appliquer aux sites choisis un modèle de transfert d'information régionale. Aux États-Unis, des ingénieurs du *United States Geological Survey* (USGS) ont proposé récemment l'utilisation d'une nouvelle procédure de régionalisation, **la régression par région d'influence** (Tasker et al., 1996) qui consiste à (1) choisir les sites jaugés à l'aide d'une approche de région d'influence et (2) utiliser une approche classique de régression pour le transfert d'information régionale.

Dans cette thèse, nous reformulons l'approche de la régression par région d'influence dans un contexte de régression non paramétrique. Nous montrons en effet que l'approche de régression par région d'influence appartient à une classe particulière de modèles de régression non paramétrique, les modèles de régression locale. Nous utilisons, dans un premier temps, les concepts de la régression locale afin de proposer une approche permettant de choisir de manière objective et optimale les paramètres de la région d'influence, des paramètres qui traditionnellement étaient choisis de manière subjective. Nous mettons aussi en évidence le fait qu'appartenant à la famille des modèles de régression locale, l'approche de la régression par région d'influence se heurte au problème, bien connu en statistique, de la raréfaction des données dans un espace à grande dimension. Nous proposons alors, par le fait même, l'utilisation d'une nouvelle approche de régionalisation non paramétrique par région d'influence qui ne soit pas affectée par ce problème de dimensionalité, **l'approche de la modélisation additive par polynômes locaux**. Nous comparons par la suite, à l'aide de données réelles, l'approche de modélisation additive par polynômes locaux à l'approche de la régression par région d'influence dont les paramètres ont été estimés de manière objective et optimale, à l'approche de la régression par région d'influence de Tasker et al. (1996) et à l'approche classique de régression régionale paramétrique du USGS.

Remerciements

Plusieurs personnes m'ont soutenu pendant mes études de doctorat. Mentionnons d'abord mon directeur de thèse, Bernard Bobée, ainsi que mon codirecteur, Peter Rasmussen. Je tiens aussi à remercier les autres membres de l'équipe de la Chaire industrielle en hydrologie statistique CRSNG / Hydro-Québec / INRS-Eau pour leurs commentaires, pour leur aide et pour avoir agrémenté mon séjour à l'INRS-Eau. Merci Taha, Luc, Mario, Hugues, Alin, Fabrice et Martyne. Il me faut aussi souligner que les commentaires des examinateurs externes ont permis d'améliorer de façon significative la qualité de ce document.

Les trois premières années de mes études ont été financées par les bourses de l'INRS et par l'INRS-Eau via mon directeur de recherche. Je tiens enfin à remercier ma conjointe Marie-Natacha d'abord pour son soutien financier mais surtout pour m'avoir supporté, appuyé et avoir su être présente, malgré son horaire chargé, lorsque j'avais besoin d'elle.

À la mémoire de Georges-Étienne Latraverse

Table des matières

Liste des figures

Liste des tableaux

1. INTRODUCTION

1.1 Problématique de recherche

La gestion et l'utilisation des eaux de surface nécessitent une connaissance des écoulements aux sites où se posent les problèmes reliés à la prévision des crues, au contrôle des inondations, à la régularisation des cours d'eau, au dimensionnement des ouvrages hydrauliques et d'une façon générale, à toute gestion de la ressource hydrique et à toute étude d'impact d'aménagement d'un cours d'eau. Pour résoudre ces différents problèmes, les hydrologues doivent (1) prévoir l'évolution des débits d'une rivière et (2) estimer l'amplitude et la fréquence de débits extrêmes (crues, étiages). Par exemple, pour l'établissement de systèmes d'alerte en cas d'inondation, on doit pouvoir émettre des prévisions à court terme de l'évolution du niveau des cours d'eau. Pour tous les problèmes reliés à la gestion des réservoirs, on doit aussi prévoir, non seulement à court mais aussi à moyen et long terme, l'évolution du débit afin de gérer efficacement le niveau des réservoirs. Par contre, pour la planification, incluant la conception, des ouvrages de contrôle des inondations ou de tout autre ouvrage soumis au risque de défaillance par les eaux, les hydrologues sont plutôt amenés à effectuer des estimations soit (1) de la **probabilité au dépassement**, aussi appelée la **probabilité** ou la **fréquence** d'une crue, c'est-à-dire la probabilité qu'une crue de grandeur donnée se produise ou soit dépassée au cours d'une année donnée ou (2) de l'amplitude de la crue associée à une probabilité au dépassement préalablement définie. Dans ce travail, nous nous intéressons spécifiquement au problème de l'estimation des amplitudes de crue.

Les hydrologues disposent d'une abondance d'approches et de méthodes d'**analyse de fréquence des crues** (AFC) permettant d'effectuer l'estimation des amplitudes de crue. Les différentes procédures d'AFC peuvent être regroupées en fonction, entre autres, (Cunnane, 1987) :

1. du nombre et de la nature des données disponibles :
   (a) données au site jaugé seulement, ou
   (b) données au site jaugé et données régionales, ou
   (c) données régionales seulement ; et
2. du type de modèle utilisé :
   (a) modèle des séries de débits maximums annuels (DMA), ou
   (b) modèle des séries de dépassements (ou séries partielles).

Dans le cadre de ce travail, nous proposons une nouvelle méthode d'AFC basée sur l'utilisation

du modèle des séries de DMA (2a) et s'attaquant spécifiquement au problème de l'estimation des amplitudes de crue en des sites non jaugés (1c).

Avec l'approche des séries de DMA, aussi appelée **approche des crues annuelles**, on ne considère qu'une seule crue par année correspondant à la valeur maximale du débit atteinte au cours de cette année. On fait ensuite l'hypothèse que le DMA d'une année quelconque est une variable aléatoire X issue d'une population P ayant une fonction de densité de probabilité f ainsi qu'une fonction de distribution stationnaire F telle que $F(x) = Pr[X \leq x]$. Selon cette approche, l'amplitude Q_p du débit de la crue ayant une probabilité préalablement définie p d'être dépassée au cours d'une année donnée correspond au $(1 - p)$ième percentile, aussi appelé **quantile** de la distribution des DMA soit

$$Q_p = F^{-1}(1 - p) \tag{1.1}$$

où $F^{-1}(.)$ est la fonction de répartition inverse de la variable aléatoire X. En hydrologie, il est courant de mesurer l'amplitude des crues en employant la notion de **période de retour**, notée T. Par définition, une crue de période de retour de T années (*T-year flood*), notée Q_T, correspond au débit de la crue qui est dépassé, en moyenne (sur une longue période) une fois toutes les T années. Avec l'approche des crues annuelles, il est possible de montrer que le débit de crue Q_T correspond à un débit dont la probabilité au dépassement est $p = 1/T$ d'où la relation

$$Q_T = Q_p = F^{-1}(1 - p) = F^{-1}(1 - 1/T) \tag{1.2}$$

et l'emploi du terme **quantile de crue** pour désigner la quantité Q_T.

Dans le cas où des mesures de DMA sont disponibles au site où l'on désire effectuer une estimation, le problème de l'estimation du quantile de crue Q_T consiste alors à estimer la distribution réelle mais inconnue F des DMA ou crues annuelles ou tout simplement, crues. De manière générale, l'**analyse (locale) de la distribution des crues** (ADC), consiste à ajuster une distribution statistique $f(x; \underline{\Theta})$ à la série $(x_1, x_2, ..., x_n)$ des DMA (où n est le nombre d'années d'enregistrements). Une fois l'estimation par $\underline{\hat{\Theta}}$ du vecteur de paramètres $\underline{\Theta}$ de la loi f effectuée, l'équation 1.2 permet d'obtenir

$$\hat{Q}_T = F^{-1}(1 - 1/T; \underline{\hat{\Theta}}) \tag{1.3}$$

En hydrologie statistique, plusieurs distributions combinées à diverses méthodes d'estimation des paramètres ont été utilisées pour estimer de manière approximative la distribution réelle des crues annuelles.

En raison des grandes étendues territoriales et du coût associé à l'installation et au maintien de stations de mesures, il arrive fréquemment que les hydrologues doivent produire une estimation de Q_T en un site où ils disposent de peu ou d'aucune information hydrométrique. Dans ces situations, il est alors possible d'effectuer une **analyse régionale de la distribution des crues** (ARDC) afin de s'attaquer à cette problématique particulière. Un projet de recherche a été mené récemment par une équipe de scientifiques canadiens (GREHYS, 1996a, 1996b) afin de comparer les principales méthodes d'ARDC utilisées en Amérique du Nord tant pour l'estimation en des sites partiellement jaugés qu'en des sites non jaugés. Puisque dans ce travail nous proposons une nouvelle approche d'estimation pour des sites **non jaugés**, nous référons le lecteur à (GREHYS, 1996a, 1996b) pour une explication détaillée de la problématique de l'estimation en des sites partiellement jaugés.

L'ARDC consiste à utiliser des procédures dites de régionalisation pour transférer l'information disponible en des sites jaugés vers le site non jaugé (site cible) où l'on désire effectuer une estimation. De manière générale, une procédure de régionalisation comporte deux étapes distinctes qui consistent à (1) choisir les sites jaugés à partir desquels s'effectuera le transfert d'information vers le site cible et (2) appliquer aux sites choisis un modèle de transfert d'information régionale.

1.1.1 Le modèle de transfert régional

Un modèle de transfert d'information régionale ou tout simplement, un modèle régional, est un ensemble d'équations qui relient entre elles une ou plusieurs caractéristiques de crue d'une région. Dans la littérature hydrologique, deux procédures de régionalisation ont plus particulièrement retenu l'attention des hydrologues : la méthode de l'indice de crue (Dalrymple, 1960) et la méthode de la régression régionale des quantiles (RRQ) (Benson, 1962).

Avec la méthode de l'indice de crue, on fait l'hypothèse que les données aux différents sites i d'une région sont indépendantes et suivent la même distribution statistique à un facteur d'échelle près, généralement le débit de crue moyen μ_i au site i (l'indice de crue). Le modèle régional intrinsèque à la méthode est alors (pour une période de retour T fixée) :

$$\frac{Q_T^i}{\mu_i} = \beta_0 + \epsilon_i \tag{1.4}$$

où Q_T^i/μ_i est le quantile de crue normalisé du site i, β_0 est une constante représentant le quantile de crue normalisé régional moyen et ϵ_i est la composante d'erreur du modèle. L'hypothèse que les données aux différents sites d'une région sont indépendantes et suivent la même distribution

statistique à un facteur d'échelle près équivaut à supposer que le coefficient de variation (CV) et tous les autres ratios de moments d'ordre supérieur sont égaux à chacun des sites régionaux. Ainsi, un autre modèle régional intrinsèque à la méthode de l'indice de crue est

$$CV_i = \overline{CV} + \epsilon_i \tag{1.5}$$

où CV_i est le CV du site i, $\overline{CV}$ représente le CV régional moyen et ϵ_i est la composante d'erreur du modèle.

Avec la méthode de la RRQ, on fait plutôt l'hypothèse que le modèle log-linéaire suivant :

$$\log(Q_T^i) = \beta_0 + \beta_1 \log(X_1^i) + \beta_2 \log(X_2^i) + ... + \beta_d \log(X_d^i) + \epsilon_i \tag{1.6}$$

où $\beta_0, \beta_1, ..., \beta_d$ sont les paramètres inconnus du modèle et ϵ la composante d'erreur, permet de décrire adéquatement la relation existant entre les quantiles de crue Q_T et les variables physiographiques et climatologiques $X_1, X_2, ..., X_d$ des différents sites d'une région donnée. En supposant une telle relation régionale, il est alors possible d'appliquer les techniques de régression linéaire multiple traditionnelles pour l'estimation des paramètres du modèle de même que pour le choix des variables explicatives.

Puisqu'en pratique aucun modèle régional ne représente parfaitement la réalité hydrologique, toute procédure de régionalisation introduit, pour l'estimation de Q_T en un site non jaugé, outre des erreurs liées à l'estimation des paramètres du modèle, une erreur causée par la forme inexacte du modèle employé. Moss (1979) a d'ailleurs montré, à l'aide de simulations, comment l'élaboration d'un meilleur modèle permet d'obtenir, pour un nombre d'observations donné, des estimations beaucoup plus précises. Selon Greis et Wood (1981), le modèle de transfert constitue d'ailleurs l'élément majeur limitant l'amélioration des procédures de régionalisation :

> "The achievement of an optimal level of regional information transfer has been frustrated by the lack of a satisfactory methodology for regionalizing statistical flood parameters such as flood recurrence intervals or quantile levels. To a point, information transfer deficiencies can be lessened by more intensive data collection activities, but the ultimate limitation is the transfer model itself, usually regression."

1.1.2 Le choix des sites

Avant de pouvoir transférer au site cible l'information régionale provenant de sites jaugés, il est primordial de bien choisir les sites jaugés à partir desquels s'effectuera le transfert d'information régionale. À cette étape, l'hydrologue doit répondre à la question suivante : de quels

sites provient l'information permettant d'estimer de la meilleure façon possible le modèle de transfert applicable au site non jaugé ? En ARDC, le choix des sites s'effectue généralement en regroupant les différents sites en régions (pas nécessairement géographiques) en fonction (1) de l'homogénéité d'une de leurs caractéristiques hydrologiques (généralement le coefficient de variation (CV)) ou (2) de la similarité de leur comportement hydrologique. Dans la littérature hydrologique, les notions de similarité hydrologique et d'homogénéité régionale sont généralement confondues. Dans ce document, on distinguera cependant clairement ces notions à l'aide des définitions suivantes de Rossi et Villani (1994a) :

> "Regionalization models use the concept of hydrologic similarity by associating basin's flood characteristics with climatic and geomorphologic factors. As only some of the geomorphologic characteristics are used as explanatory variables, regionalization models usually allow the other factors to be constants, i.e. they are combined so as not to influence the flood characteristics. In this sense, the basins in a single region are defined to be *hydrologically similar*. The particular case where no basin factor is seen to affect the flood characteristics is defined as *regional homogeneity*, with regard to the flood characteristic considered, which is thus constant in the region."

Le choix des sites est avant tout une question de respect des hypothèses du modèle de transfert. Par exemple, pour la méthode de l'indice de crue dont le modèle régional fait l'hypothèse d'un coefficient de variation (CV) constant, il est souhaitable d'avoir **homogénéité régionale** au sens du CV et le choix des sites consiste alors à regrouper les sites en régions ayant des CV similaires appelées régions homogènes. Cependant, pour une méthode comme la RRQ, les sites servant à l'établissement du modèle régional doivent plutôt avoir un **comportement hydrologique similaire**, c'est à dire qu'une seule et même relation log-linéaire avec les mêmes paramètres et les mêmes variables explicatives doit s'appliquer à tous les sites retenus.

De manière générale, la détermination de régions homogènes ou de régions ayant un comportement hydrologique similaire s'effectue soit (1) en déterminant des régions distinctes (géographiques ou non) de sites jaugés ou (2) en déterminant, pour le site cible non jaugé, son propre ensemble de sites jaugés appelé voisinage. Les voisinages peuvent être définis par une analyse des corrélations canoniques (Ribeiro-Corréa et al., 1995; Ouarda et al., 1997; Ouarda, Haché, et Bobée, 1998) ou à l'aide de la méthode de la région d'influence (Burn, 1990a). Selon la méthode de la région d'influence proposée d'abord par Burn (1990a) pour des sites cibles jaugés puis par Zrinji et Burn (1994) pour des sites non jaugés, chaque site cible est considéré comme le centre de sa propre région. Le critère permettant alors la sélection de la région d'influence est

la distance euclidienne $d_{ij} = \sqrt{\sum_{l=1}^{n} w_l (X_l^{(i)} - X_l^{(j)})^2}$ dans l'espace des variables explicatives X_l où w_l est une fonction de pondération permettant la standardisation des variables explicatives. La région d'influence au site i est alors définie par l'ensemble des sites j tels que d_{ij} est inférieure à un certain point de coupure θ_L ou, lorsque ce point de coupure est trop restrictif, par l'ensemble des K sites ayant les K plus petites distances d_{ij} où K est le nombre minimum de stations désiré. Cette méthode a d'abord été proposée pour la détermination de régions homogènes avec comme motivation le fait qu'il soit raisonnable de s'attendre à ce que des bassins versants ayant des caractéristiques physiographiques et climatologiques similaires, au sens de la distance d_{ij}, puissent avoir des réponses pluie-débit similaires et ainsi, des distributions de crue semblables (même CV) (Burn, 1990a). Cependant, d'un point de vue conceptuel, cette méthode peut tout aussi bien s'appliquer dans un contexte de détermination de sites ayant un comportement hydrologique similaire.

1.1.3 Approches actuelles

Aux États-Unis, le *United States Geological Survey* (USGS), l'organisme impliqué dans le développement des procédures de régionalisation pour chacun des États américains a d'abord utilisé la méthode de l'indice de crue durant les années quarante et ce, jusqu'au début des années soixante. Cependant, après que des études de Dawdy (1961) et de Benson (1962) aient montré que pour plusieurs régions des États-Unis, l'hypothèse d'un CV constant ne pouvait s'appliquer puisqu'empiriquement, le CV tend à diminuer lorsque l'aire des bassins versants augmente, le USGS a remplacé la méthode de l'indice de crue par celle de la régression régionale des quantiles. Cette méthode est encore employée aujourd'hui par le USGS qui produit et publie régulièrement (voir par exemple le rapport national de Jennings et al. (1994)) des équations de régression permettant aux ingénieurs et hydrologues d'estimer facilement et rapidement les quantiles de crue en chacun des sites non jaugés des États-Unis.

Durant les années quatre-vingt, la méthode de l'indice de crue a été ressuscitée dans une version utilisant les L-moments de Hosking et al. (1985a). De nombreuses études de type Monte Carlo (Hosking et al., 1985a; Lettenmaier, 1985; Lettenmaier et Potter, 1985) ont alors montré que cette version de la méthode améliore la précision des estimateurs de Q_T (en terme d'erreur quadratique moyenne) aux sites jaugés d'une région homogène. Par contre, l'hétérogénéité régionale, c'est-à-dire le non respect de l'hypothèse d'un CV régional constant, nuit à la performance de ces estimateurs (Lettenmaier et Potter, 1985). Par le fait même, l'application de cette méthode pour l'estimation en des sites non jaugés demeure problématique puisqu'il est souvent difficile d'assigner à un site non jaugé une région homogène (Fill et Stedinger, 1998).

Au cours des dernières années, de nouveaux modèles régionaux sont apparus dans la littérature afin de combler certaines des lacunes de leurs prédécesseurs. Par exemple, réalisant que l'hypothèse de base de la méthode de l'indice de crue est inconsistante avec les relations connues entre le CV et l'aire des bassins versants, Fill et Stedinger (1998) indiquent que les méthodes d'estimation régionale devraient faire intervenir davantage la dépendance des quantiles de crue avec l'aire des bassins versants et les autres caractéristiques physiographiques importantes. Ainsi, ils proposent l'utilisation de la méthode de la régression des quantiles normalisés, développée par Fill (1994), qui consiste à combiner les modèles de l'indice de crue et de la régression régionale des quantiles afin d'obtenir un modèle dont la forme est :

$$\log(\frac{Q_T^i}{\mu_i}) = \beta_0 + \beta_1 \log(X_1^i) + \beta_2 \log(X_2^i) + ... + \beta_d \log(X_d^i) + \epsilon_i \tag{1.7}$$

Ce modèle permet d'appliquer la notion de *scaling* propre à la méthode de l'indice de crue en des sites ne respectant pas nécessairement l'hypothèse de constance régionale du quantile normalisé (ou CV) puisque celui-ci se trouve modélisé à l'échelle régionale par régression linéaire. On peut remarquer que cette approche constitue une généralisation de la méthode de l'indice de crue puisqu'en présence d'un CV constant, l'estimation devrait conduire à $\beta_1 = \beta_2 = ... = \beta_d = 0$.

Considérant qu'il n'existe aucune justification physique à la sélection d'une relation linéaire multiple entre le logarithme des quantiles de crue Q_T et les différentes variables physiographiques et climatologiques des bassins versants, Gingras et al. (1995) proposent plutôt l'utilisation de la régression non paramétrique pour la modélisation de la relation régionale. La régression non paramétrique est une technique de modélisation dont la caractéristique principale est qu'aucune hypothèse n'est faite a priori sur la forme finale du modèle. Le modèle de régression non paramétrique s'exprime comme suit :

$$\log(Q_T^i) = s(X_1^i, X_2^i, ..., X_d^i) + \epsilon_i \tag{1.8}$$

où s est une fonction (courbe ou surface) de régression lisse de forme non spécifiée a priori de dimension d et ϵ_i est la composante d'erreur. La forme du modèle se détermine à partir des observations. De nombreuses méthodes ou techniques dites de lissage sont disponibles dans la littérature statistique afin de permettre l'estimation de s. Gingras et al. (1995) ont estimé le modèle de l'équation 1.4 à l'aide d'une de ces techniques, le lissage par noyau, pour des fonctions de régression de dimension $d = 1$ (courbe de régression) et $d = 2$ (surface de régression).

Lors d'une étude récente du USGS, Tasker et Slade (1994) ont proposé l'utilisation d'une approche dite interactive de régression régionale pour l'estimation de Q_T en des sites non jaugés du Texas. Il s'agit en fait d'estimer en chacun des sites non jaugés du Texas une équation de régression régionale à l'aide de la seule information provenant d'une région d'influence composée d'exactement 50 sites sur les 251 disponibles. Cette approche a permis d'obtenir de bien meilleurs résultats que l'approche traditionnelle employée aux États-Unis consistant à séparer l'État en régions hydrologiques géographiques à l'aide d'une analyse du signe des résidus des équations de régression. Plus récemment, Tasker et al. (1996) ont testé plus exhaustivement l'utilisation de cette procédure de régionalisation qu'ils nomment désormais **la méthode de la régression régionale par région d'influence**. Ils ont comparé l'utilisation de cette approche, pour la modélisation du quantile Q_{50} en Arkansas, aux autres procédures utilisées traditionnellement soit (1) l'ajustement d'un modèle de régression log-linéaire à tous les sites de l'État (2) l'ajustement de modèles log-linéaires différents pour chacune des sous-régions hydrologiques géographiques de cet État et (3) l'ajustement de modèles log-linéaires en des sous-régions identifiées selon des critères de similarité des caractéristiques de leurs bassins versants à l'aide de techniques multivariées d'analyse de regroupements (*cluster analysis*) et d'analyse discriminante (Wiltshire, 1986a). Encore une fois, les meilleurs résultats ont été obtenus par la méthode de la région d'influence.

Les résultats obtenus par Tasker et Slade (1994) et par Tasker et al. (1996) indiquent que l'approche de la régression régionale par région d'influence semble être des plus prometteuses. De plus, cette méthode fait intervenir, comme le suggèrent fortement Fill et Stedinger (1998), la dépendance entre les quantiles de crue et les caractéristiques physiographiques/climatologiques importantes. Enfin, cette méthode est robuste au non-respect de l'hypothèse de linéarité du modèle de régression (Tasker et Slade, 1994), une problématique soulevée par Gingras et al. (1995). D'ailleurs, cet avantage de l'approche de la régression régionale par région d'influence est dû au fait que la combinaison d'un modèle de régression log-linéaire et d'une approche de région d'influence revient à une approche de régression non paramétrique connue en statistique sous le nom de **régression locale** (Cleveland et Loader, 1996a) et dont un des modèles les plus populaires, le *LOESS* (Cleveland, 1979), s'est avéré être en hydrologie une méthode non paramétrique de choix pour l'analyse des tendances et la recherche de dépendances entre deux variables (Lall, 1995).

Bien que l'approche de la régression régionale par région d'influence soit des plus intéressantes, il subsiste toujours quelques facteurs susceptibles de limiter son application :

- On note une grande part de subjectivité en ce qui concerne le choix des paramètres de la région d'influence (le point de coupure θ_L, le nombre minimum K de stations désiré). Nguyen et Pandey (1996) reprochent d'ailleurs aux méthodes permettant le choix des sites, telle la méthode de la région d'influence, d'être développées pour l'analyse de la similarité hydrologique en utilisant des critères qui ne soient pas directement reliés aux objectifs de l'estimation des crues soit, par exemple, de produire une estimation des quantiles qui soit la plus précise possible.
- Puisque le modèle de régression régionale par région d'influence appartient à la famille des modèles de régression locale, il se heurte au problème, bien connu en statistique, de la raréfaction des données dans un espace à grande dimension. Ce problème entraîne un accroissement de la variance des estimateurs non paramétriques de la fonction de régression lorsque le nombre d de variables prédictives augmente. Par exemple, pour un échantillon de taille n, la vitesse optimale de convergence (en moyenne quadratique) pouvant être atteinte pour l'estimation non paramétrique d'une fonction de régression f de classe C_2 (i.e. pour laquelle on suppose l'existence et la continuité des 2 premières dérivées) est de l'ordre de $O(n^{-4/(4+d)})$ (Vieu, 1996). Dans le contexte hydrologique, cette problématique est d'autant plus importante en raison de la faible taille des échantillons (nombre de sites jaugés) généralement disponibles.
- Une autre difficulté relative à la régression locale multidimensionnelle (plusieurs variables explicatives) est la difficulté d'interprétation de la surface de régression. Il est en effet impossible de visualiser la surface de régression estimée lorsque $d > 2$.

1.1.4 L'approche de modélisation proposée

Dans le cadre de cette thèse, nous proposons, à l'instar de Gingras et al. (1995) et pour les mêmes raisons, l'utilisation d'une approche de régression non paramétrique, c'est-à-dire pour laquelle on ne fait aucune hypothèse a priori sur la forme de la fonction de régression à estimer, pour la modélisation régionale des quantiles de crue. Cependant, et ce en raison principalement du problème de dimensionalité associé à l'estimation par lissage pur d'une fonction multidimensionnelle $s(X_1^i, X_2^i, ..., X_d^i)$, nous proposons plutôt l'emploi du **modèle additif** (Hastie et Tibshirani, 1987; Buja et al., 1989; Hastie et Tibshirani, 1990) suivant :

$$\log(Q_T^i) = s_1(X_1^i) + s_2(X_2^i) + ... + s_d(X_d^i) + \epsilon_i \tag{1.9}$$

où les s_i sont des fonctions unidimensionnelles lisses, de forme non spécifiée a priori, associées respectivement à chacune des variables explicatives $(X_1, X_2, ..., X_d)$. Ce modèle se situe à mi-chemin entre la régression linéaire multiple (approche classique de Benson (1962) et du USGS) et le lissage de surfaces multidimensionnelles (approche de Gingras et al. (1995) et approche modifiée du USGS de Tasker et Slade (1994)). Ce modèle est caractérisé par le fait qu'il n'est pas affecté par le problème de dimensionalité. En effet, Stone (1985) a étudié la convergence d'estimateurs splines dans le modèle additif et a prouvé que ceux-ci atteignent la vitesse de convergence optimale pour un estimateur d'une fonction univariée ($d = 1$) et de classe C_2, c'est-à-dire de l'ordre de $O(n^{-4/5})$ (voir Stone (1982)). Une autre caractéristique de la modélisation additive qui distingue d'ailleurs le modèle additif des autres techniques de régression non paramétrique développées pour tenir compte du problème de dimensionalité (e.g. le modèle de régression sur directions révélatrices (*projection pursuit regression*) de Friedman et Stuetzle (1981), le modèle de transformations optimales utilisant l'algorithme ACE (*Alternating Conditional Expectation*) de Breiman et Friedman (1985), etc.) est qu'il est plus facile d'en interpréter les résultats. En effet, tout comme en régression linéaire multiple, il est possible avec un modèle additif d'examiner graphiquement l'effet de chacune des variables prédictives, une à la fois, conditionnellement à la présence des autres variables prédictives. Il devient alors possible d'effectuer une validation des hypothèses de linéarité associées au modèle de régression linéaire multiple.

De nombreuses techniques de lissage permettent l'estimation de chacune des fonctions s_i et le modèle additif ne requiert l'emploi d'aucune technique particulière. Nous proposons l'emploi d'une technique de lissage qui gagne en popularité tant au niveau des statisticiens en recherche fondamentale que des statisticiens en pratique, la **régression locale polynomiale** (RLP) (Loader, 1999). Avec la méthode de lissage par régression locale, on ne fait globalement aucune hypothèse quant à la forme de la fonction s. Toutefois, dans le voisinage d'un point d'estimation x, on suppose que s peut être estimée de manière approximative par une fonction paramétrique, soit un polynôme de degré p. En un point x, on définit alors la largeur de la fenêtre par $h(x)$ et la fenêtre de lissage par $(x - h(x), x + h(x))$. Pour l'estimation de $s(x)$, on n'utilise généralement que les observations à l'intérieur de cette fenêtre. De plus, on assigne à chacune de ces observations un poids qui, de manière générale, décroît à mesure que l'on s'éloigne du point d'estimation selon une fonction de pondération nommée **fonction noyau.** Mentionnons qu'en n'utilisant, pour l'estimation de $s(x)$, que les observations à l'intérieur de la fenêtre de lissage, on se trouve à faire l'hypothèse que la fonction noyau utilisée est une fonction non nulle uniquement sur l'intervalle $]-1, 1[$ (c'est-à-dire à support $[-1, 1]$).

La RLP peut être vue comme une généralisation du lissage par noyau (lorsque $p = 0$), une technique de lissage qui a fait l'objet de nombreuses publications tant en statistique qu'en hydrologie (voir la revue hydrologique de Lall (1995)). La RLP (de degré p supérieur à 0) produit cependant des estimations habituellement moins biaisées notamment aux bornes du domaine de la variable prédictive (*effets de bord*) ou lorsque la courbure de la fonction à estimer est prononcée (*effets de courbure*) (pour plus de détails, voir Hastie et Loader (1993)). On peut aussi montrer que lorsque $p = 1$ et qu'une fonction noyau rectangulaire est employée (voir, par exemple, Wand et Jones (1990)), on retrouve le modèle de régression régionale par région d'influence de Tasker et al. (1996) à une seule variable prédictive. L'utilisation de la RLP possède cependant comme avantage, entre autres, une plus grande souplesse pour représenter les diverses formes possibles de fonctions de régression à estimer.

L'approche de modélisation additive par polynômes locaux semble prometteuse puisqu'elle ne nécessite aucune hypothèse a priori sur la forme finale de la relation régionale. La forme du modèle est déterminée, à partir des observations, au moment de l'estimation. En général, cette technique de modélisation est appropriée lorsqu'on possède peu ou aucune information sur la forme exacte de la relation que l'on veut modéliser. Or, pour la modélisation des caractéristiques de crue, il n'existe aucune justification physique a priori pour l'utilisation du modèle log-linéaire de régression régionale des quantiles.

L'approche d'analyse régionale traditionnelle nécessite que l'on détermine, dans un premier temps, les sites jaugés à partir desquels s'effectue le transfert d'information régionale afin d'appliquer, par la suite, un modèle de transfert régional. L'approche de modélisation additive proposée est innovatrice en ce sens où l'étape du choix des sites se trouve intégrée à la procédure d'estimation du modèle. En effet, en utilisant une approche de régression locale pour le lissage dans l'espace de chacune des variables explicatives $X_1, X_2, ..., X_d$, les paramètres de largeur des fenêtres $h(x_1), h(x_2), ..., h(x_d)$ permettent de déterminer automatiquement des voisinages dans l'espace des différentes variables explicatives. Ainsi, avec l'approche de modélisation proposée, le modèle de transfert peut être estimé à l'aide des données de tous les sites disponibles puisque seule l'information "importante" provenant des différents voisinages aura une influence sur l'estimation de la forme du modèle.

Puisque l'étape du choix des sites se trouve intégrée à la procédure d'estimation du modèle de transfert, d'un point de vue opérationnel, la procédure de modélisation régionale est simplifiée. Cette approche permet enfin, à l'aide de procédures statistiques basées sur des critères objectifs, une estimation des paramètres du modèle additif (par exemple, la taille des différentes fenêtres

de lissage) directement reliée à l'objectif principal des procédures de régionalisation (i.e. l'estimation la plus précise possible des quantiles de crue). Ainsi, nous émettons et voulons donc vérifier, dans le cadre de cette thèse, l'hypothèse que cette procédure de régionalisation permet des estimations plus précises de Q_T en des sites non jaugés.

1.2 Objectifs de recherche

L'objectif principal de cette recherche est d'évaluer l'utilisation d'une approche de modélisation non paramétrique, la modélisation additive par polynômes locaux, comme méthode alternative d'estimation régionale des débits de crue Q_T en des sites non jaugés. Afin d'atteindre cet objectif, deux étapes principales sont identifiées :

1. présenter une description détaillée des diverses composantes de base d'une approche de modélisation non paramétrique et plus particulièrement de l'approche de modélisation additive par régression locale polynomiale. Plus spécifiquement, il s'agit de discuter des éléments à prendre en considération lors du choix :
   - des variables explicatives à inclure dans le modèle,
   - du niveau de lissage associé à chacune de ces variables explicatives,
   - du degré des différents polynômes locaux, et
   - de la fonction noyau à utiliser ; et
2. évaluer les qualités prédictives de l'approche de modélisation proposée :
   - déterminer des indices de performance, et
   - comparer l'approche proposée à d'autres approches utilisées présentement.

Les objectifs plus particuliers de cette recherche sont les suivants :

- présenter une revue des travaux reliés à la modélisation locale et régionale des quantiles de crue. Cette revue devrait permettre de présenter la problématique particulière de l'estimation régionale, de faire le point sur l'état des connaissances en régionalisation, de mettre en évidence les lacunes des méthodes existantes et ainsi de mettre en contexte la méthode proposée ;
- présenter une synthèse des principales techniques de régression non paramétrique en insistant davantage sur les approches de la régression locale et de la modélisation additive ;
- proposer une méthodologie de modélisation régionale non paramétrique des quantiles de crue, c'est-à-dire une procédure à employer pour (1) choisir les variables à inclure dans le modèle et (2) effectuer la calibration du modèle de transfert ;

- comparer l'utilisation de l'approche proposée, à l'aide de données réelles, à l'approche de la régression régionale par région d'influence de même qu'à l'approche classique de régression régionale par un modèle log-linéaire ; et
- identifier les situations pour lesquelles la méthode proposée est préférable aux autres modèles de transfert d'information régionale.

1.3 Plan de la thèse

Afin d'atteindre les différents objectifs de cette recherche, nous présentons, au **chapitre 2**, certains concepts reliés à l'approche de modélisation non paramétrique. Les principales méthodes de lissages (appelées lisseurs) sont présentées et l'approche du lissage simple (une seule variable explicative) par régression locale polynomiale est détaillée. Par la suite nous nous intéressons, au **chapitre 3**, à la régression non paramétrique multiple (ou multidimensionnelle). Nous décrivons l'approche de modélisation additive par polynômes locaux et proposons une méthodologie permettant la calibration de ce modèle. Nous présentons par la suite, au **chapitre 4**, les concepts importants de l'analyse fréquentielle locale et de l'analyse régionale de la distribution des crues. Nous y traitons, entre autres, des notions d'homogénéité régionale et de similarité hydrologique, des notions souvent confondues dans la littérature hydrologique. De plus, nous présentons une revue des principales méthodes d'estimation locale et régionale existantes. Au **chapitre 5**, nous appliquons l'approche de modélisation additive par polynômes locaux à des données de différentes régions des États-Unis. Nous discutons enfin, au **chapitre 6**, des résultats obtenus par l'approche de modélisation additive par polynômes locaux et proposons certaines perspectives de recherches futures.

2. LE LISSAGE PAR RÉGRESSION LOCALE

Dans ce chapitre, nous présentons quelques notions de base reliées à l'estimation non paramétrique des courbes de régression. Nous commençons par distinguer, dans la section 2.1, l'approche de modélisation non paramétrique de l'approche paramétrique. Avec l'approche non paramétrique, la procédure utilisée pour produire l'estimation de la forme de la fonction de régression f s'appelle un **lisseur** alors que l'estimation se nomme le **lissage**. Le lisseur est l'outil utilisé pour refléter, mathématiquement, la tendance de la variation des valeurs de la réponse Y en fonction d'une ou de plusieurs variables prédictives $X_1, X_2, ...X_d$. Dans ce chapitre, on ne s'intéresse qu'au lissage des courbes de régression, c'est-à-dire au lissage unidimensionnel avec une seule variable explicative ($d = 1$) puisque rappelons-le, l'approche de modélisation additive proposée dans cette thèse consiste à effectuer un lissage unidimensionnel sur chacune des variables explicatives du modèle additif.

Une grande variété de lisseurs est disponible dans la littérature statistique. Une propriété commune aux différentes méthodes de lissage est le caractère local de l'estimation, c'est à dire que pour estimer la fonction $f(x)$ en un point x_0, on n'utilise que les observations dans le voisinage de x_0. De plus, une autre caractéristique commune à ces méthodes est qu'il est possible d'ajuster leur niveau de lissage. Le lissage nécessite ainsi la prise de deux décisions importantes : comment s'effectuera l'estimation dans le voisinage du point x_0 ? Et, comment s'effectuera l'ajustement du niveau de lissage ? Puisque ce qui distingue principalement les différents lisseurs est leur façon d'effectuer l'estimation dans un voisinage donné, la question du type d'estimation consiste à choisir la méthode particulière de lissage à employer. Afin de faciliter le choix d'une de ces méthodes, nous présentons, dans la section 2.2, les principaux lisseurs disponibles dans la littérature statistique. Quant à la question de l'ajustement du niveau de lissage, elle sera reportée à la section 2.3 où nous décrirons de manière détaillée l'approche de modélisation retenue, l'approche du lissage par régression locale.

2.1 L'approche de modélisation non paramétrique

Considérons le modèle boîte noire suivant :

$$X \longrightarrow \blacksquare \longrightarrow Y \tag{2.1}$$

où X représente (en entrée) un vecteur (indépendant) de variables prédictives, Y représente

la réponse (ou variable dépendante) du modèle et la boîte noire représente une association quelconque (une relation de cause à effet ou simplement une relation fonctionnelle) entre les variables prédictives X et la réponse Y. Un des objectifs majeurs de la statistique est de modéliser, à l'aide d'un échantillon du couple (X,Y), cette relation entre X et Y.

Un résultat classique en statistique est que la fonction f qui minimise $E\{Y - f(X)\}^2$ (avec $E(Y^2) < +\infty$) est l'espérance conditionnelle de Y étant donné X, soit

$$f(X) = E(Y|X). \tag{2.2}$$

Cette fonction qui permet la meilleure prédiction, au sens de l'erreur quadratique moyenne, de Y étant donné X se nomme la fonction de régression de Y sur X.

Dans le cas où l'on n'a qu'une seule variable prédictive, un des modèles les plus utilisés pour résoudre ce problème est le modèle de régression linéaire suivant :

$$E(Y|X) = \alpha + \beta X \tag{2.3}$$

où α et β sont les paramètres du modèle. Le modèle linéaire (2.3) est un exemple de modèle de **régression paramétrique** puisque l'on assume que la forme de la fonction de régression est connue à l'exception près des valeurs des **paramètres** α et β que l'on doit estimer à partir de l'échantillon du couple (X,Y).

Il est facile d'envisager des situations où le modèle linéaire est inapproprié comme par exemple lorsque la dépendance entre $E(Y)$ et X est non linéaire. Il est alors possible d'ajouter au modèle un terme (par exemple, X^2) mais il est souvent difficile de déterminer a priori la forme fonctionnelle la plus appropriée en ne se fiant qu'à un examen visuel des observations. Une approche alternative, l'approche de la **régression non paramétrique**, peut alors être employée pour déterminer cette forme fonctionnelle. Cette approche est dite **non paramétrique** puisqu'on a enlevé la restriction selon laquelle la fonction de régression devait appartenir à une famille de fonctions paramétriques préalablement définie. Ainsi, le modèle de régression non paramétrique est défini par :

$$E(Y|X) = f(X) \tag{2.4}$$

où aucune hypothèse n'est émise sur la forme de la fonction $f(X)$. Le modèle de l'équation (2.4)

peut aussi être généralisé au cas où l'on considère plus d'une variable prédictive et devient :

$$E(Y|X_1, X_2, ..., X_d) = f(X_1, X_2, ..., X_d) \tag{2.5}$$

Dans le cas où $d = 1$, on appelle **courbe de régression**, la fonction de régression f alors que pour $d \geq 2$, on utilise plutôt les termes **surface de régression**. Les techniques de régression non paramétrique présentées dans ce chapitre ont pour objectif l'estimation (appelée dans ce contexte **lissage**) des courbes et surfaces de régression f.

Une motivation philosophique à l'utilisation d'une approche non paramétrique pour la régression concerne la façon dont les données sont utilisées dans un contexte de modélisation. En hydrologie, Adamowski et Feluch (1991) résument ainsi cette philosophie :

> "An important difference between the parametric and the nonparametric methods is how the data are used in model development. In the parametric method, data are used as a guide in model selection and in parameter estimation of the selected model. Thus information that can be extracted from the data is restricted to what can be obtained under the assumed parametric model. Therefore, this method is, to a certain extend, a subjective one, especially concerning the model selection. On the other hand, when little is known about the regression function, information about it is contained in the data rather than the person performing the study. As a result of this, nonparametric method is preferable, where the data speak for themselves concerning the actual form of the regression curve."

Lall (1995), dans une revue des applications récentes de l'approche non paramétrique pour l'estimation de fonctions hydrologiques est quant à lui beaucoup plus incisif envers les hydrologues statisticiens qui utilisent aveuglément l'approche paramétrique :

> "The zealotry often associated with the advocacy of particular models (e.g., floods are LP3) and parameter estimation procedures, has served to mask the basic question faced by statistical hydrologists, which is, the estimation of some function that summarizes structural relationships implicit in the data."

Ainsi, dans un contexte de régionalisation des quantiles de crue où il n'existe aucune justification physique à l'utilisation d'un modèle de régression linéaire multiple, l'approche non paramétrique est une avenue qui mérite d'être explorée plus profondément.

2.2 Les principales méthodes de lissage

2.2.1 Introduction au lissage

Afin de se faire une idée de ce que signifie "laisser les données nous montrer la forme de la fonction de régression", examinons quelques données. La figure 2.1 montre un exemple de graphiques de Y en fonction de X où en 2.1a, on a ajusté une droite de régression alors qu'en 2.1b, on a effectué un lissage par régression linéaire locale. Les données utilisées proviennent d'une étude sur le diabète réalisée par Sockett et al. (1987) et ont déjà été étudiées de manière exhaustive par plusieurs auteurs dans un contexte de régression non paramétrique (voir par exemple Hastie et Tibshirani (1990), Loader (1999)). L'étude initiale avait pour objectif l'étude de facteurs tels que l'âge et le déficit de base (une mesure d'acidité) pouvant influencer le niveau d'un sérum, le C-peptide. À la figure 2.1, on présente la relation entre le logarithme de la concentration de C-peptide (pmol/ml) et un des facteurs explicatifs, l'âge.

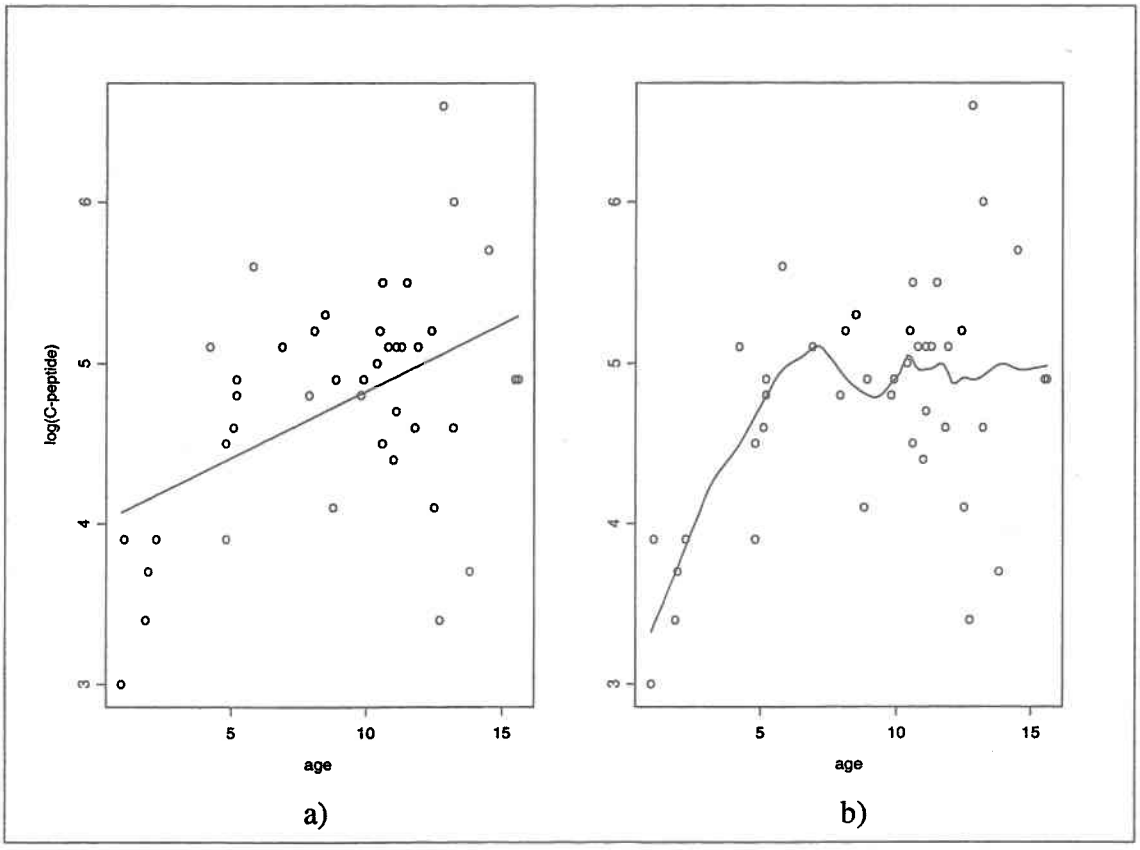

FIG. 2.1: Comparaison entre a) une régression linéaire et b) un lissage

Il semble clair, en examinant la figure 2.1a , que la droite de régression ne s'ajuste pas aux observations. Cependant, le lissage de la figure 2.1b permet d'améliorer l'apparence visuelle de la relation et ainsi d'observer une certaine tendance. Pour effectuer cette estimation, on a effectué un lissage par **droite mobile localement pondérée** selon les étapes suivantes :

1. au point X_0, on détermine les 11 points correspondant aux valeurs de X les plus voisines de X_0
2. on assigne à chaque observation un poids qui dépend de la distance $(X - X_0)$ - plus la distance est faible, plus le poids est grand
3. on estime à l'aide de ces 11 observations et par la méthode des moindres carrés pondérés, une droite de régression pour obtenir $\hat{Y}_0$, l'estimation de la droite de régression de Y sur X au point X_0.
4. on répète 1,2 et 3 pour tous les points en X

Le lissage poursuit généralement deux objectifs principaux, le premier est d'ordre descriptif, le second prédictif. Ainsi, le lissage peut être utilisé pour améliorer l'adéquation d'une relation entre Y et X et permettre ainsi l'observation de tendances sur le graphique du lissage, appelé aussi tout simplement le **lissage**. Le second objectif consiste en l'utilisation de $\hat{f}$ pour la prédiction $\hat{Y}_i$ de $E(Y|X)$ étant donné une valeur particulière X_i de la variable prédictive X. Pour la régionalisation des quantiles de crue, le lissage des relations entre quantiles de crues et variables physiographiques/climatologiques devrait dans un premier temps permettre, en examinant les différents lissages, de détecter les régions à l'intérieur desquelles l'hypothèse de linéarité inhérente au modèle de régression log-linéaire semble inadéquate. On devrait alors s'attendre à ce qu'en ces régions, une approche de régression non paramétrique soit préférable et permette de produire, dans un deuxième temps, de meilleures prédictions des quantiles de crue en des sites non jaugés.

2.2.2 Définition du lissage

Soit $\mathbf{x} = (x_1, ..., x_n)$ un vecteur de n observations de la variable prédictive X classées en ordre croissant et $\mathbf{y} = (y_1, ..., y_n)$ le vecteur des réponses correspondantes, alors le lissage des observations $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ consiste à estimer $E(Y|X = x_0)$, la moyenne de Y au point $x = x_0$. On note $S(y|x_0)$, ou simplement $s(x_0)$, le lissage de $\mathbf{x}$ et $\mathbf{y}$ évalué au point x_0. Habituellement, la procédure, appelée **le lisseur**, qui définit $s(x_0)$ est définie pour tout point x_0. Cependant, il arrive quelquefois que la fonction s ne soit définie que pour les valeurs des observations $x_1, x_2, ..., x_n$. Dans ce cas, il est alors nécessaire d'effectuer une procédure d'interpolation afin d'obtenir des estimations en d'autres valeurs de x.

2.2.3 Le régressogramme

Le régressogramme est à l'estimation d'une fonction de régression ce que l'histogramme est à l'estimation d'une fonction de densité. Il consiste à définir des points de coupure sur l'axe des x et à effectuer la moyenne des y à l'intérieur de la partition ainsi obtenue. Plus formellement, on choisit les points de coupure $c_0 < ... < c_K$ où $c_0 = -\infty$ et $c_K = +\infty$, puis on définit les indices

$$R_k = \{i; c_k \leq x_i < c_{k+1}\}; \qquad k = 0, ..., K-1, \tag{2.6}$$

qui définissent la partition. On a alors :

$$s(x_0) = moy_{i \in R_k}(y_i) \text{ si } x_0 \in [c_k, c_{k+1}]. \tag{2.7}$$

2.2.4 Le lissage par moyennes, médianes et droites mobiles

Si l'on veut produire une estimation au point x_i où on a plusieurs observations, on peut alors prendre la moyenne des valeurs de y en ces points pour estimer $s(x_i)$. Si on a une seule observation au point x_i, on peut alors prendre plutôt la moyenne des valeurs de y aux points voisins de x_i. Le choix des points voisins de x_i peut s'effectuer en prenant les r points à la gauche ainsi que les r points à la droite les plus près de x_i. Cette méthode s'appelle la **méthode des plus proches voisins symétriques** et on note les indices de ces points par $N^S(x_i)$. Dans le cas où il est impossible de prendre r points à droite ou à gauche, on en prend alors autant que possible (avec cette méthode particulière). Une définition formelle des indices du voisinage symétrique est

$$N^S(x_i) = \{max(i-r, 1), ..., i-1, i, i+1, ..., min(i+r, n)\}. \tag{2.8}$$

On peut alors définir la **moyenne mobile** par

$$s(x_i) = moy_{j \in N^S(x_i)}(y_j). \tag{2.9}$$

Et de manière équivalente, la **médiane mobile** par

$$s(x_i) = med_{j \in N^S(x_i)}(y_j). \tag{2.10}$$

Une approche alternative à la méthode des plus proches voisins symétriques consiste à définir le voisinage à l'aide des k points les plus près de x_i, peu importe le fait qu'ils se trouvent à droite ou à gauche. Cette méthode s'appelle la **méthode des k plus proches voisins.**

Pour l'estimation en un point x_0 où on n'a aucune observation, il est possible d'utiliser la méthode des k plus proches voisins. On détermine alors les valeurs x_i et x_{i+1} les plus voisines de x_0 telles que $x_i < x_0 < x_{i+1}$ et on effectue une interpolation linéaire entre $s(x_i)$ et $s(x_{i+1})$, c'est-à-dire,

$$s(x_0) = \frac{x_{i+1} - x_0}{x_{i+1} - x_i} s(x_i) + \frac{x_0 - x_i}{x_{i+1} - x_i} s(x_{i+1}) \quad \text{pour } x_i < x_0 < x_{i+1}. \tag{2.11}$$

Une généralisation simple de la méthode de la moyenne mobile est la méthode de la droite mobile où l'on ajuste une droite de régression, à l'aide de la méthode des moindres carrés, dans le voisinage du point d'estimation. On définit ainsi la **droite mobile** par

$$s(x_0) = \hat{\alpha}(x_0) + \hat{\beta}(x_0) x_0 \tag{2.12}$$

où $\hat{\alpha}(x_0)$ et $\hat{\beta}(x_0)$ sont les paramètres de la droite de régression estimés à l'aide des observations du voisinage $N^S(x_0)$.

Pour chacune des méthodes présentées dans cette section, le nombre d'observations dans le voisinage du point d'estimation (la grandeur du voisinage) contrôle l'apparence de la courbe obtenue par l'une ou l'autre de ces méthodes. Plus le voisinage sera grand (petit), plus la courbe obtenue sera lisse (irrégulière). La grandeur du voisinage constitue ainsi le paramètre, appelé paramètre de lissage, de la modélisation que l'on devra éventuellement ajuster en fonction du niveau de lissage désiré. Il est souvent plus pratique de penser en terme de grandeur relative du voisinage (GRV) (*span*), c'est-à-dire en terme de proportion des observations utilisées pour l'estimation. Ainsi, pour la méthode des k plus proches voisins GRV=k/n.

2.2.5 Le lissage par noyau

L'estimation produite, au point x_0, par la méthode de lissage par noyau peut être représentée par une moyenne pondérée des observations dans le voisinage de x_0. Pour l'estimation en un point x_0, le poids accordé à l'observation en x_j est défini par :

$$S_{0j} = \frac{c_0}{\lambda} d\left(\frac{x_0 - x_j}{\lambda}\right) \tag{2.13}$$

où $d(t)$ est une fonction, appelée **fonction noyau** (*kernel*), symétrique et qui généralement décroît en fonction de $|t|$. Le paramètre λ contrôle l'étendue du voisinage, aussi appelée la largeur de la fenêtre de lissage, et c_0 est une constante permettant de rendre unitaire la somme des pondérations.

Puisque l'estimation au point x_0 représente la moyenne de n valeurs de y pondérée par la fonction de poids S, on a :

$$s(x_0) = \sum_{j=1}^{n} \left(\frac{S_{0j}}{\sum_{j=1}^{n} S_{0j}} \right) y_j \tag{2.14}$$

puis en remplaçant S_{0j} par (2.13) on obtient :

$$s(x_0) = \frac{\sum_{j=1}^{n} d\left(\frac{x_0 - x_j}{\lambda}\right) y_j}{\sum_{j=1}^{n} d\left(\frac{x_0 - x_j}{\lambda}\right)} \tag{2.15}$$

Pour le lissage par noyau, deux paramètres sont susceptibles d'affecter le niveau de lissage obtenu : le paramètre λ qui contrôle l'étendue du voisinage et la fonction noyau $d(t)$. On présente, au tableau 2.1, les principales fonctions noyau utilisées pour le lissage des fonctions de régression. Les études réalisées jusqu'à présent suggèrent que le choix du type de noyau est relativement peu important en comparaison du choix de la largeur de la fenêtre λ (Hastie et Tibshirani, 1990). On peut aussi remarquer que le paramètre λ possède ici la même fonction que les paramètres k ou r des méthodes des plus proches voisins et sert à déterminer le voisinage. Alors qu'avec la méthode du noyau, on détermine le voisinage en fonction d'une métrique (distance), on le fait plutôt selon une distance de rang pour les plus proches voisins.

TAB. 2.1: Les principales fonctions noyau (tiré de Loader (1999))

Rectangulaire	$d(t) = 1$	$\|t\| < 1$
Triangulaire	$d(t) = 1 - \|t\|$	$\|t\| < 1$
Epanechnikov	$d(t) = 1 - t^2$	$\|t\| < 1$
Doublement quadratique	$d(t) = (1 - t^2)^2$	$\|t\| < 1$
Triplement cubique	$d(t) = (1 - \|t\|^3)^3$	$\|t\| < 1$
Triplement pondéré	$d(t) = (1 - t^2)^3$	$\|t\| < 1$
Normal	$d(t) = exp(-(2.5t)^2/2)$	$\|t\| < 1$

2.2.6 Le lissage par splines

Une **spline** est une longue lamelle de bois utilisée dans le domaine de la construction. Anciennement, les architectes navals utilisaient les splines pour la conception et le dessin des coques de navires. La forme de la coque était obtenue en attachant des pesées de plomb en certains points le long de la spline. En augmentant le nombre de pesées et en faisant varier leur position, il était possible de faire passer la spline en des points déterminés à l'avance. (Wegman et Wright, 1983)

Par analogie, une **spline mathématique**, notée $s_\Delta(x)$ est une fonction polynomiale par morceaux où $\Delta = \{\zeta_1, \zeta_2, ..., \zeta_K\}$ représente la position des différents points, appelés noeuds, qui séparent chacun de ces morceaux. En fonction des diverses contraintes que l'on impose aux splines, elles peuvent, en analyse numérique, permettre l'interpolation entre deux noeuds quelconques ou en statistique, permettre le lissage de fonctions de régression. Deux approches de lissage à l'aide des splines ont plus particulièrement retenu l'attention des statisticiens soit l'approche des splines cubiques de lissage et l'approche des splines de régression. Après avoir montré comment s'effectue le calcul des splines cubiques d'interpolation, on présente ces deux approches de lissage.

2.2.6.1 Les splines cubiques d'interpolation

Le problème de l'interpolation consiste à ajuster une courbe passant par les points $\{(x_i, y_i), i = 1, 2, ..., n\}$. Supposons que les noeuds soient positionnés aux valeurs observées en x $\{x_1, x_2, ..., x_n\}$, c'est-à-dire que $\Delta = \{\zeta_1 = x_1, \zeta_2 = x_2, ..., \zeta_n = x_n\}$ et qu'en chacun des intervalles $[\zeta_i, \zeta_{i+1}], i = 1, 2, ..., n-1$, on ait un polynôme cubique de la forme suivante :

$$y_i = a_i(x - x_i)^3 + b_i(x - x_i)^2 + c_i(x - x_i) + d_i \tag{2.16}$$

On appelle alors **spline cubique d'interpolation**, notée $s_\Delta(x)$, la fonction (dont les deux premières dérivées notées respectivement $s'_\Delta(x)$ et $s''_\Delta(x)$ sont continues) formée de polynômes cubiques y_i en chacun des intervalles $[\zeta_i, \zeta_{i+1}]$, $i = 1, 2, ..., n-1$ qui interpole $\{(x_i, y_i), i = 1, 2, ..., n\}$.

Le calcul de la fonction spline consiste à estimer les différents paramètres a_i, b_i, c_i et d_i pour $i = 1, 2, ..., n$. Posons $h_i = x_{i+1} - x_i$ et $M_i = s''_\Delta(x_i)$ pour $i = 1, 2, ..., n$. En prenant les premières dérivées de y_i et en évaluant aux noeuds, on peut montrer que

$$\begin{aligned} a_i &= (M_{i+1} - M_i)/6h_i \\ b_i &= M_i/2 \\ c_i &= \frac{y_{i+1} - y_i}{h_i} - \frac{2(h_i M_i + h_i M_{i+1})}{6} \\ d_i &= y_i \end{aligned} \tag{2.17}$$

Le problème consiste alors à trouver les valeurs des M_i. En utilisant le fait que la première dérivée de la spline est continue et les résultats de (2.17), on peut obtenir les équations suivantes qui relient entre eux, pour $i = 2, 3, ..., n-1$, les différents M_i :

$$h_{i-1}M_{i-1} + 2(h_{i-1} + h_i)M_i + h_i M_{i+1} = 6\left(\frac{y_{i+1} - y_i}{h_i} - \frac{y_i - y_{i-1}}{h_{i-1}}\right) \tag{2.18}$$

Enfin, en posant $M_1 = M_n = 0$, on obtient un système d'équations linéaires tridiagonal qui se résout facilement en $O(n)$ opérations par élimination de Gauss. On nomme **spline cubique naturelle**, cette spline ayant comme condition que $M_1 = M_n = 0$.

2.2.6.2 Les splines cubiques de lissage

Les splines d'interpolation sont utilisées principalement en analyse numérique et ont généralement peu d'intérêt en statistique. Cependant, dans un contexte d'estimation statistique par lissage, on a développé des méthodes de lissage basées sur les splines. On a cherché à obtenir des splines qui pouvaient passer près des observations sans toutefois avoir comme contrainte leur interpolation. La méthode des splines cubiques de lissage satisfait cette condition.

Contrairement aux méthodes de lissage présentées précédemment où le lissage au point x_0 consistait à appliquer une procédure définie de manière explicite, le lissage par splines cubiques est plutôt le résultat d'un problème d'optimisation. Ainsi, parmi toutes les fonctions $s(x)$ ayant leur deux premières dérivées continues, le lissage par splines cubiques, aussi appelée l'**approche de pénalité de la rugosité** (Green et Silverman, 1994), consiste à calculer celle qui minimise la somme pénalisée suivante du carré des résidus :

$$\mathrm{S} = \sum_{i=1}^{n} \{y_i - s(x_i)\}^2 + \lambda \int_a^b \{s''(t)\}^2 dt \tag{2.19}$$

où λ est une constante et $a \leq x_1 \leq ... \leq x_n \leq b$.

Le terme de gauche de l'équation (2.19) constitue une mesure de la proximité des observations avec la courbe s alors que le terme de droite pénalise la rugosité de la fonction s. Avec cette méthode de lissage, le paramètre λ permet d'ajuster le niveau de lissage désiré. De grandes valeurs de λ produisent des courbes plus lisses alors que de plus petites valeurs produisent des courbes plus ondulées, plus rugueuses. À la limite, lorsque $\lambda \to \infty$, le terme de pénalité domine ce qui oblige $s''(x) = 0$ partout et ainsi, la solution de (2.19) est la droite de régression qui minimise les moindres carrés. D'un autre côté, lorsque $\lambda \to 0$, le terme de pénalité devient négligeable et ainsi, la solution tend vers la fonction doublement dérivable qui interpole les n observations, soit la spline cubique d'interpolation.

Une caractéristique importante ayant favorisé l'utilisation de cette approche est l'unicité de la spline cubique. En effet, parmi toutes les fonctions $s(x)$ doublement dérivables ayant leur deux premières dérivées continues, celle qui minimise (2.19) est unique et de forme explicite : il s'agit d'une spline cubique naturelle dont les noeuds sont placés aux points $x_1, x_2, ..., x_n$. Ce résultat,

combiné à l'utilisation de l'algorithme de Reinsch (1967) permet l'estimation de s en $O(n)$ opérations (voir par exemple la section 2.3 de Green et Silverman (1994) ou 2.10 de Hastie et Tibshirani (1990)).

2.2.6.3 Les splines de régression

L'approche des splines de régression consiste à faire l'hypothèse que la forme de la courbe de régression est une spline, c'est-à-dire une fonction polynomiale par morceaux $s_\Delta(x)$ où $\Delta = \{\zeta_1, \zeta_2, ..., \zeta_K\}$ représente la position des différents noeuds. Les morceaux, des polynômes de degré m, sont généralement attachés de manière à ce qu'il y ait continuité des $m-1$ premières dérivées. Le régressogramme constitue un exemple d'une spline de régression ($m = 0$). Contrairement aux approches par splines présentées précédemment, le résultat obtenu ici ne minimise aucunement la courbure de la fonction de régression, n'interpole pas les observations et la position des noeuds ne coïncide pas nécessairement avec les observations. D'ailleurs, en pratique, on aura généralement beaucoup moins de noeuds que d'observations.

Au sens strict, l'approche des splines de régression est une approche paramétrique puisque la forme fonctionnelle de la fonction de régression est connue à l'exception près des valeurs de ses paramètres. Cependant, contrairement à la régression polynomiale où l'ajustement des paramètres est effectuée globalement, on conserve avec la méthode des splines de régression une caractéristique de l'approche non paramétrique soit le caractère local de l'estimation. En effet, en faisant varier certains paramètres de la modélisation (versus les paramètres du modèle) tels le degré m de la spline, le nombre de noeuds K et la position des différents noeuds ζ_i, on réussit à effectuer un certain lissage des observations même si aucun paramètre particulier ne contrôle directement le niveau de lissage.

Soit $\Delta = \{\zeta_1, \zeta_2, ..., \zeta_K\}$, alors une représentation possible de $s_\Delta^{(m)}(x_0)$, la spline de degré m ayant K noeuds aux positions données par Δ et évaluée au point x_o est :

$$s_\Delta^{(m)}(x_0) = \sum_{j=0}^{m} \beta_{0j} x_0^j + \sum_{k=1}^{K} \sum_{j=0}^{m} \beta_{kj} (x_0 - \zeta_k)_+^j \tag{2.20}$$

où

$$(u)_+ = \begin{cases} u & \text{si } u \geq 0 \\ 0 & \text{autrement} \end{cases} \tag{2.21}$$

Cette représentation est intéressante puisqu'elle permet d'exprimer le problème de l'estimation d'une spline en un problème de régression multiple ordinaire pouvant être résolu par la méthode

des moindres carrés. De plus, les tests paramétriques traditionnels permettent de tester l'hypothèse $\beta_{kj} = 0$ afin éventuellement de diminuer le degré de certains polynômes réduisant ainsi le nombre total de paramètres du modèle.

La méthode des splines de régression est intéressante en raison de la facilité de son estimation lorsque les noeuds sont donnés. Cependant, une difficulté majeure de l'utilisation de cette méthode concerne le choix du nombre de noeuds et de leurs positions (Hastie et Tibshirani, 1990), des caractéristiques qui doivent d'abord être fixées pour estimer les paramètres de la régression. De plus, cette méthode ne permet pas de faire varier globalement le niveau de lissage à l'aide d'un paramètre unique de lissage.

2.2.7 Le lissage par polynômes mobiles localement pondérés

Dans la littérature statistique, on retrouve la méthode du lissage par polynômes mobiles localement pondérés sous diverses appellations : lissage par ajustement local (Cleveland et Loader, 1994), par polynômes locaux de type noyau (Wand et Jones, 1990), par régression locale polynomiale (Opsomer, 1995), par régression localement pondérée (Cleveland, 1979) ou tout simplement par régression locale (Loader, 1999). Par souci de clarté, on privilégiera, dans ce document, l'emploi de la terminologie **régression locale**.

Avec la méthode du lissage par régression locale, on ne fait globalement aucune hypothèse quant à la forme de la fonction s. Toutefois, dans le voisinage d'un point d'estimation x, on suppose que s puisse être approximée par une fonction paramétrique, soit un polynôme de degré p. En un point x, on définit la largeur de la fenêtre par $h(x)$ et la fenêtre de lissage par $(x-h(x), x+h(x))$. Pour l'estimation de $s(x)$, on n'utilise alors que les observations à l'intérieur de cette fenêtre. De plus, on assigne à chacune de ces observations un poids qui, généralement, décroît à mesure que l'on s'éloigne du point d'estimation. De manière formelle, on assigne les poids selon la formule suivante :

$$w_i(x) = W\left(\frac{x_i - x}{h(x)}\right) \tag{2.22}$$

où $W(u)$ est une fonction noyau symétrique de support $[-1, 1]$ qui satisfait les conditions suivantes de Loader (1999) : $W(0) = 1$, $W(1) = 0$ et W décroît sur l'intervalle $[0, 1]$ (et croît sur $[-1, 0]$).

Une méthode particulière d'estimation par régression locale consiste à estimer un polynôme de degré $p = 1$ (une droite) à l'aide d'une fonction noyau triplement cubique en choisissant $h(x)$ de manière à obtenir exactement k observations (les k plus proches voisins) pour l'estimation au point x. Cette méthode, proposée par Cleveland (1979) est connue en statistique sous le nom de *lowess* ou *loess* en raison du nom des procédures d'estimation développées successivement par Cleveland pour le langage de programmation S (Becker et al., 1988).

Une des particularités intéressantes de l'approche de la régression locale est la facilité de traitement de l'hétéroscédasticité, une caractéristique généralement associée au problème de la régionalisation des quantiles de crue. Dans le cas d'un modèle de régression régionale avec hétéroscédasticité, une approche utilisée en statistique consiste à estimer les paramètres du modèle en utilisant la méthode des moindres carrés pondérés (voir par exemple Stedinger et Tasker (1985)). Avec la régression locale, puisque la fonction $w_i(x)$ est déjà une fonction de pondération, l'approche employée consiste tout simplement à modifier cette fonction $w_i(x)$ en la multipliant par l'inverse de la variance σ_i^2, c'est-à-dire en redéfinissant $w_i'(x) = w_i(x)/\sigma_i^2$. Ce concept de repondération est aussi employé pour permettre d'obtenir des estimations plus robustes de la courbe de régression. Par exemple, Cleveland (1979) propose d'effectuer d'abord une première estimation de la courbe de régression puis de repondérer de manière à diminuer le poids accordé aux observations ayant obtenues les plus grandes erreurs (ou déviations). Cette approche permet d'accorder automatiquement un poids presque nul aux observations pouvant être considérées comme des *outliers* produisant par le fait même des estimations plus robustes.

2.2.8 Les principaux lisseurs de surfaces

La plupart des méthodes présentées précédemment peuvent se généraliser au cas multidimensionnel (plusieurs variables explicatives). Il suffit généralement de déterminer le voisinage en utilisant une distance mesurée dans un espace de dimension p. Par exemple, l'approche du lissage par une droite mobile de la section 2.2.4 peut se généraliser, pour l'estimation par lissage au point $\mathbf{x}^0 = (x_1^0, x_2^0, ..., x_p^0)$, en ajustant plutôt une droite de dimension p au point $\mathbf{x}^0$ à l'aide des k plus proches voisins selon une distance euclidienne séparant deux points particuliers $\mathbf{x}^i$ et $\mathbf{x}^j$ définie par :

$$d_{ij} = (\mathbf{x}^i - \mathbf{x}^j)^T(\mathbf{x}^i - \mathbf{x}^j). \tag{2.23}$$

où T est l'opérateur de transposition de matrice. Remarquons qu'il s'agit ici de l'approche de régression par région d'influence proposée par Tasker et al. (1996) à la seule différence que

Tasker et al. (1996) ont plutôt utilisé une distance de "Mahalanobis"

$$d_{ij} = (\mathbf{x}^i - \mathbf{x}^j)^T \Sigma^{-1} (\mathbf{x}^i - \mathbf{x}^j) \tag{2.24}$$

où Σ est la matrice de variance-covariance des variables explicatives supposée diagonale par Tasker et al. (1996).

Les splines cubiques de lissage ainsi que les splines de régression se généralisent aussi au cas multidimensionnel. Ces approches se nomment alors respectivement : approche des "**splines par plaques minces**" (*thin-plate splines*) et approche des "**splines de produits tensoriels**" (*tensor product splines*). Pour une description détaillée de ces méthodes, le lecteur est référé à Green et Silverman (1994).

2.2.9 Comparaison et choix d'une approche de lissage

Les différents lisseurs unidimensionnels présentés dans cette section poursuivent tous un objectif commun qui est d'obtenir le plus précisément possible une représentation approchée, à l'aide des données d'un échantillon, de la fonction inconnue f qui représente une relation existant (ou dont on suppose l'existence) entre une variable dépendante Y et une variable explicative X. En statistique, on dénombre une quantité considérable d'études théoriques portant sur le comportement asymptotique (consistance, biais, variance, vitesse de convergence) des différents lisseurs (voir par exemple la section 2.10 de Hastie et Tibshirani (1990)). Les résultats de ces études sont constamment utilisés pour la comparaison des différents lisseurs entre eux et pour le choix d'une méthode particulière de lissage. Par exemple, Müller (1987) a démontré l'équivalence asymptotique de la méthode du noyau avec la méthode de la droite mobile localement pondérée alors que Silverman (1984) a réussi à exprimer sous la forme d'un noyau, appelé **noyau équivalent**, une spline cubique de lissage de manière à ce que ces méthodes soient asymptotiquement équivalentes. À la lumière de ces résultats, Hastie et Tibshirani (1990) indiquent d'ailleurs :

> "Some recent theorical results (e.g. Silverman (1984), Müller (1987)) suggest that for appropriately chosen smoothing parameters, there are not likely to be large differences between locally-weighted running-line, cubic smoothing-splines and kernel smoothers."

En pratique cependant, notamment en raison de la taille des échantillons généralement disponibles, il existe des différences notables entre ces approches et il est alors souhaitable de ne pas

trop accorder d'importance à ces résultats asymptotiques puisque comme l'indiquent Cleveland et Loader (1994) :

> "Asymptotic theory for smoothing has in most cases not been helpful as a guide for choosing among different procedures, since the framework that governs most asymptotic work **is not very realistic**. Methods have been put forward with a label of "optimal" that provide optimality **in no sense that is meaningful for practice**."

Il est possible d'interpréter la plupart des lisseurs présentés dans cette section comme étant des polynômes mobiles pondérés par une fonction quelconque des données avoisinantes. Alors que l'approche de la régression locale est générale et permet une modélisation pour différents degrés p des polynômes locaux, avec les autres approches, le degré p est fixé de manière explicite ou implicitement. Ainsi, les approches de la moyenne mobile et du noyau peuvent s'exprimer à l'aide de moyennes mobiles ($p = 0$) localement pondérées. La droite mobile constitue un cas particulier de régression locale lorsque $p = 1$. Avec les différentes approches par splines cubiques, il s'agit implicitement d'ajuster un polynôme de degré $p = 3$.

En régression non paramétrique, l'objectif principal de la modélisation consiste à ajuster un modèle de manière à ce que l'on obtienne un compromis adéquat entre le biais d'estimation B=$E\{\hat{f}\} - f$ et la variance de l'estimateur V=var$\{\hat{f}\}$ (cf. 2.3.5). L'ordre du polynôme local constitue un des facteurs qui influencent ce compromis biais/variance. De manière générale, un dégré p élevé produit des estimations moins biaisées mais plus variables qu'un degré p petit. La question qu'on est alors amené à se poser est : peut-on a priori déterminer le degré optimal p du polynôme local à utiliser ?

Le niveau de courbure de f, mesuré par l'intensité des changements en x dans la pente de f, constitue le premier facteur à prendre en considération pour le choix du degré p du polynôme local. Plus la fonction f est composée de pointes et de vallées, plus le degré p doit être élevé pour en permettre une estimation qui ne soit pas trop biaisée. En pratique, lorsque la courbe f est assez lisse, un polynôme de degré $p = 1$ permet généralement d'obtenir de bons résultats. Les méthodes basées sur la notion de moyenne mobile pondérée $p = 0$, telles la méthode du noyau, ne s'avèrent quant à elles que très rarement être le meilleur choix en pratique (Cleveland et Loader, 1996a). Avec ces méthodes, puisque le voisinage est asymétrique aux bornes du domaine de la variable prédictive, l'estimation est alors fortement biaisée (avec un biais fonction de la pente de f). Il peut aussi y avoir un problème de biais à l'intérieur du domaine si la dispersion des données n'est pas uniforme ou si la fonction de régression possède une courbure prononcée (Hastie et Loader, 1993).

En régression régionale des quantiles de crue, puisque les variables prédictives sont généralement transformées par une fonction logarithmique, on devrait s'attendre à ce que leur courbure ne soit pas trop prononcée. Par contre, il semble difficile a priori de déterminer si, pour chacune des variables prédictives composant le modèle, un degré $p = 1$ suffit ou si l'on doit plutôt opter pour un degré supérieur. Ainsi, nous croyons qu'il est préférable d'opter pour une approche de modélisation par régression locale où le paramètre p n'est pas fixé a priori et de plutôt déterminer les différentes valeurs de p à l'aide de tests diagnostiques.

La régression locale est une vieille méthode de lissage qui a d'abord été développée vers la fin du 19ième siècle par des actuaires pour la graduation de données de mortalité (voir Cleveland et Loader (1996a) pour une revue historique détaillée). La régression locale possède de nombreux points forts dont certains sont discutés en détail par Hastie et Loader (1993) :

1. La méthode s'adapte bien aux problèmes de biais aux bornes et dans des régions de grande courbure ;
2. Elle est facile à comprendre et à interpréter ;
3. Des méthodes produisant des calculs rapides ont été développés pour son estimation ;
4. En raison de sa simplicité, elle peut s'adapter pour prendre en considération diverses hypothèses de distribution (avec hétéroscédasticité, par exemple) ;
5. Elle ne requière pas la présence d'hypothèses strictes sur le niveau de lissage de la courbe ;
6. Elle appartient à la famille des lisseurs linéaires ; et
7. Le fait d'avoir un modèle local (plutôt qu'une seule estimation ponctuelle $\hat{f}(x)$), permet de développer des méthodes pour choisir directement la largeur de la fenêtre de même que l'ordre du polynôme local à l'aide des réponses (prédictions) obtenues par le modèle.

Selon Cleveland et Loader (1996a), aucun de ces points forts ne procure, à lui seul, une raison évidente pour choisir cette approche plutôt qu'une autre ; il s'agit plutôt de la combinaison de ceux-ci qui rend l'approche de régression locale attrayante.

2.3 La modélisation par régression locale

2.3.1 Le modèle de régression locale

La régression locale est utilisée pour modéliser une relation entre une variable prédictive (ou indépendante) X et la variable de réponse Y (ou variable dépendante) qui est reliée à

cette variable prédictive. Supposons qu'on ait un ensemble de n paires d'observations $\{(x_i, y_i), i = 1, 2, ..., n\}$, alors le modèle de régression locale est :

$$y_i = f(x_i) + \epsilon_i \tag{2.25}$$

où $f(x)$ est une fonction inconnue et ϵ_i est un terme d'erreur aléatoire. On suppose aussi que les erreurs ϵ_i sont indépendantes et identiquement distribuées de moyenne 0 ; $E(\epsilon_i) = 0$, et de variance finie ; $E(\epsilon_i^2) = \sigma^2$. De plus, bien qu'on ne fasse globalement aucune hypothèse quant à la forme de la fonction f, on suppose néanmoins que, dans le voisinage d'un point d'estimation x, f puisse être approximée par un polynôme de degré p.

2.3.2 L'estimation de la courbe de régression

Comme on l'a déjà mentionné à la section 2.2, l'estimation $\hat{f}(x)$ de la courbe de régression en un point x consiste à ajuster, à l'intérieur de la fenêtre de lissage $(x - h(x), x + h(x))$, un polynôme de degré p en accordant une pondération $w_i(x)$ à chacune des observations x_i de la fenêtre de lissage. Ainsi, soit $P(u)$ le polynôme de degré p ayant la forme suivante :

$$P(u) = \beta_0 + \beta_1(u - x) + \beta_2\frac{(u - x)^2}{2} + ... + \beta_p\frac{(u - x)^p}{p!} \tag{2.26}$$

lorsque $|u - x| < h(x)$. Alors, le vecteur de paramètres $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, ..., \hat{\beta}_p)^T$ peut être obtenu en appliquant localement la méthode des moindres carrés pondérés, ce qui consiste à obtenir le vecteur de paramètres $\hat{\boldsymbol{\beta}}$ qui minimise

$$\sum_{i=1}^{n} w_i(x)\left\{y_i - \beta_0 - \beta_1(x_i - x) - ... - \beta_p\frac{(x_i - x)^p}{p!}\right\}^2. \tag{2.27}$$

où

$$w_i(x) = W\left(\frac{x_i - x}{h(x)}\right)$$

En faisant l'hypothèse que la matrice $\mathbf{X}_x^T\mathbf{W}_x\mathbf{X}_x$ est inversible (dans le cas contraire, on peut faire en sorte qu'elle le devienne en modifiant h), on peut montrer que la solution de (2.27) donnée par la théorie classique des moindres carrés pondérés est (Wand et Jones, 1990) :

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}_x^T\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{X}_x^T\mathbf{W}_x\mathbf{Y} \tag{2.28}$$

où :

$\mathbf{Y} = (y_1, ..., y_n)^T$ est le vecteur des réponses,

$$\mathbf{X}_x = \begin{bmatrix} 1 & x_1 - x & \dots & \frac{(x_1-x)^p}{p!} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n - x & \dots & \frac{(x_n-x)^p}{p!} \end{bmatrix}$$ est la matrice de conception de dimension $n \times (p+1)$ et

$\mathbf{W}_x = \text{diag}\{w_1(x), ..., w_n(x)\}$ est la matrice diagonale de dimension $n \times n$ des pondérations.

En posant $u = x$ dans l'équation (2.26) et en substituant les paramètres par leurs estimateurs, on obtient :

$$\hat{f}(x) = \hat{\beta}_0 \tag{2.29}$$

ou en notation matricielle

$$\hat{f}(x) = \mathbf{e}_1^T \hat{\boldsymbol{\beta}} = \mathbf{e}_1^T (\mathbf{X}_x^T \mathbf{W}_x \mathbf{X}_x)^{-1} \mathbf{X}_x^T \mathbf{W}_x \mathbf{Y} \tag{2.30}$$

où $\mathbf{e}_1$ est un vecteur de dimension $(p+1) \times 1$ tel que $\mathbf{e}_1 = (1, 0, 0, ..., 0)^T$.

2.3.3 La modélisation des données

De manière générale, lorsqu'on effectue une régression linéaire (simple ou multiple), on porte principalement notre attention sur les paramètres du modèle. On suppose que le modèle estimé est approprié puis on se demande comment les estimateurs des paramètres ajustent bien les vrais paramètres. On calcule, par exemple, la variance de ces estimateurs puis on construit des intervalles de confiance sur ces paramètres. On ne s'intéresse à peu près pas à la fonction (droite) de régression comme telle. Avec l'approche de modélisation par régression locale ou, de manière plus générale, avec une approche non paramétrique, il en est tout autrement. Au lieu de se concentrer sur les paramètres du modèle, on s'intéresse à la courbe de régression f estimée. La question fondamentale que l'on doit alors se poser est : quel est le niveau de précision apporté par $\hat{f}(x)$ pour l'estimation de la vraie courbe $f(x)$? Ainsi, dans la plupart des applications, l'objectif principal de la modélisation locale consiste à obtenir une estimation de f qui permette de produire des estimations qui soient les plus précises possibles.

Avec en main une mesure quelconque de ce que désigne la précision, un objectif de la modélisation consiste alors à maximiser cette précision. Soit $\hat{\epsilon}_i = \hat{f}(x_i) - f(x_i)$ l'erreur ponctuelle d'estimation au point x_i, une des mesures généralement utilisées comme mesure de précision de l'estimation au point x_i est l'espérance de l'erreur quadratique, aussi appelée l'**erreur quadratique moyenne** (EQM) :

$$\begin{aligned} \text{EQM}(x_i) &= E\{\hat{f}(x_i) - f(x_i)\}^2 \\ &= [E\{\hat{f}(x_i)\} - f(x_i)]^2 + \text{var}\{\hat{f}(x_i)\} \end{aligned} \tag{2.31}$$

Remarquons que le terme de gauche représente le carré du biais alors que celui de droite représente la variance de l'estimateur $\hat{f}(x_i)$. Une autre quantité qui ne diffère de l'EQM que par une constante, la variance de l'erreur σ^2, est l'espérance de l'erreur quadratique de prédiction appelée tout simplement l'**erreur quadratique de prédiction** (EQP). L'EQP mesure l'efficacité des modèles à effectuer des prédictions à partir de nouvelles observations. Pour les modèles de régression, l'EQP représente l'espérance du carré de la différence entre une réponse future Y^* et sa prédiction à partir du modèle $\hat{f}(x)$. Puisque par hypothèse $Y_i^* = f(x_i) + \epsilon_i^*$ avec ϵ_i^* indépendant de $\hat{\epsilon}_i$, on obtient alors que :

$$\begin{aligned} \text{EQP}(x_i) &= E\{Y_i^* - \hat{f}(x_i)\}^2 \\ &= E\{Y_i^* - f(x_i) + f(x_i) - \hat{f}(x_i)\}^2 \\ &= E\{\epsilon_i^* - \hat{\epsilon}_i\}^2 \\ &= \sigma^2 + \text{EQM}(x_i) \end{aligned} \tag{2.32}$$

Une caractéristique commune aux différents lisseurs est qu'il est possible d'ajuster le niveau de lissage désiré. Avec l'approche des splines de lissage, on utilise le paramètre λ pour déterminer ce niveau de lissage. Pour les splines de régression, on peut faire varier le nombre K de noeuds, leur position Δ ainsi que le degré m des différentes splines. Avec l'approche de la régression locale, le lissage est influencé par la taille du voisinage (en valeur absolue h ou en nombre d'observations k), par la fonction noyau ou de pondération W utilisée ainsi que par le degré p des polynômes locaux. Ces paramètres contrôlent ce qu'on appelle en régression non paramétrique **le compromis entre le biais et la variance** puisqu'en faisant varier chacun de ces paramètres, on observe généralement soit une diminution de la variance de l'estimateur $\hat{f}(x)$ combinée à une augmentation du biais de celui-ci ou inversement, une diminution du biais combinée à une augmentation de la variance. Un des objectifs de la modélisation non paramétrique consiste alors à trouver un compromis acceptable entre le biais et la variance de l'estimateur.

En pratique, deux approches sont principalement utilisées afin de déterminer ce compromis biais/variance :

1. l'examen visuel des différents lissages produits par régression locale pour différentes valeurs de h, p et W afin de retenir le lissage qui semble le plus approprié, c'est-à-dire ni trop biaisé, ni trop variable. Nous examinerons plus en détail, à la section 2.3.5, comment les différents paramètres de lissage (h, W et p) influencent l'estimation de $\hat{f}(x)$.
2. l'optimisation de critères de sélection tels la validation-croisée (Stone, 1974), la validation croisée généralisée (Craven et Wahba, 1979) ou la statistique C_p de Mallows (Mallows, 1973) adaptée pour la régression locale par Cleveland et Devlin (1988). Nous discuterons aussi plus en détail, à la section 2.3.6, de la sélection automatique des paramètres de lissage.

2.3.4 Définitions relatives aux lisseurs linéaires

2.3.4.1 Le lissage linéaire

Par définition, un lisseur est linéaire si l'estimation par lissage au point x peut s'écrire de la manière suivante :

$$\hat{f}(x) = \sum_{i=1}^{n} l_i(x) y_i = \mathbf{L}_x^T \mathbf{Y} \tag{2.33}$$

où $\mathbf{L}_x^T = (l_1(x), l_2(x), ..., l_n(x))$ est un vecteur nommé **noyau équivalent** ou **diagramme de pondération** qui ne dépend pas des y_i. En examinant l'équation (2.30), on peut voir que le lisseur de régression locale est un lisseur linéaire avec :

$$\mathbf{L}_x^T = \mathbf{e}_1^T (\mathbf{X}_x^T \mathbf{W}_x \mathbf{X}_x)^{-1} \mathbf{X}_x^T \mathbf{W}_x. \tag{2.34}$$

En s'intéressant plus particulièrement à l'estimation de f pour les observations $x_1, x_2, ..., x_n$, on obtient :

$$\begin{bmatrix} \hat{f}(x_1) \\ \hat{f}(x_2) \\ \vdots \\ \hat{f}(x_n) \end{bmatrix} = \begin{bmatrix} \mathbf{L}_{x_1}^T \\ \mathbf{L}_{x_2}^T \\ \vdots \\ \mathbf{L}_{x_n}^T \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \tag{2.35}$$

ou en notation matricielle,

$$\hat{\mathbf{f}} = \mathbf{S}_\lambda \mathbf{Y} \tag{2.36}$$

où $\mathbf{S}_\lambda = \begin{bmatrix} \mathbf{L}_{x_1}^T & \mathbf{L}_{x_2}^T & \dots & \mathbf{L}_{x_n}^T \end{bmatrix}^T$ est une matrice de dimension $n \times n$ que l'on nomme **matrice de lissage** et le sous-indice λ désigne le ou l'ensemble des paramètres de lissage utilisé pour l'estimation.

2.3.4.2 Le biais

Pour un lisseur linéaire, le vecteur de biais est défini par :

$$\mathbf{b}_\lambda = \mathbf{f} - E\{\mathbf{S}_\lambda \mathbf{Y}\} = \mathbf{f} - \mathbf{S}_\lambda \mathbf{f} \tag{2.37}$$

Pour une fonction arbitraire et inconnue f, les lisseurs linéaires sont généralement biaisés. Par contre, ils peuvent être non biaisés pour une classe particulière de fonctions. Par exemple, le lissage par splines cubiques de lissage reproduit sans biais les fonctions linéaires (Buja et al., 1989). Pour le lissage par régression locale à l'aide d'un polynôme de degré p, on peut démontrer, en supposant que la vraie fonction $f(x)$ est $p+2$ fois dérivable, que le biais ponctuel, $\mathrm{b}(x) = E\{\hat{f}(x)\} - f(x)$, est égal à (Loader, 1999) :

$$\mathrm{b}(x) = \frac{f^{(p+1)}(x)}{(p+1)!} \sum_{i=1}^{n} l_i(x)(x_i - x)^{(p+1)} + \frac{f^{(p+2)}(x)}{(p+2)!} \sum_{i=1}^{n} l_i(x)(x_i - x)^{(p+2)} + \dots \tag{2.38}$$

La régression locale par un polynôme de degré p permet ainsi d'estimer sans biais une fonction polynomiale $f(x)$ d'ordre p puisque dans ce cas, $f^{(p+1)}(x) = f^{(p+2)}(x) = \dots = 0$.

2.3.4.3 La variance

La variance de l'estimateur linéaire de l'équation (2.33) est :

$$\mathrm{V}(x) = \sigma^2 \sum_{i=1}^{n} l_i(x)^2 = \sigma^2 \mathbf{L}_x^T \mathbf{L}_x \tag{2.39}$$

et en remplaçant $\mathbf{L}_x^T$ par (2.34), on obtient :

$$\mathrm{V}(x) = \sigma^2 \mathbf{e}_1^T (\mathbf{X}_x^T \mathbf{W}_x \mathbf{X}_x)^{-1} (\mathbf{X}_x^T \mathbf{W}_x^2 \mathbf{X}_x) (\mathbf{X}_x^T \mathbf{W}_x \mathbf{X}_x)^{-1} \mathbf{e}_1. \tag{2.40}$$

Pour un lisseur linéaire, la matrice de variance-covariance des valeurs ajustées $\hat{\mathbf{f}} = \mathbf{S}_\lambda \mathbf{Y}$ est :

$$\mathrm{COV}(\hat{\mathbf{f}}) = \sigma^2 \mathbf{S}_\lambda \mathbf{S}_\lambda^T. \tag{2.41}$$

En supposant la normalité des erreurs, il est possible d'utiliser (2.41) pour former des intervalles ponctuels d'écart-type pour le lissage. Il ne faut cependant pas considérer ces intervalles d'écart-type comme étant des intervalles de confiance puisqu'ils ne contiennent aucune information

sur le biais de $\hat{\mathbf{f}}$; il s'agit d'intervalles de confiance pour ce que le lisseur estime, c'est-à-dire $E\{\mathrm{S}_\lambda \mathbf{Y}\}$ (Buja et al., 1989). Mentionnons qu'il s'agit ici de la même problématique rencontrée, par exemple en hydrologie, pour l'estimation des intervalles de confiance des quantiles de crue estimés par une loi à deux ou à trois paramètres. Les intervalles de confiance des quantiles estimés par une loi à deux paramètres sont généralement plus petits mais seulement parce qu'ils ne tiennent pas compte du biais de modélisation introduit en utilisant une loi à deux paramètres plutôt qu'à trois paramètres.

Les résultats précédents ont été développés avec une hypothèse d'homogénéité de la variance. Cependant, comme on l'a vu précédemment, le modèle de régression locale peut s'adapter au problème de variances inégales ; $\mathrm{E}(\epsilon_i^2) = \sigma_i^2 < \infty$. Rappelons qu'avec la régression locale, puisque la fonction $w_i(x)$ est déjà une fonction de pondération, l'approche employée pour changer la procédure d'estimation consiste tout simplement à modifier la fonction $w_i(x)$ en la multipliant par l'inverse de la variance σ_i^2, c'est-à-dire en redéfinissant $w_i'(x) = w_i(x)/\sigma_i^2$. Ainsi, en posant $\mathbf{V}_x = \mathrm{diag}\{1/\sigma_1^2, 1/\sigma_2^2, \ldots, 1/\sigma_n^2\}$, on obtient le nouveau diagramme de pondération suivant :

$$\mathbf{L}_x'^T = \mathbf{e}_1^T(\mathbf{X}_x^T\mathbf{W}_x\mathbf{V}_x\mathbf{X}_x)^{-1}\mathbf{X}_x^T\mathbf{W}_x\mathbf{V}_x. \tag{2.42}$$

La variance de l'estimateur linéaire de l'équation (2.33), pour un modèle avec variances non homogènes (NH), devient :

$$\mathrm{V}_{\mathrm{NH}}(x) = \sum_{i=1}^{n} l_i(x)^2\sigma_i^2 = \mathbf{L}_x'^T\mathbf{V}_x^{-1}\mathbf{L}'_x \tag{2.43}$$

et en remplaçant $\mathbf{L}_x'^T$ par son expression en (2.42), on obtient :

$$\mathrm{V}_{\mathrm{NH}}(x) = \mathbf{e}_1^T(\mathbf{X}_x^T\mathbf{W}_x\mathbf{V}_x\mathbf{X}_x)^{-1}(\mathbf{X}_x^T\mathbf{W}_x^2\mathbf{V}_x\mathbf{X}_x)(\mathbf{X}_x^T\mathbf{W}_x\mathbf{V}_x\mathbf{X}_x)^{-1}\mathbf{e}_1. \tag{2.44}$$

2.3.4.4 L'erreur quadratique moyenne

À la section 2.3.3, nous avons présenté les notions d'erreur quadratique moyenne (EQM) et d'erreur quadratique de prédiction (EQP) pour le lissage en un point particulier x_i. En régression non paramétrique, il est souvent nécessaire d'avoir une mesure plus globale de ces erreurs. On s'intéresse ainsi à la moyenne de ces erreurs pour l'ensemble des valeurs ajustées $\hat{f}(x_i)$. Par

analogie avec les équations (2.31) et (2.32), on obtient :

$$\begin{aligned} \mathrm{EQM}(\lambda) &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{var}\{\hat{f}(x_i)\} + \frac{1}{n}\sum_{i=1}^{n}\mathrm{b}(x_i)^2 \\ \mathrm{EQM}(\lambda) &= \frac{\mathrm{tr}(\mathbf{S}_\lambda\mathbf{S}_\lambda^T)}{n}\sigma^2 + \frac{\mathbf{b}_\lambda^T\mathbf{b}_\lambda}{n} \end{aligned} \tag{2.45}$$

où tr(.), l'opérateur matriciel de trace, représente la somme des éléments de la diagonale de la matrice (i.e. ici, des variances). On obtient aussi :

$$\mathrm{EQP}(\lambda) = \left\{1 + \frac{\mathrm{tr}(\mathbf{S}_\lambda\mathbf{S}_\lambda^T)}{n}\right\}\sigma^2 + \frac{\mathbf{b}_\lambda^T\mathbf{b}_\lambda}{n} \tag{2.46}$$

Ces équations sont d'une grande utilité en régression non paramétrique puisqu'elles permettent d'ajuster le niveau de lissage, à l'aide de λ, et de choisir un niveau "optimal" de lissage λ^*. En effet, puisqu'il semble être approprié de choisir, parmi un ensemble de modèles, celui possédant la plus petite erreur quadratique de prédiction, un des objectifs de l'ajustement du niveau de lissage des modèles est d'obtenir la valeur λ^* qui minimise EQP(λ), notée

$$\lambda^* = \arg\min_{\lambda}\{\mathrm{EQP}(\lambda)\}. \tag{2.47}$$

De manière générale, en augmentant le niveau de lissage λ, on devrait s'attendre à observer une diminution de la variance combinée à une augmentation du biais du lissage. Il s'avère que cela est généralement le cas puisqu'en augmentant le niveau de lissage λ, la quantité $\mathrm{tr}(\mathbf{S}_\lambda\mathbf{S}_\lambda^T)$ a tendance à diminuer, alors que les éléments de $\mathbf{b}_\lambda$ ont plutôt tendance à augmenter (Hastie et Tibshirani, 1990).

2.3.4.5 Le nombre de degrés de liberté

Afin de permettre une comparaison des différents lisseurs linéaires sur un même base, il serait intéressant de savoir combien de **degrés de liberté** (une notion empruntée à la régression linéaire) ont été utilisés pour le lissage. Buja et al. (1989) ont proposé trois définitions, obtenues par analogie avec le modèle de régression linéaire, de la notion de degré de liberté, aussi appelée, le **nombre de paramètres effectifs** d'un lisseur linéaire.

Définition 1 : le nombre de degrés de liberté ν_1 d'un lisseur linéaire

$$\nu_1 = \mathrm{tr}(\mathbf{S}_\lambda\mathbf{S}_\lambda^T)$$

Pour le modèle de régression linéaire multiple, on a $\sum_{i=1}^{n}\mathrm{var}(\hat{y}_i) = p\sigma^2$, où le nombre de degrés de liberté p est le nombre de paramètres ou de variables prédictives du modèle. La définition analogue du nombre de degrés de liberté d'un lisseur linéaire (cf. éq. 2.45) est donc $\mathrm{tr}(\mathbf{S}_\lambda\mathbf{S}_\lambda^T)$.

Définition 2 : le nombre de degrés de liberté ν_2 d'un lisseur linéaire

$$\nu_2 = \mathrm{tr}(2\mathbf{S}_\lambda - \mathbf{S}_\lambda\mathbf{S}_\lambda^T)$$

Pour un lisseur linéaire, l'espérance de la somme du carré des erreurs SCE = $(\mathbf{f} - \hat{\mathbf{f}})^T(\mathbf{f} - \hat{\mathbf{f}})$ est :

$$\mathrm{E}\{\mathrm{SCE}\} = [n - \mathrm{tr}(2\mathbf{S}_\lambda - \mathbf{S}_\lambda\mathbf{S}_\lambda^T)]\sigma^2 + \mathbf{b}_\lambda^T\mathbf{b}_\lambda \tag{2.48}$$

Encore une fois, le nombre de degrés de liberté = $\mathrm{tr}(2\mathbf{S}_\lambda - \mathbf{S}_\lambda\mathbf{S}_\lambda^T)$ est égal à p pour l'espérance de la SCE du modèle de régression linéaire à p variables prédictives.

Définition 3 : le nombre de degrés de liberté ν_3 d'un lisseur linéaire

$$\nu_3 = \mathrm{tr}(\mathbf{S}_\lambda)$$

Une interprétation de la statistique C_p de Mallows est qu'elle corrige la SCE de manière à ce qu'elle soit non biaisée pour l'estimation de l'erreur quadratique de prédiction en ajoutant la quantité $2\,p\,\hat{\sigma}^2$, où $\hat{\sigma}^2$ est un estimateur non biaisé de σ^2 et p représente le nombre de paramètres du modèle de régression linéaire. Dans un contexte de lissage linéaire, la quantité que l'on doit alors ajouter est $2\,\mathrm{tr}(\mathbf{S}_\lambda)\,\hat{\sigma}^2$, d'où la définition 3.

2.3.5 L'ajustement du niveau de lissage : un compromis biais/variance

En régression non paramétrique, la largeur de la fenêtre de lissage est un des principaux facteurs qui contrôlent le compromis entre le biais et la variance de $\hat{f}(x)$. On présente, à la figure 2.2, le lissage des données décrites en 2.2.1 pour différentes largeurs h de la fenêtre de lissage. Lorsque la fenêtre de lissage est trop étroite (h=1), un nombre insuffisant d'observations est utilisé pour l'estimation de la courbe. Il en résulte un lissage très irrégulier avec beaucoup de bruit et une grande variabilité (variance). À l'inverse, une fenêtre de lissage trop grande (h=15) peut faire en sorte que de véritables caractéristiques de la courbe soient perdues. Le lissage est alors accompagné d'un fort biais. Pour illustrer mathématiquement ce phénomène, examinons l'exemple du lisseur par moyenne mobile présenté à la section 2.2.4.

En choisissant la fenêtre de lissage à l'aide de la méthode des (r) plus proches voisins symétriques, on peut écrire (cf. équation 2.9) :

$$\hat{f}(x_i) = \sum_{j \in N_r^S(x_i)} \frac{y_j}{2r+1}. \tag{2.49}$$

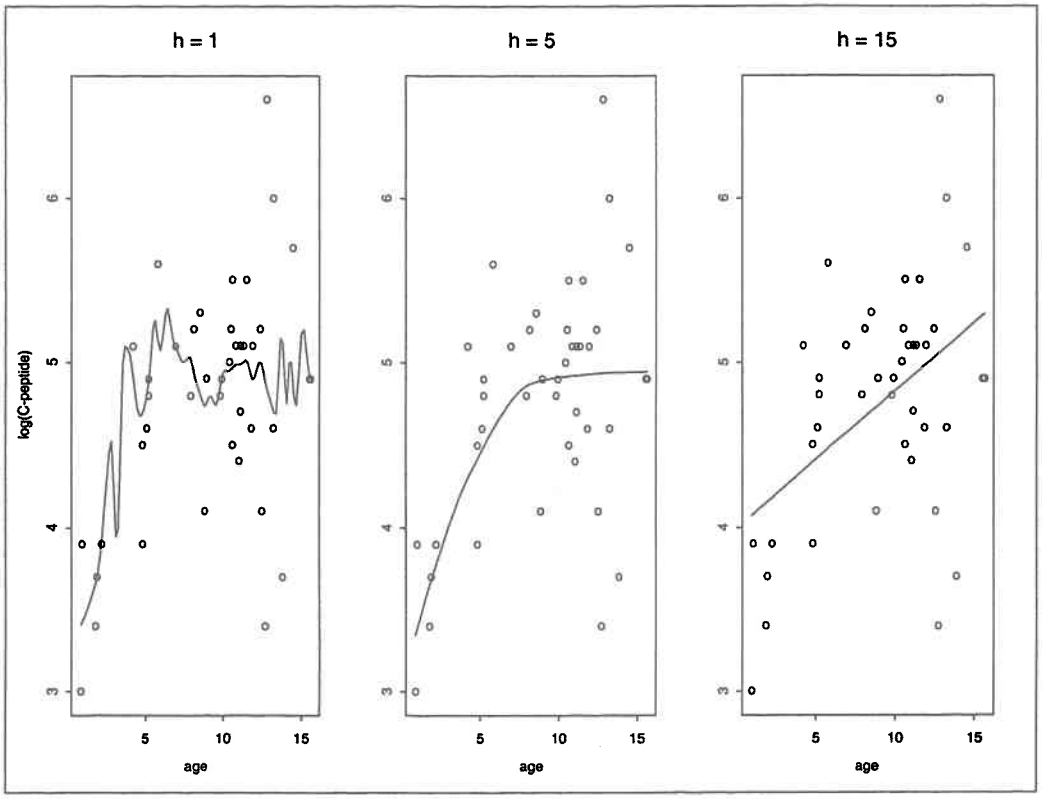

FIG. 2.2: Exemples de compromis biais/variance

Il s'agit ici de la moyenne des valeurs de y au point x_i ainsi qu'aux r plus proches voisins à droite et à gauche de x_i. La moyenne ainsi que la variance de cet estimateur sont respectivement :

$$E\{\hat{f}(x_i)\} = \frac{\sum_{j \in N_r^S(x_i)} f(x_j)}{2r+1}, \tag{2.50}$$

$$\text{var}\{\hat{f}(x_i)\} = \frac{\sigma^2}{2r+1}. \tag{2.51}$$

On peut remarquer qu'en augmentant r (ou de manière équivalente, la largeur de la fenêtre de lissage), la variance de l'estimateur diminue. Cependant, le biais a plutôt tendance à augmenter puisque l'espérance en (2.50) comporte alors plus de termes $f()$ dont la valeur diffère de $f(x_i)$. De manière équivalente, en diminuant r, on augmente la variance de l'estimateur alors que son biais à tendance à diminuer. Ce même comportement des estimateurs se produit aussi pour les différents autres lisseurs. Par exemple, pour les splines de lissage, le biais diminue et la variance augmente lorsque $\lambda \to 0$, et l'inverse se produit lorsque $\lambda \to \infty$ (Hastie et Tibshirani, 1990). Pour la régression locale, la largeur de la fenêtre de lissage h, le degré du polynôme local p de même que la fonction de pondération w influencent le compromis biais/variance.

2.3.5.1 La largeur de la fenêtre de lissage

Examinons l'effet causé par un changement de la fenêtre de lissage h à l'aide d'un exemple. Reprenons les données décrites en 2.2.1 et estimons pour différentes valeurs de h une droite en utilisant une fonction de pondération triplement cubique, c'est à dire utilisons le *loess* (cf. 2.2.7). La figure 2.3 présente le résultat de ces lissages. Pour la plus petite largeur de fenêtre, $h = 2$, la courbe estimée semble un peu trop ondulée. Pour les largeurs intermédiaires, $h = 4$ et $h = 8$, la courbe semble refléter assez bien l'allure générale des données. Par contre, lorsque $h = 16$, on peut observer ce qu'on appelle du **surlissage**, c'est-à-dire une estimation trop lisse et par le fait même biaisée, notamment dans l'intervalle d'âges à la limite inférieure de la courbe.

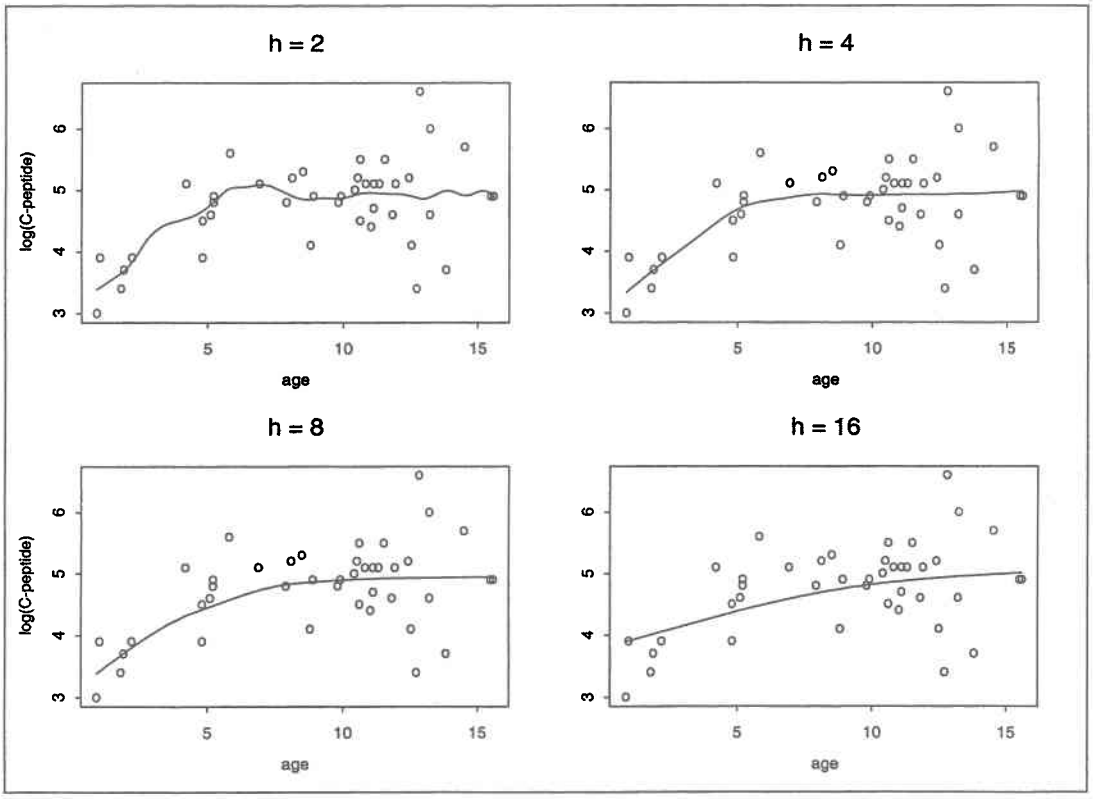

FIG. 2.3: Effet du changement de la largeur de la fenêtre de lissage

De manière générale, l'estimation aux limites est un problème associé à l'estimation non paramétrique. Une partie de ce problème est causée par le nombre généralement peu élevé de données observées en ces endroits. Pour faire face à cette problématique, une approche recommandée consiste à choisir une largeur de fenêtre qui puisse varier en fonction de la densité des observations disponibles. L'approche des k plus proches voisins permet cette variation. Avec cette approche, plutôt que de déterminer une largeur de fenêtre fixe $h(x) = h$, on la choisit de

manière à ce que $h(x)$ corresponde à la $k$ième plus petite distance $d(x, x_i) = |x - x_i|$ entre le point d'estimation x et les observations x_i. Remarquons que cette définition de $h(x)$ permet l'obtention d'une fenêtre de lissage symétrique, une contrainte et hypothèse (cf. 2.3.2) de la modélisation par régression locale. Notons d'ailleurs que la méthode des plus proches voisins symétriques n'est pas applicable en régression locale puisqu'on obtient alors généralement des fenêtres de lissage non symétriques de la forme $(x - h_1(x), x + h_2(x))$. Enfin, une autre façon de choisir $h(x)$ consiste à utiliser le maximum entre la valeur $h(x)$ obtenue par les k plus proches voisins et une valeur préalablement déterminée h. Cette approche permet de régler des problèmes de biais reliés à l'utilisation de trop faibles valeurs de h.

2.3.5.2 Le degré du polynôme local

Tout comme la largeur de la fenêtre de lissage, le degré du polynôme local utilisé influence le biais et la variance de l'estimateur $\hat{f}$. Wand et Jones (1990) ont montré que la performance asymptotique de $\hat{f}$ s'améliore lorsque la valeur de p augmente. Ils notent toutefois que puisque la variance de $\hat{f}$ augmente avec p, de très grands échantillons peuvent être nécessaires pour qu'il y ait, en pratique, une amélioration substantielle de la performance. Certains auteurs (e.g. Wand et Jones (1990), Loader (1999)) recommandent ainsi d'utiliser, au plus, un polynôme cubique.

La figure 2.4 montre des résultats de lissage pour des degrés de $p = 0$ à $p = 3$. Notons que le paramètre h a été choisi de manière à ce que chaque lissage ait un nombre de paramètres effectifs similaire, au sens de la définition 1 (cf. 2.3.4.5). En agissant ainsi, on se trouve à équilibrer la variabilité des estimateurs. On peut remarquer que même en ayant choisi ainsi le paramètre h, l'estimation locale par une constante $p = 0$ demeure moins lisse. Ce résultat est caractéristique de l'utilisation du lissage par régression locale lorsque $p = 0$. En effet, dans ce cas, un biais est introduit par la pente de la vraie fonction de régression (Loader, 1999). Plus la pente est élevée, plus l'estimation est biaisée. De plus, avec l'estimation locale par constante, notons le problème de biais d'ajustement, dont on a discuté à la section 2.2, près de la limite inférieure (à gauche). En raison du peu de différences visibles entre l'ajustement par polynômes de degrés $p = 1$, $p = 2$ et $p = 3$, l'examen visuel de la figure 2.4 suggère que l'utilisation d'un polynôme de degré $p = 1$ suffit pour estimer correctement la courbe $f(x)$.

2.3.5.3 La fonction de pondération

La fonction de pondération $W(u)$ influence le compromis biais/variance mais beaucoup moins que la largeur de la fenêtre de lissage ou que le degré du polynôme local. La fonction de

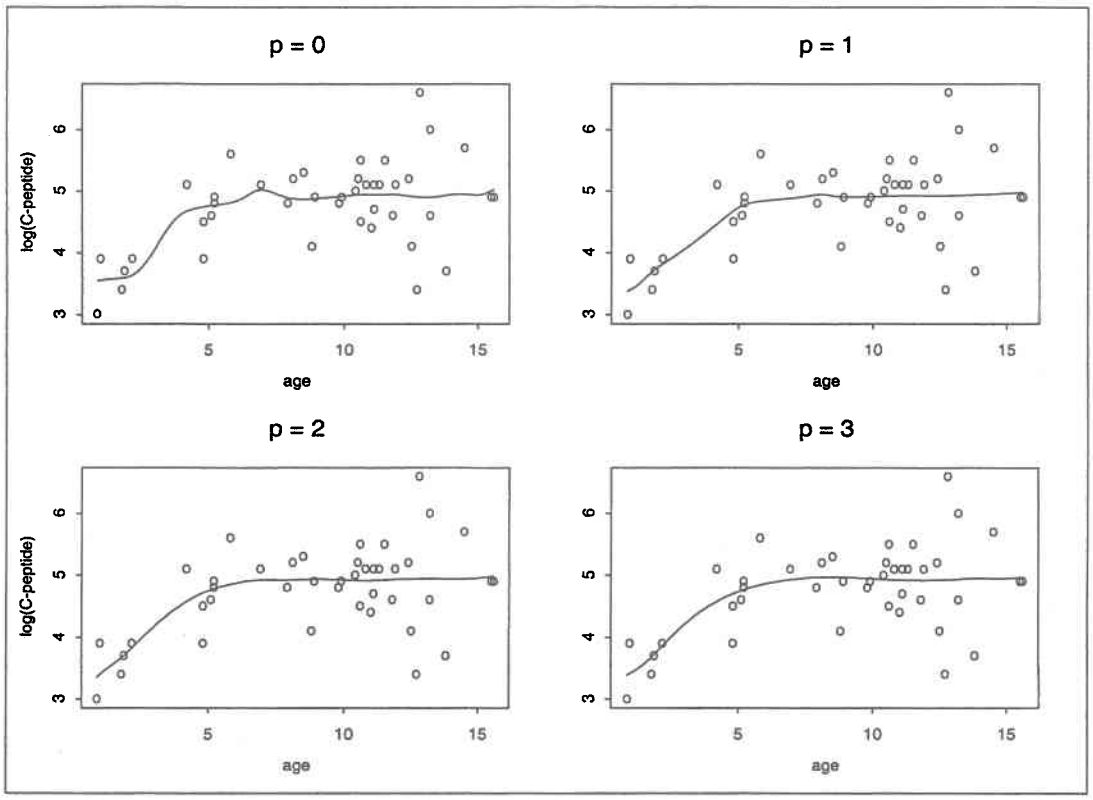

FIG. 2.4: Effet du changement du degré du polynôme local

pondération a cependant plus d'effet sur le niveau de lissage de la courbe ajustée. Par exemple, l'utilisation d'une fonction rectangulaire produit un ajustement comportant des discontinuités alors que les autres fonctions produisent généralement un ajustement beaucoup plus lisse.

Au tableau 2.2, on peut remarquer que les diverses fonctions de pondération rencontrées dans la littérature ont à peu près toutes la même efficacité asymptotique ponctuelle pour l'estimation lorsque $p = 1$ (pour plus de détails, voir le chapitre 13 de Loader (1999)). Bien que la fonction d'Epanechnikov soit la plus efficace, il est préférable d'utiliser les fonctions doublement quadratique, triplement cubique ou triplement pondérée puisqu'elles demeurent fortement efficace tout en produisant un meilleur lissage que la fonction d'Epanechnikov (Loader, 1999). De plus, il est important de rappeler que les statisticiens s'entendent sur le fait que le choix de la fonction de pondération W est relativement peu important en comparaison du choix de la largeur de la fenêtre de lissage h.

TAB. 2.2: Efficacité asymptotique ponctuelle des fonctions de pondération utilisées pour la régression locale linéaire (tiré de Loader (1999))

	eff(W)
Epanechnikov	1.000
Triplement cubique	0.998
Doublement quadratique	0.994
Triplement pondéré	0.987
Triangulaire	0.986
Normal	0.951
Rectangulaire	0.930

2.3.6 La sélection automatique des paramètres de lissage

Rappelons que puisqu'il semble approprié de choisir, parmi un ensemble de modèles, celui possédant la plus petite erreur quadratique de prédiction, un des objectifs de l'ajustement du niveau de lissage des modèles est de rechercher, parmi l'ensemble des paramètres de lissage λ, le vecteur de paramètres λ^* tel que $\lambda^* = \arg\min_\lambda\{\text{EQP}(\lambda)\}$ (cf. éq. 2.47). Pour le modèle de régression locale, on doit alors obtenir le vecteur de paramètres optimaux $\lambda^* \equiv (h^*, p^*, W^*)$.

Pour que la sélection des paramètres puisse être applicable en pratique, il est d'abord nécessaire de produire une estimation de l'EQP définie en (2.46). Intuitivement, on pourrait être porté à utiliser la moyenne du carré des erreurs (MCE) :

$$\text{MCE} = \frac{1}{n}\sum_{i=1}^{n}\{y_i - \hat{f}(x_i)\}^2 \tag{2.52}$$

mais cette mesure est beaucoup trop "optimiste" puisque les erreurs proviennent d'observations qui ont d'abord été utilisées pour l'estimation. Efron et Tibshirani (1993) nomment d'ailleurs ce type de mesure l'erreur **apparente** de prédiction. En statistique, de nombreux estimateurs de l'EQP ont été proposés : *Jackknife*, *Bootstrap*, *Bootstrap* .632, validation-croisée, statistique C_p de Mallows (voir par exemple Efron et Tibshirani (1993)). En régression non paramétrique, les estimateurs de l'EQP par validation croisée, par validation croisée généralisée ainsi qu'à l'aide d'une adaptation de la statistique C_p de Mallows constituent les méthodes les plus généralement utilisées.

2.3.6.1 La validation croisée

La validation croisée (VC) est un outil standard utilisé pour l'estimation de l'EQP. En régression non paramétrique, l'approche consiste à laisser de côté, l'un après l'autre, chacun des n couples (x_i, y_i) puis à estimer par lissage la courbe au point x_i à l'aide des $n-1$ couples restants. Le couple mis de côté est alors considéré comme une nouvelle observation au sens de l'équation (2.32). On obtient alors que l'estimateur par validation croisée de l'EQMP est :

$$\mathrm{VC}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\{y_i - \hat{f}_\lambda^{-i}(x_i)\}^2 \tag{2.53}$$

où $\lambda = (h, p, W)$ représente les paramètres de lissage et $\hat{f}_\lambda^{-i}(x_i)$ indique l'estimation par lissage au point x_i en laissant la $i$ième observation de côté. L'utilisation de cette procédure d'estimation peut être justifiée par le fait que $E\{\mathrm{VC}(\lambda)\} \approx \mathrm{EQP}(\lambda)$ (Hastie et Tibshirani, 1990).

2.3.6.2 La validation croisée pour un lisseur linéaire

En principe, le calcul de $\mathrm{VC}(\lambda)$ nécessite d'effectuer n estimations par lissage pour le calcul des n valeurs de $\hat{f}_\lambda^{-i}(x_i)$. Cependant, pour un lisseur linéaire $\hat{\mathbf{f}}_\lambda = \mathbf{S}_\lambda \mathbf{Y}$, on a la relation suivante :

$$y_i - \hat{f}_\lambda^{-i}(x_i) = \frac{y_i - \hat{f}_\lambda(x_i)}{1 - S_{ii}(\lambda)}. \tag{2.54}$$

Il n'est donc plus nécessaire d'effectuer n lissages puisque chacune des estimations $\hat{f}_\lambda^{-i}(x_i)$ peut être obtenue à partir de $\hat{f}_\lambda(x_i)$ et de $S_{ii}(\lambda)$. On a donc que pour un lisseur linéaire :

$$\mathrm{VC}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{y_i - \hat{f}_\lambda(x_i)}{1 - S_{ii}(\lambda)}\right\}^2. \tag{2.55}$$

Dans la littérature traitant de régression non paramétrique, on rencontre aussi la méthode de la validation croisée généralisée (VCG). Avec cette méthode, on obtient une approximation des éléments $S_{ii}(\lambda)$ de la diagonale de la matrice de lissage $\mathbf{S}_\lambda$ par la moyenne de ceux-ci, c'est-à-dire par la trace de la matrice $\mathbf{S}_\lambda$ divisé par n et ainsi, 2.55 s'écrit :

$$\mathrm{VCG}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{y_i - \hat{f}_\lambda(x_i)}{1 - \mathrm{tr}(\mathbf{S}_\lambda)/n}\right\}^2. \tag{2.56}$$

Cette approximation a permis, pour l'approche des splines de lissage, la réduction du temps de calcul puisqu'on ne savait pas, jusqu'à tout récemment, comment calculer les éléments $S_{ii}(\lambda)$ en $O(n)$ opérations (Hastie et Tibshirani, 1990).

2.3.6.3 La statistique C_p de Mallows

En régression linéaire multiple traditionnelle avec p variables prédictives, la statistique C_p (Mallows, 1973) constitue un estimateur de l'EQP. On définit alors cette statistique par :

$$C_p = \text{MCE} + 2p\hat{\sigma}^2/n \tag{2.57}$$

où $\hat{\sigma}^2$ est un estimateur de la variance des résidus. La statistique C_p, un cas spécial du critère d'information d'Akaike, s'obtient en ajustant la MCE de manière à ce qu'elle soit approximativement non biaisée pour l'erreur de prédiction : $E\{C_p\} \approx \text{EQP}$ (Efron et Tibshirani, 1993). En régression non paramétrique, en effectuant la même correction que précédemment, on obtient pour un lisseur linéaire :

$$C_p(\lambda) = \text{MCE}(\lambda) + 2\text{tr}(\mathbf{S}_\lambda)\hat{\sigma}^2/n. \tag{2.58}$$

Mentionnons qu'en ce qui concerne l'estimation de $\hat{\sigma}^2$, un estimateur non biaisé de σ^2 est donné par :

$$\hat{\sigma}^2 = \frac{\text{SCE}(\lambda)}{n - \nu_i} \tag{2.59}$$

où ν_i peut représenter l'une ou l'autre des trois définitions du nombre de paramètres effectifs d'un lisseur linéaire présentées précédemment (cf. 2.3.4.5). Une autre approche consiste à calculer l'estimateur en (2.59) en utilisant un paramètre de lissage λ^*, différent de λ, de manière à n'effectuer qu'un faible niveau de lissage. D'autres estimateurs ont aussi été proposés (pour plus de détails, voir par exemple la section 3.5 de Opsomer (1995)).

3. LA MODÉLISATION ADDITIVE

Dans le chapitre précédent, nous avons présenté les notions de base reliées à l'estimation non paramétrique par lissage des courbes de régression, c'est-à-dire le lissage unidimensionnel avec une seule variable explicative ($d = 1$). Bien que les différentes techniques de lissage peuvent être généralisées pour l'estimation de surface de régression ($d > 1$), cette façon de faire est souvent problématique en raison notamment du problème de la raréfaction des données dans un espace à grande dimension. Dans ce chapitre, nous proposons une approche de modélisation non paramétrique qui ne souffre pas de ce problème : la modélisation additive. Nous introduisons cette approche dans la section 3.1 pour ensuite présenter formellement le modèle additif dans la section 3.2. Par la suite, nous montrons dans la section 3.3 comment s'effectue l'estimation des modèles additifs pour enfin discuter, dans la section 3.4, de la modélisation des données, c'est-à-dire des éléments à prendre en considération lors du choix : (1) des variables explicatives à inclure dans le modèle et (2) du niveau de lissage associé à chacune de ces variables explicatives.

3.1 L'approche de modélisation additive

Rappelons le résultat classique selon lequel la fonction f qui minimise $E\{Y - f(X)\}^2$ est l'espérance conditionnelle de Y étant donné X, soit

$$f(X) = E(Y|X). \tag{3.1}$$

Dans le cas où l'on considère plusieurs variables prédictives, un des outils les plus utilisés pour représenter cette fonction est le modèle de régression linéaire multiple suivant :

$$E(Y|X_1, X_2, ..., X_d) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_d X_d \tag{3.2}$$

où α et $\beta_{i\,(i=1...d)}$ sont les paramètres du modèle. Dans le modèle linéaire (3.2), on fait une hypothèse forte à propos de la dépendance de $E(Y)$ sur $X_1, \ldots, X_d$, soit que la dépendance est linéaire pour chacune des d variables prédictives. À l'inverse, avec l'approche de régression non paramétrique présentée au chapitre précédent, aucune hypothèse n'est faite quant à la forme de la surface de régression si ce n'est le fait qu'elle soit lisse. Rappelons qu'on a alors le modèle suivant :

$$E(Y|X_1, X_2, ..., X_d) = f(X_1, X_2, ..., X_d) \tag{3.3}$$

où la fonction $f(X_1, X_2, ..., X_d)$ peut être estimée par une des techniques de lissage présentée à la section 2.2.8.

Un problème commun aux différents lisseurs de surface concerne la caractéristique de localité de l'estimation en grandes dimensions. Ce problème de dimensionalité est bien connu en régression non paramétrique sous le nom de ***curse of dimensionality*** : les voisinages avec un nombre fixé d'observations deviennent de moins en moins locaux lorsque la dimension augmente (Bellman, 1961). La figure 3.1 illustre ce phénomène. Supposons que les observations soient réparties uniformément à l'intérieur d'un cube unitaire de dimension p et que nous voulions construire un voisinage en forme de cube débutant par l'origine de manière à capturer (en moyenne) (100 x GRV)% des observations où GRV représente la grandeur relative du voisinage (i.e. le *span*) (cf. 2.2.4). On peut alors montrer que la dimension du sous-cube devrait avoir une arête de longueur $\text{GRV}^{1/p}$. Pour p=1 et GRV=0,1 on obtient une arête de 0,1. Par contre, pour p=10 (et GRV=0,1) on obtient environ 0,8 ce qui n'est plus vraiment local (même avec un sous-cube composé d'arêtes de longueur 0,8, on ne réussit, en dimension p, qu'à capturer que 10% des observations).

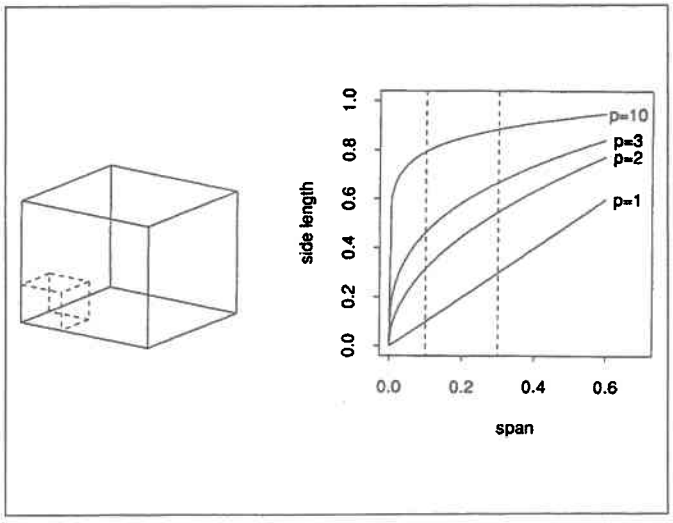

FIG. 3.1: Illustration du problème de dimensionalité (tiré de Hastie et Tibshirani (1990))

Au cours des deux dernières décennies, de nombreuses techniques de régression non paramétrique ont été développées afin de répondre à ce problème de dimensionalité (e.g. le modèle de la poursuite par projection de Friedman et Stuetzle (1981), le modèle d'alternance de l'espérance conditionnelle de Breiman et Friedman (1985), le modèle additif généralisé de Hastie et Tib-

shirani (1990), etc.) (pour une description des principales méthodes, voir, par exemple, Härdle (1989)). Un de ces modèles qui se situe à mi-chemin entre la régression linéaire multiple (équation 3.2) et le lissage de surfaces multidimensionnelles (équation 3.3), a plus particulièrement retenu l'attention des praticiens, il s'agit du **modèle additif** (Hastie et Tibshirani, 1987; Buja et al., 1989; Hastie et Tibshirani, 1990) :

$$E(Y|X_1, X_2, ..., X_d) = \alpha + f_1(X_1) + f_2(X_2) + ... + f_d(X_d) \tag{3.4}$$

où les f_i sont des fonctions unidimensionnelles lisses, de forme non spécifiée a priori, associées respectivement à chacune des variables explicatives $X_{i\,(i=1...d)}$. Ce modèle est caractérisé par le fait qu'il n'est pas affecté par le problème de dimensionalité. En effet, Stone (1982) a étudié la convergence des modèles additifs et a prouvé que la vitesse optimale de convergence pouvant être atteinte (pour f de classe C_2) est de l'ordre de $O(n^{-4/5})$, soit la même que pour l'estimation d'une fonction unidimensionnelle ($d = 1$) (cf. 1.1.3). Une autre caractéristique de la modélisation additive qui distingue d'ailleurs le modèle additif des autres techniques de régression non paramétrique est relative à la facilité d'interprétation des résultats. En effet, tout comme en régression linéaire multiple, il est possible avec un modèle additif d'examiner graphiquement l'effet de chacune des variables prédictives, une à la fois, conditionnellement à la présence des autres variables prédictives. Il devient alors possible d'effectuer une validation des hypothèses de linéarité émises par le modèle de régression linéaire multiple. De plus, l'implantation de la modélisation additive dans le logiciel S-PLUS (Chambers et Hastie, 1992) a grandement contribué à son essor.

3.2 Le modèle additif

La modélisation additive est utilisée pour représenter une relation entre d variables prédictives (ou indépendantes) $X_1, X_2, \ldots, X_d$ et une variable de réponse Y (ou variable dépendante) qui est reliée à ces variables prédictives. Supposons qu'on ait l'ensemble d'observations suivant $\{((X_1^i, X_2^i, ..., X_d^i), Y_i), i = 1, 2, ..., n\}$, alors la forme générale du modèle additif est :

$$Y_i = \alpha + f_1(X_1^i) + f_2(X_2^i) + ... + f_d(X_d^i) + \epsilon_i \tag{3.5}$$

où les erreurs ϵ_i sont indépendantes des X_j^i et identiquement distribuées de moyenne $E(\epsilon_i) = 0$, de variance $E(\epsilon_i^2) = \sigma^2$ et où chacune des fonctions f_j telle que $E(f_j) = 0$ est une fonction de forme non spécifiée pouvant être estimée par lissage univarié. Dans ce document, nous supposerons que les lisseurs utilisés pour l'estimation des fonctions f_j sont **linéaires**. De plus, nous relaxerons l'hypothèse de variance constante σ^2 et permettrons que $E(\epsilon_i^2) = \sigma_i^2$.

3.3 L'estimation des modèles additifs

En s'intéressant à l'ajustement d'un modèle additif aux valeurs observées, on peut noter le vecteur des valeurs estimées :

$$\hat{\mathbf{f}} = \hat{\alpha} + \hat{\mathbf{f}}_1 + \hat{\mathbf{f}}_2 + ... + \hat{\mathbf{f}}_d \tag{3.6}$$

où $\hat{\alpha} = \bar{\mathbf{Y}}$ et $\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, ..., \hat{\mathbf{f}}_d$ sont des lisseurs linéaires univariés tels que $\hat{\mathbf{f}}_j = \mathbf{S}_j\mathbf{Y}_{(j=1...d)}$ (cf. 2.3.4.1) où $\mathbf{S}_1$,$\mathbf{S}_2$, ..., $\mathbf{S}_d$ sont les matrices de lissage utilisées respectivement pour le lissage dans l'espace des variables explicatives $X_1, X_2, ..., X_d$. Pour des lisseurs linéaires, il est possible de montrer (Buja et al., 1989) que les estimateurs $\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, ..., \hat{\mathbf{f}}_d$ constituent les solutions du système d'équations normales suivant :

$$\begin{bmatrix} \mathbf{I} & \mathbf{S}_1 & \dots & \mathbf{S}_1 \\ \mathbf{S}_2 & \mathbf{I} & \dots & \mathbf{S}_2 \\ \vdots & & \ddots & \vdots \\ \mathbf{S}_d & \mathbf{S}_d & \dots & \mathbf{I} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{f}}_1 \\ \hat{\mathbf{f}}_2 \\ \vdots \\ \hat{\mathbf{f}}_d \end{bmatrix} = \begin{bmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \\ \vdots \\ \mathbf{S}_d \end{bmatrix} (\mathbf{Y} - \hat{\alpha}) \tag{3.7}$$

De plus, si la matrice

$$\mathbf{M} = \begin{bmatrix} \mathbf{I} & \mathbf{S}_1 & \dots & \mathbf{S}_1 \\ \mathbf{S}_2 & \mathbf{I} & \dots & \mathbf{S}_2 \\ \vdots & & \ddots & \vdots \\ \mathbf{S}_d & \mathbf{S}_d & \dots & \mathbf{I} \end{bmatrix} \tag{3.8}$$

est inversible, alors les estimateurs $\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, ..., \hat{\mathbf{f}}_d$ existent et

$$\begin{bmatrix} \hat{\mathbf{f}}_1 \\ \hat{\mathbf{f}}_2 \\ \vdots \\ \hat{\mathbf{f}}_d \end{bmatrix} = \mathbf{M}^{-1} \begin{bmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \\ \vdots \\ \mathbf{S}_d \end{bmatrix} (\mathbf{Y} - \hat{\alpha}) \tag{3.9}$$

Puisque les matrices $\mathbf{S}_{i\ (i=1,...,d)}$ sont de dimension $n \times n$ et que chacune des lignes (et colonnes) de $\mathbf{M}$ est composée de d de ces matrices, $\mathbf{M}$ est de dimension $nd \times nd$. En théorie, il est possible de résoudre le système d'équations (3.7) à l'aide d'une méthode non-itérative telle la méthode de la décomposition QR (Burden et Faires, 1989). Un problème pratique est que (3.7) est un système de dimension $nd \times nd$ et que des méthodes telles la décomposition QR requièrent $O(m^3)$ opérations pour résoudre un système de dimension $m \times m$, ce qui signifie ici

$O((nd)^3)$ opérations. Une méthode itérative, l'**algorithme de *backfitting***, peut alors être utilisée afin de diminuer le nombre d'opérations nécessaires à l'estimation. L'algorithme de *backfitting*, une méthode d'itération de type Gauss-Seidel (Burden et Faires, 1989), ne requiert, lorsque les lisseurs peuvent être estimés en $O(n)$ opérations (c'est le cas pour la régression locale linéaire), que $O(np)$ opérations (Hastie et Tibshirani, 1990).

3.3.1 L'algorithme de *backfitting*

Pour $j = 1, 2, ..., d$, notons $\mathbf{X}_j = (X_j^1, X_j^2, ..., X_j^n)^T$, $\mathbf{f}_j = (f_j(X_j^1), f_j(X_j^2), ..., f_j(X_j^n))$ et $\mathbf{S}_j$ la matrice de lissage dont les lignes sont les diagrammes de pondération (cf. section 2.3.4.1) aux valeurs observées $X_j^1, X_j^2, ..., X_j^n$. L'algorithme de *backfitting*, présenté à la figure 3.2, estime les fonctions f_j du modèle additif (3.4) $E(Y|X_1, X_2, ..., X_d) = \alpha + f_1(X_1) + f_2(X_2) + ... + f_d(X_d)$. Pour débuter l'algorithme, il est nécessaire de définir des valeurs initiales pour $\mathbf{f}_j^0$. Hastie et Tibshirani (1990) suggèrent, lorsqu'aucune information n'est disponible a priori sur la forme de ces fonctions, d'utiliser les fonctions linéaires obtenues par régression de Y sur les variables explicatives.

1. Initialiser $\mathbf{f}_j = \mathbf{f}_j^0, j = 1, 2, ..., d, \hat{\alpha} = \bar{\mathbf{Y}}$

2. Faire pour $j = 1, 2, ..., d$
   $\mathbf{f}_j = \mathbf{S}_j(\mathbf{Y} - \hat{\alpha} - \sum_{k \neq d} \mathbf{f}_k)$

3. Répéter 2. jusqu'à ce que les fonctions $\mathbf{f}_j$ convergent

FIG. 3.2: Algorithme de *backfitting*

Buja et al. (1989) ont donné des conditions suffisantes, basées sur la notion de rétrécissement (*shrinking*), garantissant la convergence de l'algorithme. Par définition, pour toute norme matricielle $||.||$, on nomme **lisseur rétrécissant** la matrice $\mathbf{S}$, si pour tout vecteur $\mathbf{y}$, $||\mathbf{Sy}|| \leq ||\mathbf{y}||$. Pour un **lisseur strictement rétrécissant**, on a $||\mathbf{Sy}|| < ||\mathbf{y}||$ pour tout $\mathbf{y}$. Buja et al. (1989) ont montré que si tous les lisseurs utilisés dans l'algorithme sont rétrécissants, la convergence est garantie et si de plus, les lisseurs sont strictement rétrécissants, alors la convergence vers une solution unique est assurée. Dans le cas des lisseurs par régression locale, il n'y a pas garantie de convergence bien que les contre-exemples soient difficiles à trouver.

Un résultat intéressant, obtenu par Hastie et Tibshirani (1990), est qu'un lisseur rétrécissant peut être transformé en un lisseur strictement rétrécissant, ce qui garantit l'unicité de la solution, en effectuant un centrage, c'est-à-dire en remplaçant les matrices de lissage $\mathbf{S}_j$ par des **lisseurs centrés** $\mathbf{S}_j^* = (\mathbf{I} - \mathbf{1}\mathbf{1}^T/n)\mathbf{S}_j$ où $\mathbf{1}$ est une matrice unitaire de dimension $n \times n$. Notons qu'en procédant ainsi, nous nous trouvons alors à employer l'algorithme de *backfitting* modifié, présenté à la figure 3.3, qui estime plutôt les fonctions f_j du modèle additif sans constante suivant : $E(Y|X_1, X_2, ..., X_d) = f_1(X_1) + f_2(X_2) + ... + f_d(X_d)$.

1. Initialiser $\mathbf{f}_j = \mathbf{f}_j^0, j = 1, 2, ..., d$

2. Faire pour $j = 1, 2, ..., d$
   $\mathbf{f}_j = \mathbf{S}_j^*(\mathbf{Y} - \sum_{k \neq d} \mathbf{f}_k)$

3. Répéter 2. jusqu'à ce que les fonctions $\mathbf{f}_j$ convergent

FIG. 3.3: Algorithme de *backfitting* modifié

3.3.2 Définitions relatives aux modèles additifs de lisseurs linéaires

Par analogie avec les résultats présentés à la section 2.3.4 sur les lisseurs linéaires, nous présentons ici des définitions analogues pour un modèle additif composé de lisseurs linéaires.

3.3.2.1 Des estimateurs linéaires

Notons d'abord que le fait d'effectuer un centrage des lisseurs linéaires revient à estimer non pas le système d'équations (3.7) mais plutôt le système suivant :

$$\begin{bmatrix} \mathbf{I} & \mathbf{S}_1^* & \dots & \mathbf{S}_1^* \\ \mathbf{S}_2^* & \mathbf{I} & \dots & \mathbf{S}_2^* \\ \vdots & & \ddots & \vdots \\ \mathbf{S}_d^* & \mathbf{S}_d^* & \dots & \mathbf{I} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{f}}_1 \\ \hat{\mathbf{f}}_2 \\ \vdots \\ \hat{\mathbf{f}}_d \end{bmatrix} = \begin{bmatrix} \mathbf{S}_1^* \\ \mathbf{S}_2^* \\ \vdots \\ \mathbf{S}_d^* \end{bmatrix} \mathbf{Y} \tag{3.10}$$

En supposant que la matrice

$$\mathbf{M}^* = \begin{bmatrix} \mathbf{I} & \mathbf{S}_1^* & \dots & \mathbf{S}_1^* \\ \mathbf{S}_2^* & \mathbf{I} & \dots & \mathbf{S}_2^* \\ \vdots & & \ddots & \vdots \\ \mathbf{S}_d^* & \mathbf{S}_d^* & \dots & \mathbf{I} \end{bmatrix} \tag{3.11}$$

est inversible, alors les estimateurs $\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, ..., \hat{\mathbf{f}}_d$ existent et sont uniques ; ils sont donnés par :

$$\begin{bmatrix} \hat{\mathbf{f}}_1 \\ \hat{\mathbf{f}}_2 \\ \vdots \\ \hat{\mathbf{f}}_d \end{bmatrix} = \mathbf{M}^{*-1} \begin{bmatrix} \mathbf{S}_1^* \\ \mathbf{S}_2^* \\ \vdots \\ \mathbf{S}_d^* \end{bmatrix} \mathbf{Y} \tag{3.12}$$

L'équation (3.12) permet d'observer que les estimateurs sont linéaires en Y. Les estimateurs produits par la procédure de *backfitting*, lorsque la convergence est atteinte, sont aussi linéaires en Y. Il est donc possible d'obtenir $\hat{\mathbf{f}}_1 = \mathbf{R}_1\mathbf{Y}$, $\hat{\mathbf{f}}_2 = \mathbf{R}_2\mathbf{Y}$, et ainsi de suite d'où :

$$\hat{\mathbf{f}} = \hat{\alpha}\mathbf{1} + \hat{\mathbf{f}}_1 + \hat{\mathbf{f}}_2 + \ldots + \hat{\mathbf{f}}_d = \mathbf{RY} \tag{3.13}$$

où $\hat{\mathbf{f}}$ est le vecteur des valeurs ajustées. Avec la procédure de *backfitting*, les matrices $\mathbf{R}_j$ et $\mathbf{R}$ peuvent être obtenues en applicant l'algorithme à la séquence de vecteurs de réponses $\mathbf{e}_1 = (1, 0, 0, ..., 0)^T, \mathbf{e}_2 = (0, 1, 0, ..., 0)^T, ..., \mathbf{e}_n = (0, 0, ..., 1)^T$ (pour plus de détails, voir Hastie et Tibshirani (1987)).

3.3.2.2 Le biais, la variance et l'erreur quadratique moyenne

En considérant que les estimateurs du modèle additif sont linéaires en Y, on obtient, par analogie avec l'équation (2.37), comme expression du vecteur de biais :

$$\mathbf{b} = \mathbf{f} - E\{\mathbf{RY}\} = \mathbf{f} - \mathbf{Rf} \tag{3.14}$$

Mentionnons que pour une fonction arbitraire et inconnue f, les modèles additifs sont biaisés bien qu'ils puissent toutefois être non biaisés pour une classe particulière de fonctions.

En considérant que les observations sont indépendantes, identiquement distribuées et de variance constante σ^2, alors la matrice de variance-covariance des valeurs ajustées $\hat{\mathbf{f}} = \mathbf{RY}$ est :

$$\mathrm{COV}(\hat{\mathbf{f}}) = \sigma^2\mathbf{R}\mathbf{R}^T. \tag{3.15}$$

En supposant la normalité des erreurs, il est possible d'utiliser (3.15) pour obtenir des intervalles ponctuels d'écart-type pour les estimateurs du modèle additif. De plus, il est possible d'obtenir des intervalles ponctuels d'écart-type pour chacun des lisseurs univariés $\hat{\mathbf{f}}_j$ en utilisant le fait que

$$\mathrm{COV}(\hat{\mathbf{f}}_j) = \sigma^2\mathbf{R}_j\mathbf{R}_j^T. \tag{3.16}$$

Rappelons qu'il ne faut cependant pas considérer ces intervalles d'écart-type comme étant des intervalles de confiance puisqu'ils ne contiennent respectivement aucune information sur le biais de $\hat{\mathbf{f}}$ et de $\hat{\mathbf{f}}_j$; il s'agit d'intervalles de confiance pour ce qui est estimé (cf. 2.3.4.3).

Les résultats précédents ont été développés avec une hypothèse d'homoscédasticité. Cependant, comme on l'a mentionné précédemment, le modèle additif peut s'adapter au cas des variances inégales ; $\mathrm{E}(\epsilon_i^2) = \sigma_i^2$. La procédure consiste alors à ramener ce problème d'hétéroscédasticité au cas traditionnel avec homoscédasticité en effectuant les transformations suivantes : $\mathbf{Y}' = \mathbf{V}^{1/2}\mathbf{Y}$ et $\mathbf{f}_j' = \mathbf{V}^{1/2}\mathbf{f}_j$ où $\mathbf{V} = \mathrm{diag}\{1/\sigma_1^2, 1/\sigma_2^2, \ldots, 1/\sigma_n^2\}$ (pour plus de détails, voir la section 5.4 de Hastie et Tibshirani (1990)).

En ce qui concerne les notions d'erreur quadratique moyenne et d'erreur quadratique de prédiction, il est aussi possible d'obtenir, par analogie avec les équations (2.45) et (2.46) :

$$\mathrm{EQM} = \frac{\mathrm{tr}(\mathbf{R}\mathbf{R}^T)}{n}\sigma^2 + \frac{\mathbf{b}^T\mathbf{b}}{n} \tag{3.17}$$

et

$$\mathrm{EQP} = \left\{1 + \frac{\mathrm{tr}(\mathbf{R}\mathbf{R}^T)}{n}\right\}\sigma^2 + \frac{\mathbf{b}^T\mathbf{b}}{n} \tag{3.18}$$

3.3.2.3 Le nombre de degrés de liberté

Chacune des 3 définitions présentées en 2.3.4.5 a son analogue naturel ici : $\nu_1 = \mathrm{tr}(\mathbf{R}\mathbf{R}^T)$, $\nu_2 = \mathrm{tr}(2\mathbf{R} - \mathbf{R}\mathbf{R}^T)$ et $\nu_3 = \mathrm{tr}(\mathbf{R})$. Deux facteurs limitent cependant l'utilisation pratique de ces définitions. D'abord, le calcul de ces définitions est très exigeant du point de vue du nombre d'opérations. De plus, un inconvénient majeur avec ces définitions est qu'elles ne déterminent que le nombre total de degrés de liberté (ou de paramètres effectifs) de la modélisation additive. En pratique, il est plutôt souhaitable de connaître le nombre de degrés de liberté associé à chacun des lisseurs, c'est-à-dire à chacune des variables explicatives. En ce sens, Hastie et Tibshirani (1990) suggèrent d'utiliser :

$$\nu_2^j = \mathrm{tr}(2\mathbf{S}_j - \mathbf{S}_j\mathbf{S}_j^T) - 1 \tag{3.19}$$

ou

$$\nu_3^j = \mathrm{tr}(\mathbf{S}_j) - 1 \tag{3.20}$$

comme mesure du nombre de degrés de liberté associé à la variable prédictive X_j. Ces définitions sont basées respectivement sur les approximations suivantes :

$$\nu_2 = \mathrm{tr}(2\mathbf{R} - \mathbf{R}\mathbf{R}^T) \approx 1 + \sum_{j=1}^{p} \left\{ \mathrm{tr}(2\mathbf{S}_j - \mathbf{S}_j\mathbf{S}_j^T) - 1 \right\} \tag{3.21}$$

et

$$\nu_3 = \mathrm{tr}(\mathbf{R}) \approx 1 + \sum_{j=1}^{p} \left\{ \mathrm{tr}(\mathbf{S}_j) - 1 \right\} \tag{3.22}$$

Ainsi, le nombre total de degrés de liberté est représenté par la somme des degrés de liberté associés à chacune des variables explicatives à laquelle on rajoute un pour tenir compte de l'estimation de la constante. Ces approximations donnent généralement de bons résultats sauf dans le cas de fortes corrélations entre les variables prédictives ou lorsque de très petits paramètres de lissage sont utilisés (Hastie et Tibshirani, 1990).

3.4 La modélisation des données

Au chapitre précédent (cf. 2.3.5, 2.3.6), nous avons discuté de l'ajustement du niveau de lissage des modèles de régression locale à une seule variable prédictive. Rappelons que dans cette situation, une approche retenue pour la sélection du modèle final consiste à choisir, parmi un ensemble de modèles $\hat{\mathbf{f}} = \mathbf{S}_\lambda \mathbf{Y}$, celui possédant la plus petite erreur de prédiction, c'est-à-dire qu'il s'agit de rechercher, parmi l'ensemble des paramètres de lissage λ, le vecteur de paramètres λ^* tel que $\lambda^* = \arg\min_\lambda \{\mathrm{EQP}(\lambda)\}$. Rappelons aussi que pour que la sélection des paramètres puisse être applicable en pratique, des mesures (estimations) de l'EQP sont nécessaires et ont d'ailleurs été présentées (cf. 2.3.6).

De manière équivalente, pour un modèle additif, il semble approprié de choisir, parmi un ensemble de modèles $\hat{\mathbf{f}} = \mathbf{S}_{\lambda_1}\mathbf{Y} + \mathbf{S}_{\lambda_2}\mathbf{Y} + ... + \mathbf{S}_{\lambda_d}\mathbf{Y}$, celui possédant la plus petite erreur de prédiction. Il s'agit ici de rechercher, parmi l'ensemble de vecteurs de paramètres de lissage $\Lambda = (\lambda_1, \lambda_2, ..., \lambda_d)$, le vecteur de paramètres Λ^* tel que $\Lambda^* = \arg\min_\Lambda \{\mathrm{EQP}(\Lambda)\}$. Encore une fois, pour que la sélection des paramètres puisse être applicable en pratique, des mesures de l'EQP pour un modèle additif sont nécessaires.

3.4.1 Les mesures de l'ajustement

Rappelons qu'intuitivement, on pourrait être porté à utiliser, comme mesure de l'ajustement, la moyenne du carré des erreurs (MCE) :

$$\mathrm{MCE}(\Lambda) = \frac{1}{n}\sum_{i=1}^{n}\left\{y_i - \sum_{j=1}^{p}\hat{f}_{j,\lambda_j}(x_{ij})\right\}^2 \tag{3.23}$$

mais que cette mesure est beaucoup trop "optimiste" puisque les erreurs proviennent d'observations qui ont d'abord été utilisées pour l'estimation. Les estimateurs de l'EQP par validation croisée, par validation croisée généralisée ainsi qu'à l'aide d'une adaptation de la statistique C_p de Mallows développées dans un contexte de régression non paramétrique constituent aussi des méthodes généralement utilisées en modélisation additive.

Pour un modèle additif, on peut définir, par analogie à l'équation (2.56), le critère de validation croisée généralisée (VCG) par :

$$\mathrm{VCG}(\Lambda) = \frac{\mathrm{MCE}(\Lambda)}{(1 - \mathrm{tr}(\mathbf{R}_\Lambda)/n)^2}. \tag{3.24}$$

Gu et Wahba (1988) ont développé un algorithme permettant, pour des splines de lissage, le calcul de $\mathrm{VCG}(\Lambda)$. Cependant, en raison de la difficulté à calculer de manière efficace le dénominateur de l'expression (3.24), le calcul de $\mathrm{VCG}(\Lambda)$ requiert $O(n^3)$ opérations. Il est cependant possible, afin de réduire le temps de calcul, d'utiliser l'approximation (3.22) afin d'obtenir :

$$\mathrm{VCG}^*(\Lambda) = \frac{\mathrm{MCE}(\Lambda)}{(1 - [1 + \sum_{j=1}^{p}\{\mathrm{tr}(\mathbf{S}_{\lambda_j}) - 1\}]/n)^2} \tag{3.25}$$

qui ne requiert que $O(n)$ opérations.

Dans le contexte de l'utilisation de l'analyse discriminante pour la sélection de modèles, Hastie et al. (1993) ont proposé, comme mesure permettant la discrimination, l'utilisation de

$$\mathrm{VCG}^*(c, \Lambda) = \frac{\mathrm{MCE}(\Lambda)}{(1 - [1 + c\sum_{j=1}^{p}\{\mathrm{tr}(\mathbf{S}_{\lambda_j}) - 1\}]/n)^2} \tag{3.26}$$

où c représente un facteur de pénalité (une fonction de coût) associé au nombre de paramètres d'un modèle additif. Notons que VCG^* implique un facteur de pénalité unitaire alors que de nombreux auteurs (e.g. Friedman (1991), Owen (1991), Hastie et al. (1993)) mentionnent que $c = 2$ constitue plutôt une valeur appropriée.

Une dernière mesure, connue dans la littérature sous le nom de AIC (*Akaike Information Criterion*) de Hastie (Hastie et Tibshirani, 1990), est en fait la statistique de C_p de l'équation (2.58) adaptée à la modélisation additive. On définit cette statistique par :

$$\mathrm{AIC}(\Lambda) = \mathrm{MCE}(\Lambda) + 2\sum_{j=1}^{p}\left\{\mathrm{tr}(\mathbf{S}_{\lambda_j}) - 1\right\}\hat{\sigma}^2/n \tag{3.27}$$

où $\hat{\sigma}^2$ est un estimateur de la variance des résidus. Mentionnons qu'il s'agit de la mesure employée dans le logiciel S-Plus pour la discrimination de différents modèles additifs. En ce qui concerne l'estimation de $\hat{\sigma}^2$, de nombreux estimateurs sont disponibles dont par exemple, l'estimateur non biaisé suivant :

$$\hat{\sigma}^2 = \frac{\mathrm{SCE}(\Lambda)}{n - \nu_i} \tag{3.28}$$

où ν_i peut représenter l'une ou l'autre des définitions, présentées en 3.3.2.3, du nombre total de paramètres effectifs d'un modèle additif.

3.4.2 Les procédures de sélection du modèle

Supposons que l'on ait à sélectionner un modèle optimal, au sens de l'un des critères énumérés précédemment, parmi l'ensemble des modèles additifs composés de polynômes locaux dont la forme générale est :

$$E(Y|X_1, X_2, ..., X_d) = \mathbf{S}_{\lambda_1}\mathbf{Y} + \mathbf{S}_{\lambda_2}\mathbf{Y} + ... + \mathbf{S}_{\lambda_d}\mathbf{Y} \tag{3.29}$$

où l'on a pour un lisseur composé de polynômes locaux, $\lambda_{i\,(i=1...d)} \equiv (h_i, p_i, W_i)$. Supposons maintenant que l'on détermine *a priori* et $\forall i$ la fonction de pondération W_i employée, que l'on ne s'intéresse qu'aux fonctions polynomiales d'ordre 1 et 2 (i.e. $p_i \in \{1, 2\}$) et qu'enfin, l'on détermine la largeur de la fenêtre de lissage à l'aide de la méthode des k plus proches voisins (i.e. que pour un échantillon de taille n, $h_i \in \{1, 2, ..., n\}$). Supposons de plus que l'on permette que chacun des termes du modèle puisse être soit ignoré ou remplacé par un terme paramétrique linéaire de la forme $\beta_j X_j$, alors un exercice d'analyse combinatoire permet d'obtenir que pour un échantillon de taille n, le nombre total de modèles potentiels est égal à $(2n + 2)^d$. Ainsi, par exemple, pour un échantillon de taille $n = 49$ et $d = 5$ variables explicatives potentielles, il existe environ $100^5 = 10^{10}$ soit 10 milliards de modèles différents. Cet exercice montre bien que dans cette situation, il serait en pratique trop exigeant, en temps de calcul, de tenter de déterminer le modèle optimal en estimant chacun de ces modèles potentiels et en calculant pour chacun d'entre eux le critère retenu.

Puisqu'en pratique il semble irréalisable d'examiner chacun des modèles potentiels, de nombreuses procédures plus ou moins automatiques ont été développées et/ou proposées dont les procédures de type pas-à-pas par en avant, par en arrière ou dans les deux directions (Hastie et Tibshirani, 1990) et les procédures employant de manière adaptative l'algorithme de *backfitting* telles l'algorithme BRUTO (Hastie et Tibshirani, 1990) ou la méthode de sélection automatique des paramètres de lissage de Opsomer (1995). Bien que ces dernières approches soient plus efficaces au niveau du nombre d'opérations à effectuer, leur utilisation pratique est généralement limitée par le manque d'outils informatiques conviviaux permettant leur application. De plus, l'algorithme BRUTO est une nouvelle méthode qui nécessite d'être testée en profondeur (Hastie et Tibshirani, 1990) alors que la méthode de Opsomer (1995) est une méthode de type *plug-in*, un type de méthode amené à disparaître en raison des mauvais résultats généralement obtenus comme l'indiquent Cleveland et Loader (1996b) :

> "Plug-in methods should be considered an idea that failed, and allowed to die a natural death."

Dans ce travail, et ce en raison, entre autres, de la disponibilité d'outils informatiques efficaces et conviviaux dans le logiciel S-Plus, nous adopterons une procédure de type pas-à-pas dans les deux directions. Afin de pouvoir utiliser, en pratique, une procédure de type pas-à-pas, l'on doit d'abord effectuer un certain ordonnancement des paramètres de lissage associés à chacune des variables explicatives. Cet ordonnancement s'effectue généralement selon l'échelle des degrés de liberté en utilisant la relation directe existant entre la notion de paramètre de lissage (via $\mathbf{S}_\lambda$) et la notion de degré de liberté, notée désormais dl, (une fonction de $\mathbf{S}_\lambda$). Ainsi, pour un modèle additif composé de polynômes linéaires, en supposant que l'on permette que chacun des termes du modèle puisse être soit (1) ignoré (i.e. $dl = 0$), (2) remplacé par un terme paramétrique linéaire de la forme $\beta_j X_j$ (i.e. $dl = 1$) ou (3) lissé à l'aide de ses k plus proches voisins (i.e. $dl > 1$ et est une fonction décroissante de k), alors, un ordonnancement naturel des paramètres de lissage potentiels pour la variable prédictive X_j est le vecteur :

$$\begin{aligned} \Lambda_j &= (\lambda_j^1, \lambda_j^2, \ldots, \lambda_j^{i_j}, \ldots, \lambda_j^n, \lambda_j^{n+1}, \lambda_j^{n+2}) \\ &= ((1,1,w),(2,1,w),\ldots,(i_j,1,w),\ldots,(n,1,w),(n,1,\text{rect}),(0,1,w)) \end{aligned} \quad (3.30)$$

où w est une fonction de pondération quelconque et rect est la fonction de pondération rectangulaire. Remarquons que l'emploi de $\lambda_j^{n+1} = (n, 1, \text{rect})$ équivaut à estimer le terme linéaire $\beta_j X_j$ alors que l'emploi de $\lambda_j^{n+2} = (0, 1, w)$ équivaut à ignorer la variable explicative X_j (puisque $k = 0$).

La procédure de type pas-à-pas consiste à estimer, dans un premier temps, le modèle correspondant au vecteur de paramètres de lissage $\Lambda^{(0)} = (\lambda_1^{i_1}, \lambda_2^{i_2}, ..., \lambda_p^{i_p})$. On estime ensuite tous les modèles dont un seul des paramètres de lissage diffère de $\Lambda^{(0)}$ d'une seule position (selon l'ordonnancement établi en 3.30). Il s'agit ainsi des modèles correspondant aux vecteurs $(\lambda_1^{i_1-1}, \lambda_2^{i_2}, ..., \lambda_p^{i_p})$, $(\lambda_1^{i_1+1}, \lambda_2^{i_2}, ..., \lambda_p^{i_p})$, $(\lambda_1^{i_1}, \lambda_2^{i_2-1}, ..., \lambda_p^{i_p})$, $(\lambda_1^{i_1}, \lambda_2^{i_2+1}, ..., \lambda_p^{i_p})$, ... , $(\lambda_1^{i_1}, \lambda_2^{i_2}, ..., \lambda_p^{i_p-1})$, $(\lambda_1^{i_1}, \lambda_2^{i_2}, ..., \lambda_p^{i_p+1})$. On choisit ensuite le meilleur de tous ces modèles que l'on compare au modèle initial. Si le modèle initial est le meilleur, la procédure s'arrête. Sinon, on recommence les étapes précédentes avec le meilleur de ces modèles comme nouveau modèle initial. Notons qu'avec cette procédure, il n'y a pas nécessairement convergence vers la solution optimale ; il peut s'agir d'un maximum local et l'atteinte de cet optimum est alors relié aux conditions initiales. Il est donc recommandé, en pratique, d'effectuer la procédure avec différents paramètres initiaux. Mentionnons aussi qu'avec cette procédure, il n'est pas possible de créer un ordonnancement qui tienne compte du degré des polynômes locaux. Nous recommandons ainsi d'analyser les graphiques obtenus par la modélisation additive afin de vérifier que les polynômes de degré 1 décrivent bien les observations.

3.4.3 L'aspect graphique et la linéarité des modèles de régression

Rappelons qu'en pratique, l'utilisation d'une approche de modélisation non paramétrique, comme la modélisation additive, poursuit généralement deux objectifs principaux, le premier d'ordre descriptif, le second prédictif. Ainsi, l'utilisation d'une approche de modélisation additive permet non seulement (1) d'effectuer une modélisation des données observées afin de produire un lisseur additif pouvant être utilisé pour la prédiction, mais aussi (2) de suggérer l'emploi d'un modèle paramétrique quelconque. Par exemple, un examen visuel du graphique de la figure 3.4a pourrait suggérer l'emploi d'une droite de régression.

Mentionnons que pour la régionalisation des quantiles de crue, le lissage des relations entre quantiles de crues et variables physiographiques / climatologiques devrait permettre, dans un premier temps, en examinant les différents lissages, de détecter les régions à l'intérieur desquelles l'hypothèse de linéarité inhérente au modèle traditionnel de régression log-linéaire est inadéquate. On devrait alors s'attendre à ce qu'en ces régions, une approche de régression non paramétrique comme l'approche de modélisation additive soit préférable et permette de produire, dans un deuxième temps, de meilleures prédictions des quantiles de crue en des sites non jaugés. La figure 3.4 constitue d'ailleurs un exemple du lissage de la relation entre le logarithme décimal du quantile de crue Q_{50} et le logarithme décimal de l'aire (A) des bassins versants pour deux régions hydrologiques des États-Unis. La figure 3.4a suggère que pour la région 1, le

modèle log-linéaire traditionnel peut être adéquat alors que pour la région 12 (fig. 3.4b), le graphique de droite, l'emploi d'une modélisation additive peut permettre l'obtention de meilleures prédictions.

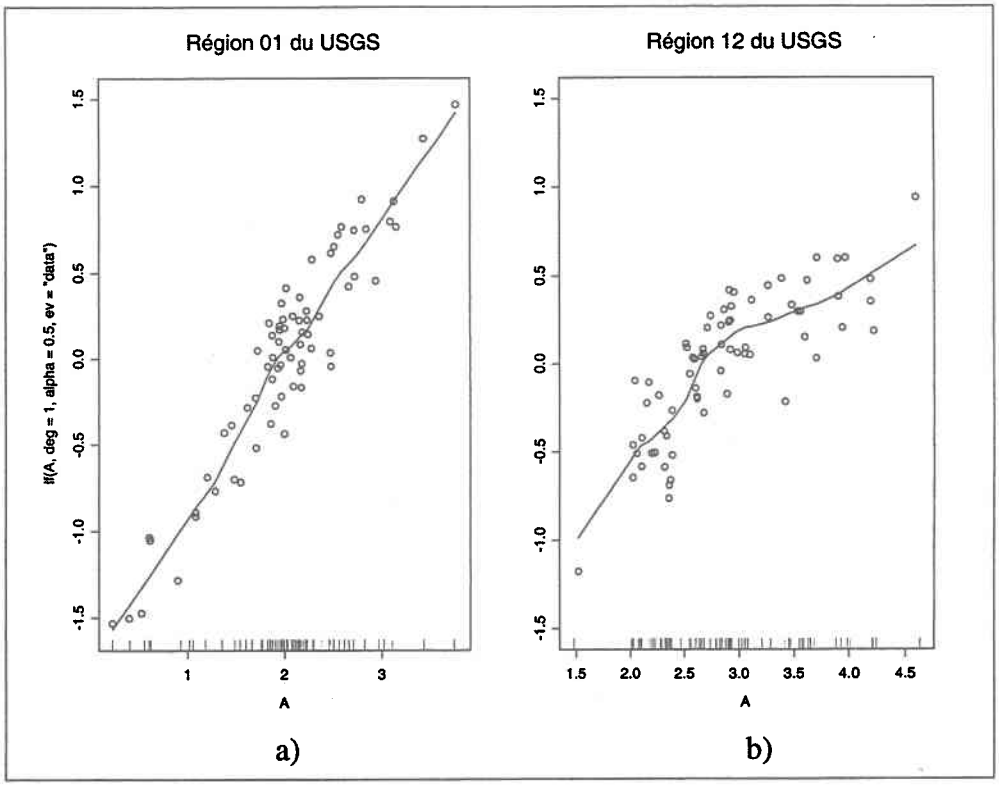

FIG. 3.4: Visualisation de l'hypothèse de linéarité pour 2 régions hydrologiques définies par le USGS (cf. 5.1 pour une description détaillée des régions)

En régression non paramétrique, c'est-à-dire en modélisation locale et en modélisation additive, des tests d'hypothèses peuvent être utilisés afin de valider l'hypothèse qu'un lisseur est significativement supérieur à un autre lisseur. Pour tester l'hypothèse :

$$H_0 : \hat{\mathbf{f}}_{\Lambda_0} = \mathbf{R}_{\Lambda_0}\mathbf{Y}$$

contre l'hypothèse alternative

$$H_1 : \hat{\mathbf{f}}_{\Lambda_1} = \mathbf{R}_{\Lambda_1}\mathbf{Y}$$

il faut d'abord calculer la statistique :

$$F = \frac{(\mathrm{SCE}(\Lambda_0) - \mathrm{SCE}(\Lambda_1))/(\nu^{(0)} - \nu^{(1)})}{\mathrm{SCE}(\Lambda_1)/(n - \nu^{(0)})} \tag{3.31}$$

où n est le nombre d'observations, $\nu^{(0)}, \mathrm{SCE}(\Lambda_0)$ et $\nu^{(1)}, \mathrm{SCE}(\Lambda_1)$ respectivement le nombre de degré de liberté et la somme du carré des erreurs des lisseurs $\hat{\mathbf{f}}_{\Lambda_0}$ et $\hat{\mathbf{f}}_{\Lambda_1}$. On utilise ensuite le fait que F est approximativement distribuée selon une loi $F_{\nu^{(0)}-\nu^{(1)},n-\nu^{(0)}}$ et on rejète H_0 à un niveau de confiance α lorsque $\alpha(F) < \alpha$ (c'est-à-dire lorsque $F > F_\alpha$).

Une façon possible de valider l'hypothèse de linéarité consiste à utiliser le test précédent avec l'hypothèse $H_0 : \hat{\mathbf{f}} = \boldsymbol{\beta}\mathbf{Y}$ pour laquelle on a $\nu^{(0)} = d + 1$ où d est le nombre de variables explicatives du modèle de régression linéaire multiple. L'application de tels tests pour valider les hypothèses de linéarité des lissages de la figure 3.4 ont donné comme résultats, selon le logiciel S-Plus, que $\alpha(F) = 0,289$ pour la région 01 (cf. fig. 3.4a) et $\alpha(F) = 0,017$ pour la région 12 (cf. fig. 3.4b). Ces résultats indiquent, tels que l'on devrait s'attendre, que pour la région 12, l'application d'un lisseur plutôt qu'un modèle linéaire, diminue significativement la variance résiduelle alors qu'aucune diminution significative n'est observée pour la région 01. Le test F approximatif en 3.31 donne généralement de bons résultats pour des lisseurs par polynômes locaux de degré 1 (Hastie et Tibshirani, 1990) (pour d'autres approximations plus précises voir, par exemple, Cleveland et Devlin (1988)).

4. LA RÉGIONALISATION DES QUANTILES DE CRUE

La **régionalisation** hydrologique est un processus par lequel on déduit des caractéristiques en des sites où elles ne sont pas mesurées. En hydrologie de surface, on s'intéresse, par exemple, à la régionalisation des débits de crue, des étiages, des précipitations, etc. Dans le cadre de cette thèse, on s'intéresse plus particulièrement à la régionalisation des quantiles de crue, c'est-à-dire à l'estimation, en un site non jaugé, d'un quantile particulier, le quantile de période de retour T, noté Q_T, de la distribution de fréquence des crues annuelles. Mentionnons que dans ce document, nous ne traitons que de la régionalisation, conçue pour l'estimation en des sites non jaugés, que nous distinguons de l'analyse régionale, conçue pour l'estimation en des sites non jaugés mais aussi en des sites partiellement jaugés contenant peu d'information hydrologique. Nous référons le lecteur à (GREHYS, 1996a, 1996b) pour une description du problème plus général de l'analyse régionale et de l'estimation en des sites partiellement jaugés.

Dans le cas où des mesures de DMA sont disponibles au site où l'on désire produire une estimation, le problème de l'estimation du quantile de crue Q_T consiste à approximer la distribution de fréquence cumulée F, puisque $Q_T = F^{-1}(1-1/T)$, à l'aide de la série $(x_1, x_2, ..., x_n)$ des DMA (où n est le nombre d'années passées pour lesquelles on connaît le débit maximum annuel). Il s'agit alors de modéliser la relation existant entre les quantiles de crue et leur période de retour. Dans la littérature hydrologique, on qualifie généralement de **temporelle** ce type de modélisation aux sites jaugés (locale). Lorsqu'aucune information hydrométrique n'est disponible au site où l'on désire produire une estimation, l'on se doit de s'intéresser davantage à la modélisation **spatiale** (régionale) des caractéristiques de crue.

Le but de ce chapitre est de présenter certains concepts de base liés au problème de la régionalisation des quantiles de crue. Il s'agit d'effectuer une revue de la littérature des approches de modélisation jugées pertinentes pour le travail en cours soient les approches statistiques de modélisation temporelle (locale) et spatiale (régionale) des quantiles de crue. Après avoir décrit, dans la section 4.1, la problématique hydrologique étudiée, nous présentons respectivement, dans les sections 4.2 et 4.3, les principales approches locales et régionales de modélisation.

4.1 Description de la problématique

En pratique, des estimations de quantiles de crue Q_T sont requises, entre autres, pour la gestion des plaines inondables, pour le dimensionnement des structures de protection et de contrôle des crues de même que pour le dimensionnement de toute autre structure soumise à un risque de défaillance (par exemple débordement). Ainsi, aux États-Unis, en matière de gestion des plaines inondables, le *National Flood Insurance Program* oblige l'achat d'une police d'assurance-inondation aux gens habitant les zones dont le débit Q_{100} (débit de période de retour T = 100 ans) est susceptible de causer des dommages. On se doit alors d'obtenir l'estimation de Q_{100} afin de déterminer les zones à risque. Un autre exemple d'utilisation de Q_T est le calcul du débit de conception, un paramètre permettant le dimensionnement des ouvrages construits en rivière. On s'intéressera ici tout particulièrement à cette problématique.

En général, les ouvrages construits en rivière sont dimensionnés de façon à pouvoir résister à une crue que l'on nomme **crue de conception** (*design flood*) ou **crue de projet**. La crue de conception est définie comme étant le débit maximum qu'une structure peut supporter sans qu'il en résulte des dommages appréciables. En hydrologie, il est de pratique courante de dimensionner les ouvrages de manière à ce qu'ils puissent résister à une crue d'une période de retour T donnée, notée Q_T. On peut d'ailleurs montrer que le risque annuel de défaillance p (c'est-à-dire la probabilité au dépassement) d'un ouvrage conçu pour résister à une crue de période de retour T est $p = 1/T$ (Bobée et Rasmussen, 1994). Par exemple, le risque qu'un ouvrage dimensionné pour résister à une crue centennale ($T = 100$) soit détruit au cours d'une année donnée est de une chance sur 100.

À l'étape de la planification des ouvrages, on doit d'abord fixer le niveau de risque que l'on est prêt à accepter. Idéalement, une analyse du risque, une analyse supportant la prise de décision par la quantification des conséquences ainsi que de leur probabilité d'occurrence (NRC, 1988) devrait être effectuée afin de déterminer le niveau de risque à accepter (pour plus de détails sur l'analyse du risque, voir par exemple Duckstein et Parent (1997)). Cependant, en pratique, lorsqu'aucune norme n'impose le niveau de risque, on le détermine généralement a priori en intégrant intuitivement l'ampleur des conséquences dommageables d'une défaillance. Ainsi, on peut accepter qu'un pont desservant une route très secondaire soit submergé une fois tous les 10 ans ; les ponts d'autoroutes sont plutôt conçus pour résister à une crue survenant une fois toutes les 50 ou 100 années. Enfin, lorsque les conséquences peuvent être désastreuses (destruction d'une usine nucléaire ou du barrage d'un grand réservoir), on utilise plutôt la crue millennale ($T = 1\,000$) ou même la crue décamillennale ($T = 10\,000$) afin de concevoir ces ouvrages.

Pour un niveau de risque spécifié, il est important d'obtenir l'estimation la plus précise possible de Q_T en raison d'une part, des coûts de construction injustifiés qu'entraîne une surestimation du débit de conception et d'autre part, de l'augmentation du risque réel de défaillance et par le fait même, d'occurrence de dommages qu'entraîne une sous-estimation de Q_T. En règle générale, bien qu'il soit impliqué dans le processus décisionnel, l'hydrologue n'est pas la personne qui prend les décisions en matière de dimensionnement des ouvrages. L'hydrologue est plutôt le détenteur des informations et connaissances techniques dans les domaines de l'hydrologie et de l'hydraulique (Bernier, 1997). Les décisions sont quant à elles prises à un niveau plus élevé par un gestionnaire (ou décideur) qui demande à l'hydrologue de lui fournir l'information hydrologique nécessaire à son analyse. Généralement, cette information consiste en une estimation de Q_T accompagnée d'un intervalle de confiance, c'est-à-dire d'une mesure de l'incertitude reliée à son estimation. Le gestionnaire a par la suite comme tâche l'évaluation des diverses incertitudes tant économiques, socio-économiques, stratégiques qu'hydrologiques susceptibles d'affecter sa prise de décision, c'est-à-dire le choix du débit de conception qui ultérieurement, servira à la construction de l'ouvrage.

Pour le gestionnaire de projet, le problème du choix du débit de conception est un problème d'analyse de la décision en avenir incertain. Diverses théories tant économiques (théorie de l'utilité (Bowers et al., 1986)) que statistiques (théorie Bayesienne de la décision (Fortin, 1997)) permettent la prise de décision en environnement incertain. Cunnane (1987) a déjà mis en évidence le fait qu'en pratique, en dépit du caractère économique relié au dimensionnement des ouvrages, l'estimation de Q_T est généralement séparée du problème décisionnel de nature plus économique que constitue la conception des ouvrages. De plus, Bernier (1990) a montré l'intérêt d'utiliser une approche permettant une analyse intégrée des problèmes d'estimation hydrologique et des problèmes décisionnels de planification en tenant compte rationnellement des diverses incertitudes intervenant dans ces problèmes. L'objectif de ce travail n'est cependant pas l'étude des diverses théories décisionnelles. Il s'avère toutefois essentiel d'insister sur l'effet des incertitudes sur la prise de décision. En effet, peu importe la théorie retenue, une réduction des diverses incertitudes permet généralement la prise de décisions économiquement plus réalistes. Par exemple, en matière de dimensionnement des ouvrages, une approche classique de prise en compte des incertitudes consiste à surdimensionner les ouvrages en prenant non pas l'estimation de Q_T comme débit de conception mais plutôt la limite supérieure de l'intervalle de confiance à 95% de cette estimation (voir, par exemple, Bernier (1990)). Il est donc clair qu'une diminution de l'incertitude entourant l'estimation de Q_T (donc de l'intervalle de confiance), incertitude quantifiée dans cet exemple par la variabilité échantillonnale de l'estimation, permet de réduire

le surdimensionnement des ouvrages et par le fait même, les coûts de construction de ceux-ci. Dans le cadre de ce travail, nous ne nous intéresserons qu'à la réduction des **incertitudes hydrologiques**, aussi appelées **incertitudes technologiques** (Bernier, 1990) entourant l'estimation de Q_T.

4.1.1 Les incertitudes hydrologiques

Puisque dans le domaine des ressources en eau, une grande confusion existe concernant la différence entre les notions d'incertitude, d'imprécision, d'erreur et d'aléa que l'on regroupe souvent sous le terme incertitude (Abi-Zeid, 1997), il s'avère important de définir ces notions dans ce document. On emploie le terme incertitude pour traduire notre faible degré de connaissance d'un phénomène. L'incertitude peut découler de l'erreur ou de l'imprécision mais n'est synonyme d'aucun de ces termes. L'erreur est une déviation par rapport à une valeur correcte tandis que l'imprécision fait plutôt appel à la notion de "vague" où l'affirmation n'est pas en terme clair (Abi-Zeid, 1997). On mentionne généralement que parmi les nombreuses incertitudes susceptibles d'affecter la prise de décision d'un gestionnaire de la ressource eau, on retrouve les incertitudes naturelles ou aléas provenant des fluctuations temporelles et spatiales des phénomènes hydrométéorologiques. Cependant, notre définition de l'incertitude n'inclut pas cette notion d'aléa. On a choisi de distinguer les aléas des autres sources d'incertitudes en raison de leur irréductibilité. En effet, puisque les aléas hydrologiques reflètent généralement la stochasticité de phénomènes soumis aux forces de la nature (précipitation, écoulement en rivière), l'hydrologue n'a pas de moyen d'intervenir sur ces phénomènes naturels ; il ne peut que les mesurer. Il a cependant comme tâche la modélisation, sous diverses incertitudes hydrologiques, de ces phénomènes aléatoires.

L'incertitude hydrologique comporte trois sources distinctes d'incertitude (Bernier, 1990) :

1. l'**incertitude échantillonnale** provenant de données en nombre insuffisant ou comportant des erreurs de mesure,
2. l'**incertitude d'estimation** provenant de méthodes d'estimation mal adaptées, et
3. l'**incertitude de modélisation** provenant de la difficulté à modéliser des phénomènes complexes.

Mentionnons qu'en ce qui concerne l'estimation de Q_T en un site non jaugé, des incertitudes hydrologiques se retrouvent tant au niveau de la modélisation locale et de l'estimation locale des caractéristiques de crue aux sites jaugés (voir la section 4.2) qu'au niveau de la modélisation spatiale et de l'estimation des relations régionales (voir la section 4.3).

La littérature hydrologique abonde de modèles et de méthodes d'estimation tant locales que régionales permettant l'estimation de Q_T. Cependant, comme l'indiquent d'ailleurs de nombreux auteurs (e.g. Potter (1987), Bobée et al. (1993), Fortin (1994)), peu a été fait au niveau de la comparaison (évaluation) des diverses procédures d'estimation. L'objectif ici n'est pas de proposer une nouvelle procédure de comparaison mais bien de clarifier deux éléments généralement ignorés par les hydrologues mais cependant fondamentaux pour l'évaluation des diverses procédures d'estimation : la distinction entre (1) la forme d'un modèle, responsable de l'incertitude de modélisation et (2) l'estimation de ce dernier, reliée aux incertitudes échantillonnale et d'estimation. En hydrologie

Puisque notre objectif principal dans ce travail est la réduction des diverses incertitudes hydrologiques entourant l'estimation de Q_T en des sites non jaugés, l'on se doit aussi de proposer une mesure de ces incertitudes comme critère de comparaison. Dans la littérature hydrologique, la majorité des études publiées ne s'intéressent généralement qu'aux problèmes de biais d'estimation et de variabilité échantillonnale, c'est-à-dire que l'on néglige l'incertitude de modélisation. Une mesure de comparaison alors couramment utilisée est l'erreur quadratique moyenne (EQM) (ou la racine carrée de celle-ci) :

$$\mathrm{EQM}(\hat{Q}_T) = E\{\epsilon_T^E\}^2 = E\{\hat{Q}_T - Q_T\}^2 = [\mathrm{b}(Q_T)]^2 + \mathrm{var}(\hat{Q}_T) \tag{4.1}$$

où $\epsilon_T^E = \hat{Q}_T - Q_T$ représente l'erreur d'estimation entre la valeur réelle inconnue Q_T et l'estimation $\hat{Q}_T$, $\mathrm{b}(Q_T) = E\{\hat{Q}_T - Q_T\}$ représente le biais d'estimation et $\mathrm{var}(\hat{Q}_T) = E\{\hat{Q}_T - E\{\hat{Q}_T\}\}^2$ représente la variance de l'estimateur. Puisqu'en pratique, la vraie valeur de Q_T est inconnue, une approche généralement employée pour le calcul de l'EQM consiste, par exemple pour une estimation paramétrique aux sites jaugés (voir la section 4.2), à simuler des échantillons de taille prédéterminée n à partir d'une distribution parente quelconque D (on connaît alors Q_T) et à estimer $\hat{Q}_T$ à l'aide d'une méthode particulière d'estimation M et de la loi parente D utilisée pour les simulations. En procédant ainsi, on se trouve à négliger le biais (ou l'erreur, i.e. l'incertitude) de modélisation et à ainsi ne s'intéresser qu'aux erreurs (incertitudes) d'estimation (incluant l'incertitude échantillonnale). Par analogie aux mesures de précision présentées à la section 2.3.3, l'EQM calculée ainsi ne constitue qu'une mesure descriptive permettant de comparer différentes procédures d'estimation. Cette façon de faire s'est d'ailleurs attirée de nombreuses critiques en hydrologie (e.g. Landwehr et al. (1987), Bobée et al. (1993), Fortin (1994)) et ce notamment puisque l'on fait alors l'hypothèse non vérifiable et donc difficilement justifiable que les débits de crues appartiennent à une distribution parente connue D. Dans ce document, nous emploierons la notion de **précision de l'estimation** pour faire référence à toute mesure d'incertitude n'incluant pas l'incertitude de modélisation.

En pratique, il est généralement plus opportun d'avoir en notre possession une mesure de l'incertitude hydrologique, c'est-à-dire une mesure qui n'inclut pas seulement les incertitudes échantillonnale et d'estimation mais aussi l'incertitude de modélisation. Une revue sommaire de la littérature hydrologique indique que Weber et al. (1973) ont été les premiers à discuter de l'importance de tenir compte de l'incertitude de modélisation lors de l'évaluation des modèles hydrologiques. Ces auteurs mentionnent ainsi qu'il est extrêmement important de se rappeler que toute prédiction à partir d'un modèle est conditionnelle à la forme du modèle (responsable de l'incertitude de modélisation) ainsi qu'à la méthode d'estimation des paramètres (reliée aux incertitudes échantillonnale et d'estimation) de ce modèle.

Weber et al. (1973) indiquent qu'un modèle est sujet à deux sources distinctes d'erreur : une erreur systématique ou biais et une erreur aléatoire ou variance. Ils utilisent le terme *accuracy* (exactitude) pour faire référence à la qualité qu'un modèle a de produire des estimations non biaisées (sans biais de modélisation) alors que le terme *precision* (précision de l'estimation) fait plutôt référence à la qualité de produire ces estimations avec une faible variabilité. Un point important à prendre alors en considération est qu'il est possible d'améliorer la précision (de l'estimation) d'un modèle en augmentant la taille de l'échantillon servant à sa calibration ou en utilisant des méthodes d'estimation mieux adaptées alors que l'exactitude de ce modèle ne peut être améliorée qu'en changeant la forme de celui-ci. Il est donc important et même nécessaire lors de l'évaluation de modèles de tenir compte de ces deux sources d'erreur (aléatoire et systématique) puisque des méthodes d'estimation différentes feront généralement varier la précision de l'estimation du modèle (partie aléatoire) alors que différentes formes de modèles feront plutôt varier l'exactitude de celui-ci (partie systématique).

Par analogie aux mesures de précision présentées à la section 2.3.3, nous proposons l'emploi, comme mesure de l'incertitude hydrologique associée à l'estimation de Q_T, de l'erreur quadratique de prédiction (EQP) :

$$\mathrm{EQP}(\hat{Q}_T) = E\{\epsilon_T^P\}^2 = E\{\hat{Q}_T - Q_T\}^2 \tag{4.2}$$

où $\epsilon_T^P = \hat{Q}_T - Q_T$ représente l'erreur de prédiction. Mentionnons que la différence entre l'EQP, définie en 4.2, et L'EQM, définie en 4.1, est que l'espérance de l'erreur en 4.1 est calculée pour des erreurs obtenues lors de l'estimation de Q_T sous l'hypothèse que l'on connaît exactement le modèle (sa forme) alors que l'espérance de l'erreur en 4.2 est plutôt calculée pour des erreurs obtenues sous l'hypothèse que le modèle est inconnu ou inexact. En hydrologie, une telle mesure de l'EQP peut être obtenue, par exemple, en simulant des échantillons à l'aide d'une loi parente générale D mais en estimant plutôt à l'aide de diverses combinaisons possibles de lois parentes

D' (différentes de D) et de méthodes d'estimation M. Dans la littérature hydrologique, on qualifie alors de **robustes** (Kuczera, 1982) les procédures d'estimation possédant les plus faibles EQP. La robustesse constitue présentement une propriété qui reçoit une attention bien méritée de la part des hydrologues pour la comparaison de diverses procédures d'estimation paramétriques. Dans ce document, nous privilégierons l'emploi du terme **précision prédictive**, plutôt que robustesse, pour faire référence à cette mesure de l'incertitude hydrologique (i.e. incluant l'incertitude de modélisation).

4.2 L'analyse locale de la distribution des crues annuelles

4.2.1 Les objectifs

L'objectif premier de l'analyse locale (au site jaugé) de la distribution des crues annuelles est d'obtenir une approximation la plus précise possible de la distribution de fréquence réelle cumulée F des crues annuelles, à l'aide de la série $(x_1, x_2, ..., x_n)$ des DMA disponible au site jaugé où l'on désire produire une estimation ; n est le nombre de valeurs du débit maximum annuel dans l'échantillon. De manière générale, cette estimation a pour objectif l'estimation, au site jaugé, des quantiles de crue Q_T ou, dans un contexte d'analyse régionale (voir la section 4.3), de paramètres de cette distribution.

4.2.2 Les hypothèses

Avec l'approche de modélisation des crues annuelles, on ne considère qu'une seule crue par année correspondant à la valeur maximale du débit atteinte au cours de cette année. On fait ensuite l'hypothèse que le débit maximum annuel d'une année quelconque est une variable aléatoire X issue d'une population P ayant une fonction de densité de probabilité cumulée stationnaire F telle que :

$$F(x) = Pr[X \leq x] \tag{4.3}$$

Selon cette approche, l'amplitude Q_T du débit de la crue ayant une probabilité au dépassement préalablement définie $p = 1/T$ d'être dépassée au cours d'une année donnée correspond au $(1 - p)$ième quantile de la distribution des débits maximums annuels, soit :

$$Q_T = F^{-1}(1 - p) = F^{-1}(1 - 1/T) \tag{4.4}$$

Après avoir extrait d'une série de débits journaliers mesurés les débits maximums pour chaque années, il est nécessaire de vérifier que la série $(x_1, x_2, ..., x_n)$ respecte certaines hypothèses de

base afin que l'analyse de fréquence locale soit théoriquement acceptable. On doit alors s'assurer que (Bobée et Ashkar, 1991; Roy, 1993) :

1. **le débit est naturel**, c'est-à-dire que les fluctuations des observations sont naturelles et non induites artificiellement (réservoir contrôlé, rivière régularisée, etc).
2. **les observations de l'échantillon (ordonnées chronologiquement) sont indépendantes.** Généralement, les valeurs maximums annuels de débits ne sont pas autocorrélées mais si, par exemple, une crue (au sens large du terme) se produisait à la fin d'une année donnée et se poursuivait jusqu'au début de l'année suivante, l'hypothèse d'indépendance ne serait pas respectée.
3. **l'échantillon est homogène**, c'est-à-dire que la population P des crues annuelles n'est composée que d'éléments homogènes soit par exemple, des crues printanières causées par la fonte de neige, des crues causées par des ouragans, des crues causées par des précipitations extrêmes, etc.
4. **la série est stationnaire**, c'est-à-dire que les probabilités de dépassement annuelles ne varient pas avec les années. On rejette, par exemple, l'hypothèse qu'il peut y avoir des cycles hydro-météorologiques interannuels (changements climatiques tels le réchauffement de la planète, phénomènes du type El-Niño, etc.).
5. **la série est dépourvue de valeurs singulières.** Il arrive parfois que des erreurs grossières de mesure ou de transcription se produisent. Il faut alors retrancher ces valeurs de l'échantillon.

4.2.3 L'estimation de la distribution des crues annuelles

Lorsque l'on dispose d'un échantillon d'observations de DMA homogènes et indépendantes $(x_1, x_2, ..., x_n)$, deux approches permettant l'estimation de F sont principalement utilisées par les hydrologues : l'approche non paramétrique d'estimation à l'aide d'une formule de probabilité empirique (FPE) ou d'une technique de lissage et l'approche paramétrique classique consistant à ajuster une distribution paramétrique statistique $F(x; \underline{\Theta})$ où $\underline{\Theta}$ représente le vecteur de paramètres de cette distribution.

4.2.3.1 L'approche non paramétrique

Plusieurs méthodes d'estimation non paramétrique de la distribution F sont basées sur l'ordre qu'occupe chaque observation dans l'échantillon classé des débits, noté $(x_{(1)}, x_{(2)}, ..., x_{(n)})$. on peut obtenir une approximation de $F(x_{(i)})$ à l'aide d'une fonction de i et de n appelée **formule**

de probabilité empirique (FPE) (*plotting position*) dont la forme générale est donnée par :

$$F(x_{(i)}) = \frac{i-a}{n+1-2a} \tag{4.5}$$

où a est une constante. Le tableau 4.1 présente les FPE les plus employées en hydrologie.

TAB. 4.1: Formules de probabilité empirique (tiré de Bobée et Ashkar (1991))

Nom de la formule	Équation	Référence
Hazen ($a = 0,5$)	$F(x_{(i)}) = \frac{i-0,5}{n}$	Hazen (1914a)
Weibull ($a = 0$)	$F(x_{(i)}) = \frac{i}{n+1}$	Weibull (1939)
Chegodayev ($a = 0,3$)	$F(x_{(i)}) = \frac{i-0,3}{n+0.4}$	Chow (1964)
Cunnane ($a = 0,4$)	$F(x_{(i)}) = \frac{i-0,4}{n+0,2}$	Cunnane (1978)

Pour l'estimation en des points non observés, on procède généralement par interpolation linéaire. L'estimation par une FPE se limite alors aux cas où $1 - 1/T$ est inférieur à $F(x_{(n)})$. En pratique, l'estimation par FPE n'est généralement utilisée, comme l'indique d'ailleurs son appellation anglaise, que pour produire un graphique de la relation entre les valeurs de crue observées et leur probabilité de dépassement (ou période de retour) associée. Ce graphique permet éventuellement de vérifier l'adéquation de fonctions F paramétriques aux fréquences observées (selon la FPE).

Une autre approche non paramétrique apparue dans la littérature hydrologique au milieu des années quatre-vingt pour l'estimation de la distribution des crues annuelles est l'approche par lissage. Que ce soit pour l'estimation d'une fonction de régression, d'une fonction de densité ou d'une fonction de répartition, l'approche de modélisation non paramétrique par lissage poursuit toujours le même objectif soit celui d'estimer, avec une erreur qui diminue lorsque la taille des échantillons augmente, une fonction inconnue et arbitraire des données ; chaque estimation étant locale, c'est-à-dire influencée seulement par les données environnantes (Lall, 1995). La philosophie est aussi la même : l'information se retrouve dans les données et il faut les laisser nous indiquer la forme de la vraie fonction de densité (ou de répartition) des crues. De plus, les différentes techniques de lissage présentées au chapitre 2 peuvent être appliquées, presque intégralement, pour l'estimation de fonctions de densité ou de fonctions de répartition.

L'approche de modélisation non paramétrique par lissage comporte deux étapes : (1) le choix d'une méthode particulière de lissage, appelé un lisseur et (2) le choix d'une méthode d'ajustement du niveau de lissage (e.g. le choix du type de noyau, de la taille du voisinage, etc.). Dans une revue des principales applications hydrologiques des méthodes d'estimation non paramétrique, Lall (1995) mentionne que Yakowitz (1985) et Adamowski (1985) ont été les premiers, au congrès de l'AGU de l'automne 1983, à introduire de manière indépendante l'estimation par noyau pour l'estimation de distributions de crues. En hydrologie, on a utilisé différents types de noyau : noyau fixe de type rectangulaire et de Cauchy (Adamowski, 1985), noyau fixe de Gumbel (Bardsley, 1989), noyau variable (Adamowski, 1989). On a aussi utilisé différentes méthodes pour choisir la fenêtre optimale : méthode d'Adamowski (Adamowski, 1985), validation croisée (Adamowski et Feluch, 1991; Gingras et al., 1995), validation croisée par maximum de vraisemblance (Adamowski, 1989), validation croisée par moindres carrés (Adamowski, 1996). De plus, on a lissé tantôt la fonction de densité (Adamowski, 1989), tantôt la fonction de probabilité empirique, aussi appelée fonction de quantile (Moon et Lall, 1994; Wu et Woo, 1989). Une des seules exceptions à l'utilisation du lissage par noyau revient à Wu et Woo (1989) qui ont proposé l'utilisation des séries de Fourier, une méthode de la famille des séries orthogonales, pour le lissage de la fonction de probabilité empirique.

4.2.3.2 L'approche paramétrique

L'approche paramétrique d'estimation de F comporte aussi deux étapes. La première consiste à choisir une distribution D pour représenter les débits maximums annuels. Chaque distribution paramétrique est caractérisée par une fonction de densité de probabilité cumulée différente $F(x;\underline{\Theta})$, où $\underline{\Theta}$ représente le vecteur des paramètres de la distribution D. En utilisant la relation en (4.4), on obtient alors comme expression du quantile de crue :

$$Q_T = F^{-1}(1-p\,;\underline{\Theta}) = F^{-1}(1-1/T\,;\underline{\Theta}) \tag{4.6}$$

La deuxième étape consiste à obtenir l'estimateur $\underline{\hat{\Theta}}$ du vecteur de paramètres $\underline{\Theta}$, à partir de l'échantillon $(x_1, x_2, ..., x_n)$ des DMA, en utilisant une méthode d'ajustement M. Puis, en remplaçant le vecteur de paramètres $\underline{\Theta}$ par son estimateur $\underline{\hat{\Theta}}$ dans (4.6), on obtient l'expression de l'estimateur local paramétrique du quantile de crue :

$$\hat{Q}_T = F^{-1}(1-p\,;\underline{\hat{\Theta}}) = F^{-1}(1-1/T\,;\underline{\hat{\Theta}}) \tag{4.7}$$

En hydrologie statistique, plusieurs distributions combinées à diverses méthodes d'estimation des paramètres ont été utilisées pour approximer la distribution des crues annuelles ou de manière plus générale, la distribution d'événements hydrologiques extrêmes. De nombreuses revues de ces diverses distributions et méthodes d'estimation ont d'ailleurs déjà été effectuées (e.g. Greis (1983), Potter (1987), Cunnane (1987), Bobée et Ashkar (1991), Stedinger et al. (1993), Bobée et Rasmussen (1994, 1995)). Soulignons qu'en hydrologie, la loi normale semble avoir été la première loi à être employée par Horton (1913). Par la suite, réalisant que les séries de crues annuelles étaient asymétriques, Hazen (1914b) a montré que la loi log-normale ajustait mieux les séries de crues annuelles. Puis, dans une perspective historique, la loi de Gumbel (aussi appelée loi de valeur extrême de type 1 (EV1)) a probablement été, par la suite, la loi la plus utilisée pour décrire les données de crues (Bobée et Rasmussen, 1994). Au début des années quarante, réalisant que la rapidité de la décroissance de la fonction de densité de probabilité des lois usuelles, pour les très grandes valeurs de la variable, s'avérait parfois en désaccord avec les observations empiriques, Halphen (1941) a proposé de nouvelles lois de probabilité à 3 paramètres (lois de type A et B) et Morlat (1956) en a présenté une extension pour obtenir la famille des lois de Halphen (types A, B et B^{-1}). Toutefois, en raison de la complexité de la forme analytique de leur fonction de densité de probabilité, l'ajustement des lois de Halphen nécessitait de laborieux calculs ce qui, à cette époque, a fait en sorte qu'elles n'ont pas retenu l'attention des praticiens. De nos jours, les lois à deux paramètres (e.g. log-normale, Gumbel, etc.) sont généralement remplacées par d'autres distributions plus flexibles telles la loi généralisée des valeurs extrêmes (GEV) (Jenkinson, 1955) ou la loi log-Pearson à 3 paramètres (LP3) (Bobée, 1975). L'application de la loi GEV est d'ailleurs recommandée au Royaume-Uni (NERC, 1975) alors qu'aux États-Unis (USWRC, 1981) et en Australie (IEA, 1987), la loi LP3 proposée dès 1968 (Benson, 1968) est imposée.

Pour l'estimation des distributions de crues annuelles, de nombreuses méthodes d'estimation ont été employées. Notons qu'une problématique particulière à l'estimation de la distribution des crues annuelles est la faible taille des échantillons disponibles (généralement entre 20 et 70 observations). Durant les années soixante, l'estimation s'effectuait principalement à l'aide de la méthode des moments (MM) ou en ajustant, graphiquement ou par moindres carrés, la FPE à la distribution $F(x;\Theta)$. La méthode du maximum de vraisemblance (MMV) était aussi utilisée occasionnellement pour la loi log-normale à 2 paramètres (LN2) et la loi de Gumbel. Durant les années soixante-dix, la MMV est devenue plus populaire en raison notamment, de son efficacité asymptotique, bien que l'on ait déjà observé des faiblesses en présence d'échantillons de faible taille ou lorsque la loi retenue n'ajustait pas bien les observations (Fill, 1994). Diverses mé-

thodes s'apparentant à la méthode classique des moments (i.e. la méthode des moments mixtes, la méthode des moments généralisés, etc.) ont aussi été développées pour l'estimation de lois hydrologiques (voir par exemple le livre de Bobée et Ashkar (1991)). Depuis le début des années quatre-vingt, une méthode a plus particulièrement retenu l'attention des hydrologues : la méthode des L-moments (Hosking, 1990), présentée à l'annexe A, qui se dérive de la théorie des moments pondérés par probabilités (Greenwood et al., 1979; Hosking, 1986). La principale caractéristique de l'estimation à l'aide de L-moments est que l'on accorde une moins grande importance aux grandes et aux petites observations lors de l'estimation des L-moments et des ratios de L-moments (l'équivalent des moments et des ratios de moments). Ceci se traduit par des estimations moins sensibles à la présence de valeurs singulières dans les échantillons et par le fait même, à des estimations généralement d'une plus grande robustesse en présence d'échantillons de petite taille. Toutefois, selon Bernier (1993), pour l'estimation de quantiles élevés, cette approche d'estimation peut être trop robuste dans la mesure où l'on accorde peu d'importance à des informations importantes provenant de la queue de la distribution parente. Enfin, la MMV a été ressuscitée récemment pour l'ajustement des paramètres de la loi de Halphen et ce, en raison des propriétés statistiques intéressantes de la MMV pour l'estimation des paramètres d'une loi de la famille exponentielle pour laquelle il existe des statistiques exhaustives (pour plus de détails, voir Perreault et al. (1999a, 1999b)).

4.2.4 Le choix d'une procédure d'estimation locale

Puisque la crue d'une rivière est la résultante d'une multitude de facteurs et de phénomènes physiques complexes (conditions météo, caractéristiques physiques du bassin versant, contribution des eaux souterraines, ...), il est en pratique impossible de déterminer théoriquement la distribution statistique réelle du débit maximum annuel et même si tel était le cas, elle serait composée d'un trop grand nombre de paramètres pour avoir une quelconque utilité en pratique (Stedinger et al., 1993). L'objectif premier de l'analyse de la distribution des crues annuelles est donc d'obtenir la meilleure distribution approximative possible de la distribution réelle mais inconnue F des débits maximums annuels et ce, plus particulièrement dans le domaine d'intérêt de la variable aléatoire X (i.e. près de $F(Q_T)$ pour l'estimation de Q_T).

En hydrologie, Cunnane (1987) a été un des premiers auteurs à élaborer sur les critères à prendre en considération lors de la sélection d'une procédure d'estimation (paramétrique), c'est-à-dire d'une combinaison D/M (distribution / méthode d'estimation). Cunnane (1987) a ainsi introduit les notions de capacité **descriptive** et **prédictive** d'une distribution. Selon Cunnane (1987), un modèle possède une bonne capacité descriptive s'il permet de préserver les caractéristiques sta-

tistiques des données de crues observées alors que les capacités prédictives font plutôt référence aux propriétés statistiques des estimateurs des quantiles et ce, en ignorant temporairement les données observées. Remarquons que l'emploi par Cunnane (1987) du terme prédictif est inconsistant avec la notion de précision prédictive définie précédemment. En effet, au sens où nous l'entendons dans ce document, la notion de capacité prédictive constitue une mesure de la précision de l'estimation et non de la précision prédictive. Par ailleurs, selon notre terminologie, la notion de capacité descriptive constitue une mesure qualitative permettant de soupçonner la présence d'un biais de modélisation.

Dans la littérature hydrologique, une des propriétés statistiques des quantiles les plus couramment utilisées comme mesure de la capacité prédictive (au sens de Cunnane (1987)) d'une procédure d'estimation est l'erreur quadratique moyenne (ou la racine carrée de celle-ci), définie à l'équation 4.1. Rappelons que puisqu'en pratique, la vraie valeur de Q_T est inconnue, une approche généralement employée pour le calcul de l'EQM consiste à simuler des échantillons de taille prédéterminée n à partir d'une distribution parente quelconque D (on connaît alors Q_T) et à estimer $\hat{Q}_T$ à l'aide d'une méthode particulière d'estimation M et de la loi parente D utilisée pour les simulations. Cette approche a été critiquée en raison notamment de la non prise en compte de l'incertitude de modélisation. Rappelons enfin qu'une véritable mesure de la précision prédictive (ou robustesse) des procédures d'estimation peut être obtenue en simulant des échantillons à l'aide d'une loi parente générale D mais en estimant plutôt à l'aide de diverses combinaisons possibles D'/M. Mentionnons, de plus, que rien n'empêche d'utiliser cette approche de simulation proposée pour évaluer la précision prédictive des procédures d'estimation tant paramétriques que non paramétriques.

Pour l'estimation locale des quantiles de crue, la forme du modèle est représentée par la fonction de répartition inverse $F^{-1}(1 - 1/T)$ (cf. éq. 4.4). La précision de l'estimation d'un modèle ne peut être améliorée qu'en augmentant la taille de l'échantillon servant à sa calibration ou en utilisant des méthodes d'estimation mieux adaptées (cf. 4.1.1). L'exactitude d'un modèle ne peut quant à elle être améliorée qu'en changeant la forme de ce dernier (cf. 4.1.1). Examinons maintenant l'évolution des différentes procédures employées pour l'estimation locale des quantiles de crue en tenant compte de ces notions d'exactitude et de précision de l'estimation.

D'abord, en ce qui concerne le choix de la forme (paramétrique ou non) du modèle, on peut remarquer que pour l'obtention de meilleures qualités descriptives (c'est-à-dire reproduire le mieux possible les caractéristiques statistiques (coefficient de variation (CV), coefficient d'asymétrie (CS), etc.) des échantillons observés), on a eu recours à des lois possédant de plus en

plus de paramètres, le cas limite étant l'application du lissage où l'adéquation avec les données observées peut être presque parfaite. D'ailleurs, une motivation mentionnée pour l'emploi d'une procédure non paramétrique est que ce type d'estimation permet l'estimation (puisque fortement descriptive) de fonctions de densité multimodales auxquelles on fait même un lien avec le nombre de mécanismes différents pouvant causer les crues (Gingras et Adamowski, 1993; Adamowski et al., 1994; Moon et Lall, 1994).

De manière générale, on remarque que le prix à payer pour l'utilisation de lois avec de plus en plus de paramètres est cependant une diminution des qualités prédictives des modèles. En effet, Bobée et Rasmussen (1994) rapporte que des résultats de Kuczera (1982) et d'autres chercheurs sur la robustesse des estimateurs de quantiles indiquent que généralement, de meilleurs résultats sont obtenus, pour des tailles d'échantillons représentatifs de la réalité, avec des lois à 2 paramètres plutôt qu'à 3 paramètres. De plus, même pour des lois à valeurs de CS fixées (comme la loi EV1) pour lesquelles des estimations fortement biaisées peuvent être produites, la diminution au niveau de la variance en raison d'une parcimonie de paramètres, plus que contrebalance l'effet du biais et produit habituellement des estimations plus robustes.

En ce qui concerne le choix d'une méthode particulière d'estimation, il semble que la méthode des L-moments procure les meilleurs résultats au niveau de la précision de l'estimation. On devrait aussi s'attendre à ce que cette méthode soit aussi efficace, en raison de la taille des échantillons généralement disponible, au niveau de la précision prédictive sauf peut-être, comme l'indique Bernier (1993), dans le cas de l'estimation de quantiles élevés où un biais de modélisation risque d'être introduit par la non prise en compte d'informations importantes dans les queues.

En conclusion, mentionnons que l'analyse locale de la distribution des crues annuelles (ADC) a fait l'objet de nombreuses publications au cours des dernières décennies et demeure un sujet d'intérêt et de grande importance pour les chercheurs. Lors du dernier Rapport National présenté à la International Union of Geodesy and Geophysics (IUGG) traitant des développements récents (1991-1994) de l'ADC, Bobée et Rasmussen (1995) ont fait remarquer que la recherche sur l'ADC a varié en intensité au cours des dernières décennies. Ils mentionnent en effet que durant les années soixante-dix et quatre-vingts, les efforts ont été concentrés sur le développement de nouvelles procédures efficaces d'estimation aux sites jaugés. Puis, les chercheurs réalisant de plus en plus que le manque d'information disponible aux sites jaugés (i.e. une grande incertitude échantillonnale causée par des données en nombre insuffisant) limite le degré de sophistication que peuvent atteindre de façon justifiée ces méthodes d'estimation, on s'intéresse maintenant da-

vantage aux méthodes permettant l'intégration d'information supplémentaire. Il est par exemple possible d'utiliser l'information sur des crues survenues avant la période de jaugeage (information historique). Cette information peut provenir des souvenirs des résidents de longue date, de documents ou de journaux publiés avant la période de jaugeage (Tasker et Stedinger, 1987). Des techniques de paléohydrologie permettent aussi d'obtenir de l'information supplémentaire en étudiant le mouvement de l'eau à travers les dépôts sédimentaires (Stedinger et Cohn, 1986). Dans ce travail, nous nous intéressons plutôt à l'intégration de l'information régionale à l'aide de méthodes d'analyse régionale de la distribution des crues.

4.3 L'analyse régionale de la distribution des crues annuelles

4.3.1 Les objectifs

L'analyse régionale de la distribution des crues annuelles (ARDC) a été développée (1) pour permettre l'estimation de Q_T, le débit ayant une période de retour de T années, en des sites non jaugés et (2) pour améliorer la précision des estimations en des sites contenant peu de données. En ce qui nous concerne, nous ne nous intéresserons qu'au premier objectif, c'est-à-dire à l'estimation de Q_T en des sites non jaugés. Ce type d'estimation (site non jaugé) est d'ailleurs, et de loin, le cas le plus rencontré en pratique (Linsley, 1986).

L'ARDC consiste à utiliser des procédures de régionalisation afin de transférer l'information disponible en des sites jaugés vers le site non jaugé, appelé le **site cible**, où l'on désire produire une estimation. De manière générale, une procédure de régionalisation comporte deux étapes distinctes qui consistent à (1) choisir les sites jaugés à partir desquels s'effectuera le transfert d'information et (2) appliquer aux sites choisis un modèle de transfert d'information régionale, aussi appelé le **modèle régional**. Généralement, un modèle régional est une équation ou un ensemble d'équations qui relient entre elles une ou plusieurs caractéristiques de crue d'une région.

L'ARDC consiste à modéliser la variabilité des crues dans l'espace plutôt que dans le temps. La variabilité entre les caractéristiques de crues calculées aux sites jaugés est composée :

1. de **variabilité spatiale** due aux différences entre les caractéristiques (physiographiques et météorologiques) des bassins versants,
2. de **variabilité temporelle** due à l'échantillonnage aux sites jaugés, et
3. d'une part inconnue de **variabilité due aux erreurs de modélisation**.

Selon Riggs (1990), une procédure de régionalisation devrait expliquer, le plus précisément possible (i.e. avec la plus petite erreur de modélisation possible), les variations dues aux caractéristiques des bassins versants tout en "moyennant" les variations dues à l'échantillonnage aux sites jaugés. Ainsi, en régionalisation, l'intérêt devrait être accordé principalement au modèle régional (relié à 1.) plutôt qu'au modèle local $Q_T = F^{-1}(1 - 1/T)$ (relié à 2.) employé.

4.3.2 Le choix des sites

Avant de pouvoir transférer au site cible l'information régionale provenant de sites jaugés, il est primordial de bien choisir les sites jaugés à partir desquels s'effectuera le transfert d'information régionale. À cette étape, l'hydrologue doit répondre à la question suivante : de quels sites provient l'information permettant d'estimer de la meilleure façon possible le modèle de transfert applicable au site non jaugé ? L'étape du choix des sites comprend généralement deux sous-étapes. La première consiste à regrouper différentes stations afin de former des régions hydrologiques. Il faut ensuite assigner la station non jaugée à une de ces régions ainsi formées.

L'étape du regroupement de sites en régions consiste à définir et déterminer (1) des régions dont les stations ont un comportement hydrologique similaire ou (2) des régions homogènes, c'est-à-dire des régions dont les stations ont une de leurs caractéristiques de crue, généralement le CV, constante à l'échelle régionale. Rappelons que dans la littérature hydrologique, les notions de similarité hydrologique et d'homogénéité régionale sont généralement confondues (cf. 1.1.2 pour une distinction détaillée). Dans ce document, nous emploierons la terminologie **région hydrologique** lorsqu'il ne sera pas nécessaire de distinguer entre homogénéité régionale et similarité hydrologique.

Un projet de recherche a été mené par une équipe de scientifiques canadiens (GREHYS, 1996a, 1996b) afin de comparer les principaux modèles d'estimation régionale des crues utilisés en Amérique du Nord. Les résultats de l'intercomparaison indiquent que, pour l'étape du choix des sites, la méthode de la région d'influence (Burn, 1990a; Zrinji et Burn, 1994) et la méthode d'analyse des corrélations canoniques (Cavadias, 1989, 1990; Ribeiro-Corréa et al., 1995; Ouarda et al., 1997) se distinguent des autres. Nous présentons un aperçu, inspiré de Ouarda et al. (1999), des principales méthodes de détermination de régions hydrologiques employées et décrivons brièvement la méthode des régions d'influence et de l'analyse canonique des corrélations. Pour une revue plus détaillée de l'évolution des procédures de détermination de régions hydrologiques, voir par exemple la section 1.3.1 de Roy (1993).

4.3.2.1 Les méthodes de détermination de régions hydrologiques

On distingue deux types d'approche pour la détermination de régions hydrologiques :

1. **approche basée sur des régions fixes** : ensemble de stations formant une même région hydrologique. Deux sous-groupes peuvent être considérés :
   - *Régions géographiquement contiguës.* Ces régions géographiques peuvent être définies a priori par des délimitations territoriales (États, provinces, ...) ou administratives (régions du USGS (cf. 5.1)). Ces régions peuvent aussi être définies, par exemple, à partir d'une analyse du signe des résidus de modèles de régression (Jennings et al., 1994) ou à partir de la similarité des densités non-paramétriques des débits de crue (Gingras et Adamowski, 1993).
   - *Régions non-contiguës.* Ces régions sont généralement définies en fonction de la similarité de leurs caractéristiques physiques et/ou hydrologiques. Ces régions peuvent être déterminées, par exemple, par une analyse discriminante (DeCoursey, 1973), une analyse factorielle des correspondances (White, 1975), une analyse de regroupement (*cluster analysis*) (Mosley, 1981; Tasker, 1982), etc.
2. **approche basée sur des régions du type voisinage** : où on associe à chaque station cible son propre voisinage. La méthode de la région d'influence (Burn, 1990a; Zrinji et Burn, 1994) définit la proximité de deux sites par une distance dans un espace multidimensionnel dont les axes sont des caractéristiques hydrologiques, physiographiques et météorologiques. Les voisinages peuvent aussi être définis par une analyse des corrélations canoniques (Cavadias, 1989, 1990).

4.3.2.2 L'assignation de la station cible aux régions hydrologiques

Lorsque les régions sont déterminées selon des critères géographiques ou selon une approche de voisinage, le problème de l'assignation de la station cible à une région particulière ne se pose pas. Cependant, lorsqu'il s'agit de régions dont les stations sont regroupées sur la base de leurs caractéristiques physiques et/ou hydrologiques, il est généralement nécessaire de définir une approche d'assignation. Remarquons d'ailleurs que pour l'assignation d'une station non jaugée où aucune information hydrologique n'est disponible, un problème supplémentaire survient : il n'est pas possible de déterminer les caractéristiques hydrologiques de la station non jaugée. Afin de tenir compte de cette problématique, les méthodes de détermination de régions hydrologiques conçoivent les similarités dans l'espace des variables physiographiques/météorologiques et non hydrologiques. On peut aussi avoir recours à une analyse discriminante (Tasker, 1982)

afin d'établir la probabilité d'appartenance d'une station non jaugée à une région préalablement définie. Il est alors possible d'assigner la station à la région dont la probabilité d'appartenance est maximale. Une approche alternative, appelé concept d'appartenance fractionnaire (Tasker, 1982; Wiltshire, 1986b), consiste à appliquer le modèle régional en chacune des régions et à estimer le quantile de crue au site cible à l'aide d'une moyenne pondérée (par les probabilités d'appartenance) des quantiles calculés en chacune des régions hydrologiques.

4.3.2.3 La méthode de la région d'influence

Selon la méthode de la région d'influence proposée d'abord par Burn (1990a) pour des sites cibles jaugés puis par Zrinji et Burn (1994) pour des sites non jaugés, chaque site cible est considéré comme le centre de sa propre région. Le critère permettant alors la sélection de la région d'influence est la distance euclidienne

$$d_{ij} = \sqrt{\sum_{l=1}^{n} w_l (X_l^{(i)} - X_l^{(j)})^2} \tag{4.8}$$

dans l'espace des variables explicatives X_l où w_l est une fonction de pondération permettant la standardisation des variables explicatives. La région d'influence au site i est alors définie par l'ensemble des sites j tels que d_{ij} est inférieure à un certain point de coupure θ_L ou, lorsque ce point de coupure est trop restrictif, par l'ensemble des K sites ayant les K plus petites distances d_{ij} où K est le nombre minimum de stations désiré.

Burn (1990a) propose aussi l'utilisation d'une fonction de pondération (fonction noyau) afin de refléter la proximité relative des différents sites voisins du site cible dont la forme est :

$$\nu_{ij} = 1 - \left(\frac{d_{ij}}{\mathrm{TP}}\right)^n \tag{4.9}$$

où TP et n sont des paramètres de la fonction de pondération.

4.3.2.4 La méthode d'analyse des corrélations canoniques

La méthode d'analyse des corrélations canoniques permet d'identifier les sites dont le régime de crue est similaire au site cible (Cavadias, 1989, 1990; Ribeiro-Corréa et al., 1995; Ouarda et al., 1997, 1998). L'analyse des corrélations canoniques est un outil d'analyse statistique multivariée qui permet de décrire la relation de dépendance existant entre deux ensembles de variables aléatoires. Cette méthode permet de déterminer des paires de combinaisons linéaires

de chaque ensemble de variables, que l'on appelle des variables canoniques, telles que la corrélation entre les variables canoniques d'une paire est maximisée, et la corrélation entre les variables de paires différentes est nulle. On obtient ainsi un ensemble de variables canoniques pour les deux ensembles de variables aléatoires associées à des coefficients de corrélation canonique. Il est alors possible d'inférer sur les variables canoniques d'un ensemble connaissant les variables canoniques de l'autre ensemble. On peut aussi calculer une distance entre deux variables canoniques ce qui permet de déterminer des voisinages hydrologiques. Lorsqu'on ne dispose que d'une seule variable dépendante, l'analyse des corrélations canoniques revient à une analyse de régression multiple.

Plus formellement, on note $\mathbf{Y}^T = (Y_1, Y_2, ..., Y_p)$ et $\mathbf{X}^T = (X_1, X_2, ..., X_q)$ respectivement l'ensemble des variables hydrologiques et l'ensemble des variables caractérisant la géomorphologie et la météorologie des bassins versants. On note $\mathbf{W}$ le vecteur de variables canoniques hydrologiques, $\mathbf{V}$ le vecteur de variables physiographiques, $(\lambda_1, \lambda_2, ..., \lambda_p)$ les coefficients de corrélation canonique, et Λ la matrice diagonale formée par les coefficients de corrélation canonique. On suppose que les vecteurs des variables canoniques $\mathbf{W}$ et $\mathbf{V}$ suivent une distribution multi-normale. La distribution conditionnelle de $\mathbf{W}$ étant donné $\mathbf{V}$ est alors p-normale. Par conséquent, des bassins physiographiquement semblables et représentés par un vecteur canonique commun $\mathbf{V}$, seront répartis autour d'une position moyenne $\Lambda\mathbf{V}$ dans l'espace canonique hydrologique. La distance à la position moyenne est contrôlée par la forme quadratique de la distribution conditionnelle représentée par une distance de Mahalanobis avec une distribution du Khi-deux à p degrés de liberté (χ^2_p). Pour les bassins non jaugés, la position moyenne est déterminée, en utilisant une estimation de Λ, par $\Lambda\mathbf{V_0}$ où $\mathbf{V_0}$ est le vecteur canonique physiographique connu. On peut ainsi définir le voisinage d'un bassin non-jaugé à un niveau de confiance $(1-\alpha)$ par l'ensemble de bassins dont la position $\mathbf{W}$ dans l'espace canonique hydrologique vérifie la relation :

$$(\mathbf{W} - \hat{\Lambda}\mathbf{V_0})^T(\mathbf{I}_p - \hat{\Lambda}\hat{\Lambda}^T)^{-1}(\mathbf{W} - \hat{\Lambda}\mathbf{V_0}) \leq \chi^2_{\alpha,p} \tag{4.10}$$

où $\mathbf{I}_p$ est la matrice identité d'ordre p.

4.3.3 Les principaux modèles régionaux

Supposons qu'en chacune des N stations de jaugeage d'une région donnée, l'on dispose d'une série de taille $n_{j\,(j=1...N)}$ d'observations homogènes et indépendantes $(x_1, x_2, ..., x_{n_j})$ de DMA. Supposons de plus qu'en chacune de ces N stations, l'on dispose d'une série $(X_1, X_2, ..., X_p)$ de

p caractéristiques physiographiques et/ou climatologiques du bassin versant de la station. Supposons enfin que l'on désire produire une estimation d'un quantile de crue Q_T en un site non jaugé mais où l'on dispose néanmoins des p caractéristiques physiographiques et/ou climatologiques du bassin versant du site cible. Dans cette situation, deux approches de modélisation régionale ont principalement été utilisées par les hydrologues : la méthode de l'indice de crue et la méthode de la régression régionale des quantiles.

4.3.3.1 La méthode de l'indice de crue

Thomas Jr. (1994) rapporte que la méthode de l'indice de crue (*index flood*) a été développée durant les années 40 par les ingénieurs du USGS. Gupta et Waymire (1997) associent la méthode à Kinnison et Colby (1945) alors que la majorité des auteurs font plutôt référence à Dalrymple (1960) pour l'introduction de cette méthode en hydrologie. L'hypothèse de base de la méthode est que les données aux différents sites d'une région sont indépendantes et suivent la même distribution statistique à un facteur d'échelle près. La méthode de l'indice de crue comprend trois étapes. La première étape consiste à standardiser les données, c'est-à-dire qu'à chaque site i et pour chaque année t, les données $x_{i,t}$ sont standardisées en divisant par un indicateur de tendance centrale (moyenne, médiane, etc.) μ_i (l'indice de crue). À l'étape 2, les données standardisées $x_{i,t}/\mu_i$ des différents sites sont regroupées afin d'estimer la distribution régionale $F_{\mathrm{R}}(x;\hat{\Theta})$ puis le quantile régional correspondant $\hat{Q}_T^{\mathrm{R}} = F_{\mathrm{R}}^{-1}(1-1/T;\hat{\Theta})$. L'étape 3 consiste à estimer, généralement par régression en fonction des caractéristiques physiographiques, l'indice de crue μ_S du site non jaugé S où l'on désire produire une estimation (le site cible). Le quantile de crue désiré $\hat{Q}_T^S$ au site cible S est alors obtenu par :

$$\hat{Q}_T^S = \hat{Q}_T^{\mathrm{R}}\,\mu_S \tag{4.11}$$

L'hypothèse que les données aux différents sites d'une région sont indépendantes et suivent la même distribution statistique à un facteur d'échelle près équivaut à supposer que le coefficient de variation (CV) et tous les autres ratios de moments d'ordre supérieur sont égaux à chacun des sites régionaux. Une autre hypothèse intrinsèque à la méthode est l'égalité des quantiles de crue normalisés Q_T^i/μ_i en chacun des sites i d'où, pour une valeur de T fixée, le modèle régional :

$$\frac{Q_T^i}{\mu_i} = \beta_0 + \epsilon_i \tag{4.12}$$

où β_0 est une constante représentant le quantile de crue normalisé régional moyen et ϵ_i est la composante d'erreur du modèle.

Aux États-Unis, le USGS a d'abord utilisé la méthode de l'indice de crue durant les années quarante et ce, jusqu'au début des années soixante. Cependant, les études de Dawdy (1961) et de Benson (1962) ont montré que pour plusieurs régions des États-Unis, l'hypothèse d'un CV constant ne pouvait s'appliquer puisque empiriquement, CV tend à diminuer lorsque l'aire des bassins versants augmente. C'est pourquoi le USGS a remplacé la méthode de l'indice de crue par celle de la régression régionale des quantiles. Mentionnons qu'à cette époque, le terme région n'était utilisé que pour désigner une région géographique. Avec le développement des méthodes permettant la détermination de régions homogènes où le CV est constant à l'échelle régionale, les critiques du USGS quant à la constance du CV n'avaient plus leur raison d'être. La méthode de l'indice de crue a alors été réintroduite en hydrologie et a fait l'objet de nouveaux développements : nouvelles distributions régionales, utilisation des moments pondérés par probabilités (réf. Annexe A) plutôt que des moments ordinaires, etc. La performance de la méthode dépend cependant de deux facteurs importants : il doit être possible (1) d'identifier adéquatement des régions homogènes et (2) d'assigner correctement le site non jaugé à une région homogène. Gupta et al. (1994) mentionnent à cet effet qu'en abandonnant la méthode de la régression régionale pour la méthode de l'indice de crue, on se trouve à transférer le problème de l'estimation du modèle vers la détermination des régions homogènes et non à le résoudre. Fill et Stedinger (1998) indiquent quant à eux que l'application de cette méthode pour l'estimation en des sites non jaugés demeure problématique puisqu'il est souvent difficile d'assigner un site non jaugé à une région homogène.

4.3.3.2 La méthode de la régression régionale des quantiles

La méthode de la régression régionale permet d'établir une relation directe entre les quantiles de crue Q_T et les variables explicatives physiographiques et météorologiques $X_1, X_2, ..., X_p$ des bassins versants. La méthode a l'avantage d'être simple, rapide et de permettre d'utiliser des distributions de crue différentes pour représenter les débits de crue dans chacun des sites d'une même région. De plus, la méthode n'est pas sensible à l'hétérogénéité qui peut exister dans la région considérée. En pratique, la fonction ayant la forme de puissance suivante :

$$Q_T = \alpha_0 X_1^{\beta_1} X_2^{\beta_2} ... X_p^{\beta_p} \tag{4.13}$$

de paramètres $\alpha_0, \beta_1, \beta_2, ...$ et β_p est la plus généralement utilisée. L'estimation de cette fonction non linéaire est généralement effectuée en utilisant une transformation logarithmique de façon à obtenir le modèle classique de régression linéaire multiple suivant :

$$\log(Q_T^i) = \beta_0 + \beta_1 \log(X_1^i) + \beta_2 \log(X_2^i) + ... + \beta_p \log(X_p^i) + \epsilon_i \tag{4.14}$$

où $\beta_0, \beta_1, \beta_2, ..., \beta_p$ sont des paramètres à estimer et ϵ_i est la composante d'erreur. En supposant une telle relation régionale, il est alors possible d'appliquer la technique de régression linéaire multiple pour l'estimation des paramètres du modèle de même que pour le choix des variables explicatives.

Benson (1962) a utilisé la méthode des moindres carrés ordinaires pour l'estimation des paramètres de (4.14). Cependant, puisque la variable dépendante Q_T du modèle de régression est en réalité un estimateur ayant déjà fait l'objet d'une estimation locale, l'hypothèse de variance constante inhérente à la méthode des moindres carrés ordinaires est peu justifiable. En effet, les estimateurs locaux des quantiles, étant basés sur des tailles d'échantillon (de DMA) différentes, devraient avoir des variances (inversement) proportionnelles à la taille des échantillons. Afin de remédier à ce problème, il est alors possible d'utiliser la méthode des moindres carrés pondérés (Stedinger et Tasker, 1985, 1986) qui tient compte de ces variances différentes. Un autre problème associé à l'utilisation de la régression régionale est qu'en raison de la dépendance spatiale entre les stations, il est probable que les erreurs du modèle soient corrélées. La méthode des moindres carrés généralisés a alors été proposée pour faire face à cette problématique (Stedinger et Tasker, 1985, 1986). Enfin, un dernier problème relatif à l'estimation régionale par régression des quantiles de crue est la dépendance généralement observée entre les différentes variables explicatives. Afin de pallier à ce problème de collinéarité, Roy et al. (1989) ont proposé l'utilisation de la régression ridge afin d'estimer de manière plus appropriée les paramètres de (4.14). Mentionnons enfin que l'équation 4.14 suppose un terme d'erreur multiplicatif en 4.13. Nguyen et Pandey (1994) ont utilisé un algorithme d'optimisation non-linéaire afin d'estimer directement 4.13 sous l'hypothèse d'un terme d'erreur additif.

4.3.3.3 Les principaux modèles alternatifs

Au cours des dernières années, de nouveaux modèles régionaux sont apparus dans la littérature afin de combler certaines des lacunes de leurs prédécesseurs. Par exemple, réalisant que l'hypothèse de base de la méthode de l'indice de crue est inconsistante avec les relations connues entre le CV et l'aire des bassins versants, Fill et Stedinger (1998) indiquent que les méthodes d'estimation régionale devraient faire intervenir davantage la dépendance observée des quantiles de crue avec l'aire des bassins versants et les autres caractéristiques physiographiques importantes. Ils proposent l'utilisation de la méthode de la régression des quantiles normalisés, développée par Fill (1994), qui consiste à combiner les modèles de l'indice de crue et de la régression régionale des quantiles afin d'obtenir un modèle dont la forme est :

$$\log(\frac{Q_T^i}{\mu_i}) = \beta_0 + \beta_1 X_1^i + \beta_2 X_2^i + ... + \beta_p X_p^i + \epsilon_i \tag{4.15}$$

Ce modèle permet d'appliquer la notion de *scaling* propre à la méthode de l'indice de crue en des sites ne respectant pas nécessairement l'hypothèse de constance régionale du quantile normalisé (ou CV) puisque celui-ci se trouve modélisé à l'échelle régionale par régression linéaire. On peut remarquer que cette approche constitue une généralisation de la méthode de l'indice de crue puisqu'en présence d'un CV constant, l'estimation devrait conduire à $\beta_1 = \beta_2 = ... = \beta_p = 0$.

Considérant qu'il n'existe aucune justification physique à la sélection d'une relation linéaire multiple entre le logarithme des quantiles de crue Q_T et les différentes variables physiographiques et climatologiques des bassins versants, Gingras et al. (1995) proposent plutôt, pour la modélisation régionale des quantiles de crue, l'utilisation de la régression non paramétrique suivante :

$$\log(Q_T^i) = s_\lambda(X_1^i, X_2^i, ..., X_p^i) + \epsilon_i \tag{4.16}$$

où s est une fonction (courbe ou surface) de régression lisse (*smooth*) de forme non spécifiée *a priori* de dimension p, λ est le paramètre de lissage et ϵ_i la composante d'erreur. Gingras et al. (1995) ont estimé le modèle de l'équation 4.16 à l'aide du lissage par Noyau pour des fonctions de régression de dimension $p = 1$ et $p = 2$.

L'analyse régionale est implicitement ou explicitement basée sur la présence d'une distribution de probabilité des crues annuelles F dépendante de paramètres régionaux $\underline{\Theta}$, d'où la relation (Rossi et Villani, 1994a) :

$$Q_T = F^{-1}(1 - 1/T; \underline{\Theta}) \tag{4.17}$$

où $\underline{\Theta}$ est un vecteur de longueur p de caractéristiques de crue (paramètres de la distribution, moments, ratios de moments, L-moments, etc.). Pour procéder à l'estimation de Q_T, de nombreuses études régionales (e.g. Rossi et Villani (1994b) en Italie, Mimikou (1990) en Grèce) consistent à utiliser un modèle de transfert, par exemple de régression régionale, afin d'estimer chacun des paramètres de $\underline{\Theta}$ au site cible. Le modèle de cette approche de régionalisation est donc représenté par un ensemble de p relations régionales où chacune a pour objectif de transférer un paramètre particulier de $\underline{\Theta}$ au site cible.

Suite aux nombreuses recommandations (e.g. Yevyevich (1974), Potter (1987), NRC (1988), Bobée et Rasmussen (1995)) se retrouvant dans la littérature à l'effet que l'on devrait développer davantage les approches d'analyse de la distribution des crues basées sur les phénomènes hydrologiques (physiques) responsables des crues plutôt que les approches purement statistiques ou mathématiques, Roy (1993) a développé la méthode HYBRIIDS, une méthode combinant les approches statistique et déterministe (physique) basée sur l'estimation statistique de données simulées à partir d'un modèle déterministe de bassin versant. Gupta et al. (1994) ont quant à eux développé une théorie physique-statistique régionale des crues basée sur les notions d'invariance d'échelle simple et multiple.

La théorie de l'invariance d'échelle simple suppose la présence d'un CV constant et est donc étroitement reliée à la méthode de l'indice de crue. Pour les théories d'invariance d'échelle proposées par Gupta et al. (1994), seule l'aire des bassins versants (A) est utilisée comme facteur d'échelle ($\approx$ indice de crue). Il est possible de montrer (Gupta et al., 1994) que lorsque les hypothèses d'invariance d'échelle simple sont respectées alors (1) la relation entre $\log(A)$ et $\log(Q_T)$ est linéaire et (2) la pente de cette relation ne dépend pas de T. En utilisant la notation $Q_T(A)$ pour insister sur la notion de dépendance de l'aire sur les quantiles de crue, (1) et (2) peuvent se traduire par :

$$Q_T(A) = c(T)A^{\theta} \tag{4.18}$$

où $c(T)$ est un coefficient qui dépend de la période de retour T alors que l'exposant de mise à l'échelle θ n'en dépend pas. Lorsque ces hypothèses ne sont pas respectées, il est alors possible d'utiliser le modèle d'invariance d'échelle multiple :

$$Q_T(A) = c(T)A^{\theta(T)} \tag{4.19}$$

pour lequel l'exposant de mise à l'échelle θ dépend de T. Avec le modèle d'invariance d'échelle multiple, le CV varie en fonction de l'aire des bassins versants selon la relation représentée à la figure 4.1. On peut remarquer la présence d'une relation différente pour les petits ($\approx< 50\text{km}^2$) et les grands bassins versants. Gupta et Dawdy (1995) ont étudié les variations régionales de l'exposant de mise à l'échelle θ à l'aide des équations de régression de trois États américains. Leurs conclusions suggèrent que pour les régions où les crues sont causées principalement par la fonte de neige, l'hypothèse d'invariance d'échelle simple semble vérifiée alors que pour les crues causées davantage par des épisodes de précipitations, le modèle d'invariance d'échelle multiple doit plutôt être employé.

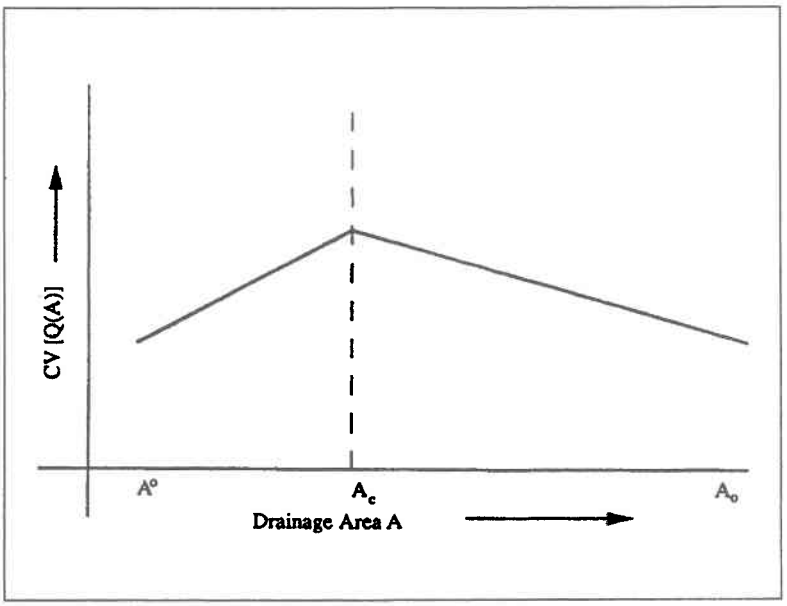

FIG. 4.1: Variation régionale du CV en fonction de l'aire (A) des bassins versants (tiré de Gupta et al., 1994)

5. APPLICATIONS ET COMPARAISON

Dans ce chapitre, nous appliquons et évaluons l'approche de modélisation additive par polynômes locaux à trois régions hydrologiques des États-Unis. Nous débutons ce chapitre par une description, dans la section 5.1, des diverses données et bases de données utilisées pour la modélisation. Nous effectuons par la suite, dans la section 5.2, une présentation de la méthodologie d'évaluation et de comparaison de l'approche de modélisation par polynômes locaux, de même que des diverses approches de modélisation retenues pour fin de comparaison. Cette méthodologie est ensuite appliquée, dans les sections 5.3 à 5.5, respectivement aux régions du Texas, de la Nouvelle-Angleterre et de l'Arkansas. Nous apportons finalement, dans la section 5.6, certaines conclusions sur les résultats obtenus lors de ces applications.

5.1 Présentation des données

Aux États-Unis, la gestion du réseau de mesures hydrologiques est assurée par le *United States Geological Survey* (USGS), le même organisme, présenté au chapitre 1, impliqué dans le développement des procédures de régionalisation pour chacun des États américains. Au début des années quatre-vingt dix, le USGS a regroupé diverses bases de données régionales de manière à constituer une base de données nationale afin de permettre l'étude des conditions de l'écoulement suite à des fluctuations de conditions climatiques. Cette base de donnée, appelée *Hydro-Climatic Data Network* (HCDN), est constituée d'observations de débits provenant de 1659 sites à travers les États-Unis et ses Territoires (Slack et al., 1993). Le choix des divers sites de même que des diverses observations de débit a été effectué de manière rigoureuse en respectant des critères précis sur l'exactitude des mesures et sur les conditions naturelles de l'écoulement : aucune série de débit n'a fait l'objet de quelque reconstruction que ce soit (pour plus de détails, voir Slack et Landwehr (1992)). Les sites retenus devaient aussi généralement contenir au moins 20 années d'observations. À chacun des sites de la base de données est associé un vecteur de caractéristiques physiographiques et climatologiques tels que la superficie du bassin versant, l'élévation moyenne du bassin, la pente et la longueur du cours d'eau principal et la précipitation moyenne annuelle survenant sur le bassin versant.

Dans le cadre du projet HCDN, les États-Unis ont été divisés en 21 grandes régions hydrologiques, présentées dans la figure 5.1. Afin de déterminer des stations propices à l'étude, nous avons effectué, dans chacune de ces régions, une analyse préliminaire de la log-linéarité existant

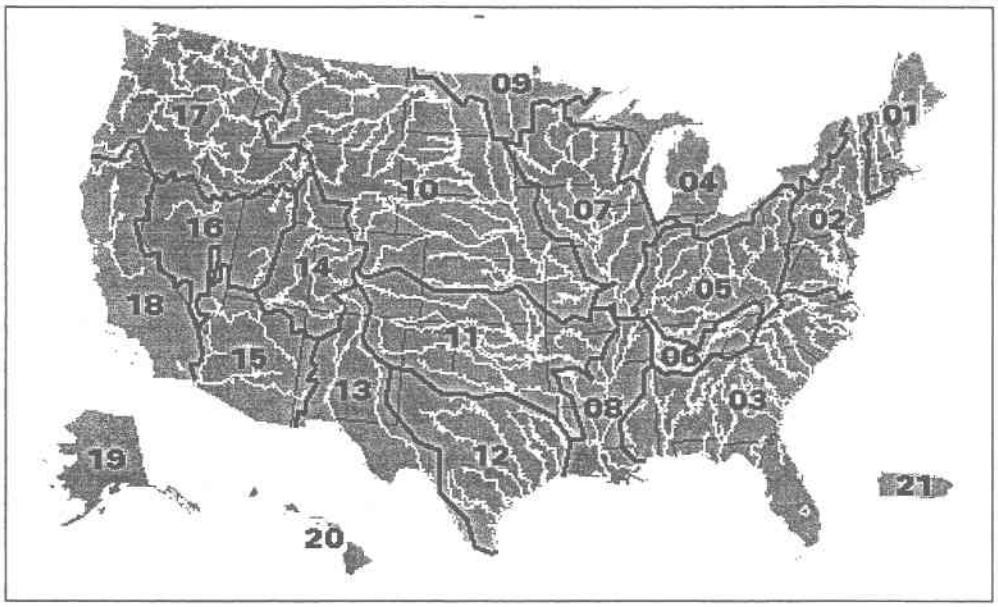

FIG. 5.1: Les régions hydrologiques Américaines du HCDN

entre le quantile de crue Q_{50}, le quantile étudié lors des études de régression par région d'influence de Tasker et Slade (1994) et de Tasker et al. (1996), et la superficie des bassins versants, la variable explicative apparaissant le plus fréquemment et le plus significativement lors d'études de régression des quantiles (Jennings et al., 1994). Le tableau 5.1 montre les résultats obtenus par l'application du test F présenté dans la section 3.4.3. Nous avons indiqué en caractère gras les régions où l'on observe, selon ce test F, une réduction significative au niveau de confiance $\alpha = 0,05$, c'est-à-dire pour $\alpha(F) < 0,05$ (cf. 3.4.3), de la variance de l'erreur en utilisant la modélisation additive plutôt qu'un modèle paramétrique log-linéaire. Pour fins d'analyse, nous avons choisi de retenir une région où le test est significatif et une région où il ne l'est pas. Nous avons ainsi choisi, comme région où la modélisation additive est susceptible de produire de bons résultats, la région 12, la région du Texas et du Golfe du Mexique. Nous avons arrêté notre choix sur la région 12 puisque cette région est composée principalement de sites du Texas, la région étudiée par Tasker et Slade (1994). Quant à l'autre région, nous avons choisi la région 01, la région de la Nouvelle-Angleterre, en raison de sa proximité du Québec.

Une caractéristique propre aux diverses régions du HCDN est la faible densité spatiale des stations, c'est-à-dire par le fait même, la présence d'une faible corrélation spatiale entre les mesures de débits des sites de la région. Afin de vérifier l'effet de la corrélation spatiale sur les résultats obtenus par la modélisation additive, nous avons aussi voulu appliquer l'approche de modélisation additive à une région caractérisée par un plus fort niveau de corrélation spatiale.

TAB. 5.1: Les régions hydrologiques du HCDN : Analyse préliminaire de la log-linéarité

Région	Nom de la région	Nombre de stations	$\alpha(F)$
01	*New England*	71	0,289
02	***Mid-Atlantic***	**167**	**0,042**
03	***South Atlantic-Gulf***	**193**	**0,011**
04	*Great Lakes*	57	0,749
05	***Ohio***	**108**	**0,004**
06	*Tennessee*	44	0,231
07	*Upper Mississippi*	127	0,060
08	*Lower Mississippi*	23	0,763
09	*Souris-Red-Rainy*	39	0,247
10	***Missouri***	**144**	**0,017**
11	*Arkansas-White-Red*	87	0,239
12	***Texas-Gulf***	**90**	**0,017**
13	*Rio Grande*	22	0,081
14	*Upper Colorado*	44	0,329
15	*Lower Colorado*	17	0,056
16	*Great Basin*	32	0,289
17	***Pacific Northwest***	**191**	**0,004**
18	*California*	115	0,716
19	***Alaska***	**31**	**0,017**
20	*Hawaii*	42	0,953
21	*Caribbean*	15	0,123

Pour ce faire, nous avons utilisé les mêmes données que Tasker et al. (1996), c'est-à-dire les données provenant de 204 stations de l'Arkansas. Les données du Texas (région 12) et de la Nouvelle-Angleterre (région 01) du HCDN, disponibles sur internet via FTP à l'adresse internet **ftp ://ftprvares.er.usgs.gov/hcdn92/**, ont été reproduites respectivement aux annexes B et C. Les données de l'Arkansas, publiées intégralement dans Hodge et Tasker (1995), se retrouvent quant à elles à l'annexe D.

5.2 La méthodologie d'évaluation et de comparaison

L'objectif principal de cette recherche est d'évaluer l'utilisation de la modélisation additive par polynômes locaux comme méthode alternative d'estimation régionale des débits de crue Q_T en des sites non jaugés. Afin d'atteindre cet objectif, nous avons choisi d'évaluer les qualités **prédictives** de cette approche et de la comparer à l'approche de la régression régionale par

région d'influence de même qu'à l'approche classique de régression par un modèle log-linéaire. Ces trois approches d'estimation régionale des débits de crue Q_T peuvent être représentées par le modèle de type boîte noire suivant :

$$\log(X_1), \log(X_2), ..., \log(X_d) \longrightarrow \blacksquare \longrightarrow \log(Q_T) \tag{5.1}$$

où $\log(X_1), \log(X_2), ..., \log(X_d)$ représentent en entrée le logarithme décimal de variables physiographiques / climatologiques prédictives $X_1, X_2, ..., X_d$ associées à un site particulier et $\log(Q_T)$ représente la réponse (ou variable dépendante) du modèle, le logarithme décimal du quantile de crue Q_T du site en question.

5.2.1 Les simulations vs les procédures de séparation de données

De manière générale, deux grandes approches peuvent être utilisées pour la comparaison. Il est d'abord possible d'effectuer des simulations de type Monte Carlo à partir de relations fonctionnelles (représentées par la boîte noire) paramétriques connues. Gingras et al. (1995) ont utilisé ce genre d'approche afin d'évaluer l'utilisation de la régression par noyau pour la régionalisation des quantiles de crue. Ils ont simulé des relations linéaires, quadratiques et exponentielles qu'ils ont par la suite estimées à l'aide de modèles de régression paramétriques et non paramétriques. Cette façon de faire est cependant discutable puisqu'en pratique, les véritables relations entre les variables de réponse et les variables explicatives sont inconnues. En ce sens, nous proposons plutôt l'utilisation de **procédures de séparation de données** (*data splitting procedures* ou *split-sample experiments*) qui consistent à séparer les observations disponibles en un **échantillon d'estimation** servant à la calibration du modèle et en un **échantillon de prédiction** servant à la validation du modèle, c'est-à-dire à l'évaluation de ses propriétés prédictives.

Les procédures de séparation de données, aussi appelées **techniques de rééchantillonnage**, simulent la collection de nouvelles données (ici, de nouveaux sites) afin de vérifier les habilités prédictives des modèles. En pratique, il n'existe pas de règles générales sur la façon de séparer les données. Une technique de rééchantillonnage couramment utilisée, appelée ***jackknife*** (Quenouille, 1956), consiste à répéter, lorsqu'on dispose de n observations (sites), n fois la procédure de séparation consistant à laisser de côté tour à tour chacune des observations. Il s'agit alors d'utiliser des échantillons d'estimation composés de $n-1$ observations et des échantillons de validation composés d'une seule observation. Nous utiliserons cette approche pour les trois régions sous étude.

Mentionnons que pour la comparaison de procédures de régression paramétriques à des procédures de régression non paramétriques, sur la base des erreurs de prédictions obtenues dans l'échantillon de prédiction, le choix de la taille de l'échantillon peut jouer un rôle important sur l'évaluation comparative des différentes approches. En effet, puisque rappelons-le, les estimateurs paramétriques convergent plus rapidement (par exemple, en moyenne quadratique) que les estimateurs non paramétriques vers la véritable fonction de régression, une taille d'échantillon trop petite devrait avoir tendance à désavantager l'approche non paramétrique. En ce sens, la procédure de *jackknife* constitue une approche de comparaison des plus objectives puisqu'il s'agit de la procédure employant la plus grande taille possible pour l'échantillon d'estimation.

5.2.2 Les critères de comparaison

Les procédures de séparation de données permettent de vérifier les capacités prédictives des modèles. Il suffit de procéder à une analyse des prédictions effectuées par les différents modèles. Il est d'abord possible d'effectuer un examen visuel de la distribution des erreurs (absolues, relatives, quadratiques, etc.) de prédictions afin d'en mesurer, par exemple, la dispersion, à l'aide de graphiques en boîte (*box-plot*). Il est cependant plus courant d'effectuer une moyenne des différentes erreurs de prédictions. Dans ce document, nous avons retenu comme mesure d'évaluation des capacités prédictives des modèles, la racine carrée des erreurs quadratiques moyennes (RMSE) (pour *root mean square error*) de **prédiction**, calculée dans l'espace logarithmique,

$$\mathrm{RMSE}^{\mathrm{PRED}} = \sqrt{1/N_p \sum_{i \in \mathrm{PRED}} \left(\log(Q_{T,i}) - \log(\hat{Q}_{T,i})\right)^2} \tag{5.2}$$

de même que la déviation relative moyenne (DRM), exprimée en pourcentage, des **prédictions**,

$$\mathrm{DRM}^{\mathrm{PRED}} = 1/N_p \sum_{i \in \mathrm{PRED}} \frac{|Q_{T,i} - \hat{Q}_{T,i}|}{Q_{T,i}} \tag{5.3}$$

où les sommations sont effectuées sur les sites de l'échantillon de prédiction (PRED), $Q_{T,i}$ est l'estimateur du quantile de crue (de période de retour T) du site i basé sur les données de DMA du site en question, $\hat{Q}_{T,i}$ représente l'estimation régionale, selon le modèle de régression retenu, du quantile de crue au site i et N_p est le nombre de sites dans l'échantillon (ou les échantillons) de prédiction.

5.2.3 La calibration des différents modèles

La modélisation non paramétrique se distingue de l'approche paramétrique principalement par deux caractéristiques importantes : (1) le temps de calcul nécessaire à l'estimation et (2) le

nécessité d'obtenir un aperçu visuel des estimations. Il est donc nécessaire d'avoir en notre possession des outils informatiques efficaces. Le logiciel S-Plus est un de ces logiciels offrant une interface graphique intéressante de même que des procédures statistiques efficaces d'estimation par lissage et par modélisation additive. La calibration des différents modèles retenus pour la comparaison a donc été effectuée avec le logiciel S-Plus. Mentionnons qu'ici, la calibration des différents modèle consiste à :

1. choisir les variables prédictives à inclure dans le modèle,
2. choisir le niveau de lissage (pour les modèles non paramétriques), et
3. procéder à l'estimation des différents modèles.

Nous présentons, dans ce qui suit, les différentes approches de calibration retenues. Dans bien des cas, le choix de l'approche de calibration repose sur les particularités et limites des différentes procédures disponibles dans le logiciel S-Plus.

5.2.3.1 Le modèle de régression log-linéaire et de régression par région d'influence

Pour la calibration des modèles de régression log-linéaire et de régression par région d'influence, nous utilisons le fait que ces modèles appartiennent à la famille des modèles de régression locale polynomiale multidimensionnelle (Loader, 1999). Le modèle de régression par région d'influence peut ainsi être représenté par un modèle de régression locale polynomiale de degré 1. Le modèle de régression log-linéaire ne représente quant à lui qu'un cas particulier de modèle de régression locale : il s'agit d'utiliser une fonction noyau rectangulaire et de choisir le paramètre de largeur de fenêtre de manière à ce que toutes les observations appartiennent à celle-ci. Rappelons enfin que la seule distinction pouvant être apportée entre le modèle classique de régression locale et le modèle de régression par région d'influence est que ce dernier permet qu'en chacun des points estimés, l'estimation de la fonction de régression puisse être basée sur des variables (ou ensembles de variables) explicatives différentes alors qu'avec le modèle de régression locale, une fois choisies, les variables explicatives sont obligatoirement toutes utilisées en chacun des points (sites) où l'on désire effectuer une estimation.

En régression régionale des quantiles de crue, il est bien connu (Stedinger et Tasker, 1985, 1986) qu'un modèle de moindres carrés pondérés ou généralisés est plus représentatif de la problématique réelle de l'estimation régionale des quantiles en raison respectivement de la variance inégale des estimateurs locaux $Q_{T,i}$ (basés sur des tailles d'échantillon différentes) et de la corrélation spatiale annuelle pouvant exister entre des DMA de sites géographiquement voisins. Le modèle de régression locale ne permet l'utilisation que d'une approche particulière d'estimation

par moindres carrés pondérés. L'estimation requiert une connaissance a priori de la pondération accordée à chacune des observations (chacun des sites). Nous utiliserons une telle approche en pondérant chacun des sites selon la taille des échantillons à ces sites.

En S-Plus, deux procédures de régression locale multidimensionnelle (et unidimensionnelle) sont disponibles : la procédure **loess**, incluse automatiquement dans le logiciel et la procédure **locfit**, un module externe de Clive Loader, l'auteur d'un livre paru récemment sur la régression locale (e.g. Loader (1999)). La procédure loess est plus efficace au niveau du temps de calcul des estimations mais ne permet pas de choisir une fonction de pondération autre que la fonction triplement cubique alors que la procédure **locfit** permet l'utilisation de chacune des fonctions de pondération (noyau) présentées dans le tableau 2.1. Puisque nous avons, dans une certaine mesure, comme objectif de reproduire les résultats de Tasker et Slade (1994) et de Tasker et al. (1996) qui utilisent une fonction de pondération rectangulaire, nous avons donc choisi d'utiliser la procédure **locfit**.

La procédure **locfit** possède une option intéressante. Il est en effet possible, en utilisant l'option d'estimation par validation-croisée (*evaluation=cross*), d'effectuer (très rapidement) une estimation en chacun des n sites en laissant de côté l'information du site en question et en n'utilisant que l'information provenant des $n-1$ autres sites. Cette option permet ainsi de calculer rapidement les critères de comparaison des équations 5.2 et 5.3. Nous utiliserons tout particulièrement le RMSE prédictif de l'équation 5.2 tant pour choisir les variables prédictives à inclure dans les différents modèles que pour déterminer le nombre optimal de sites à inclure dans la région d'influence.

5.2.3.2 La modélisation additive par polynômes locaux

Avec S-Plus, le module **gam** (*generalized additive models*) permet la modélisation additive. L'estimation s'effectue à l'aide de l'algorithme de *backfitting* (cf. 3.3.1). Nous employons le lisseur par polynômes locaux du module **locfit**. Mentionnons que par défaut, pour un échantillon de taille n, la procédure **locfit** n'effectue pas une estimation en chacun des n points. En effet, de manière à diminuer le temps de calcul, la procédure effectue plutôt une estimation en un sous-ensemble de points $n_1 < n$ et procède par la suite par interpolation pour produire les autres estimations. Il en résulte alors des matrices de lissage non pas de dimension n mais plutôt de dimension n_1. En procédant ainsi, on se trouve aussi à faire une approximation du nombre ν de paramètres effectifs des lisseurs (puisque ν est fontion des matrices de lissage (cf. 3.3.2.3)). La procédure **locfit** permet cependant, en utilisant l'option (*evaluation=data*),

d'effectuer le lissage en chacun des n points. Nous utilisons cette option pour la calibration du modèle. Par contre, pour la validation, il est impossible d'utiliser cette option puisqu'il n'est alors pas possible d'effectuer des prédictions en dehors des observations initiales. Nous utilisons donc l'option d'évaluation par défaut pour la validation.

En ce qui concerne la calibration du modèle additif, nous utilisons la procédure de sélection pas-à-pas présentée à la section 3.4.2. Comme critères de sélection, nous avons choisi d'évaluer les résultats obtenus par le critère de validation croisée de l'équation 3.26 pour des valeurs de $c = 1$, ce qui revient à l'équation 3.25, et de $c = 2$, la valeur jugée appropriée par plusieurs auteurs (cf. 3.4.1). En S-Plus, une procédure de sélection pas-à-pas est disponible (**step.gam**). Le critère utilisé par cette procédure est le critère d'AIC de l'équation 3.27. Cependant, puisque S-Plus est avant tout un langage de programmation, il nous a été possible d'effectuer une modification mineure (une seule ligne de code a été changée) à la procédure **step.gam** afin de permettre l'utilisation des critères de validation croisée retenus.

5.3 Application à la région du Texas

Afin de pouvoir appliquer les diverses approches de régionalisation des quantiles de crue retenues pour fin de comparaison, nous devons dans un premier temps procéder à l'estimation des quantiles de crue Q_T aux différents sites jaugés régionaux. Pour ce faire, nous avons choisi d'utiliser la procédure d'estimation locale GEV/PWM, décrite à l'annexe A. Nous avons calculé les quantiles de crue Q_2, Q_5, Q_{10}, Q_{25} et Q_{50} aux différents sites jaugés. Ces résultats sont présentés à l'annexe B (cf. tab. B.1). La modélisation régionale nécessite aussi la présence de variables explicatives (ou prédictives). Le tableau 5.2 présente les variables explicatives retenues pour la région du Texas. Il s'agit de variables que l'on retrouve généralement dans les diverses équations de régression produites par le USGS. Nous avons d'ailleurs indiqué entre parenthèses, au tableau 5.2, le nombre d'États (sur 51) où ces variables étaient jugées significatives (Jennings et al., 1994). Mentionnons enfin que la région du Texas, i.e. la région 12 du USGS, comporte 90 sites mais qu'en raison du manque de certaines variables explicatives en plusieurs sites, la modélisation n'a été effectuée qu'avec les 69 sites présentés à l'annexe B.

TAB. 5.2: Description des variables physiographiques et climatologiques

Variable	Symbole	Définition
Superficie (*Area*)	A (51)	Surface de drainage du bassin versant (km^2)
Pente (*Slope*)	S (27)	Pente du cours d'eau principal, en m/km, mesurée entre le 10 ième et le 85ième percentile de la longueur du cours d'eau
Précipitation (*Precipitation*)	P (19)	Précipitation moyenne annuelle, en cm, survenant sur le bassin versant
Élévation (*Elevation*)	EL (13)	Élévation moyenne du bassin versant, en m au dessus du niveau moyen de la mer
Longueur (*Length*)	L (6)	Longueur du cours d'eau principal, en km, mesurée le long du canal à partir des limites du bassin versant et jusqu'à la station de jaugeage

5.3.1 La modélisation par régression log-linéaire

Le modèle de régression log-linéaire étudié est :

$$\log(Q_T) = \beta_0 + \beta_A \log(A) + \beta_S \log(S) + \beta_P \log(P) + \beta_{EL} \log(EL) + \beta_L \log(L) \qquad (5.4)$$

Rappelons que nous avons choisi d'utiliser le critère de RMSE de l'équation 5.2 afin de déterminer les variables explicatives à inclure dans le modèle. La procédure utilisée consiste à calculer le critère de RMSE pour différents modèles, i.e. avec des variables explicatives différentes, et à retenir le meilleur de ces modèles. Puisque la variable explicative A, l'aire du bassin versant, est généralement la variable explicative la plus significative dans les équations de régression régionale et qu'elle se retrouve dans les équations de régression de toutes les régions des États-Unis (cf. tab. 5.2), nous avons choisi de l'inclure par défaut dans les différents modèles. En procédant ainsi, il ne reste que 16 modèles différents, présentés au tableau 5.3, pour lesquels on doit calculer le RMSE. La légende du tableau 5.3 permet de visualiser graphiquement le RMSE des différents modèles à une, deux, trois, quatre ou cinq variables explicatives. Par exemple, à la figure 5.2, on voit que pour la modélisation du quantile Q_{50}, le modèle numéro 1 à 4 variables explicatives, c'est-à-dire, d'après la légende, le modèle composé des variables A,S,P et EL est optimal au sens du RMSE. Les graphiques correspondant aux périodes de retour (T=2,5,10 et 25) se retrouvent respectivement aux figures B.1, B.2, B.3 et B.4 de l'annexe B. Enfin, le tableau 5.4 présente les valeurs des paramètres des différents modèles optimaux déduits des figures B.1 (T=2), B.2 (T=5), B.3 (T=10), B.4 (T=25) et 5.2 (T=50). Alors que le tableau 5.5 présente les mesures de l'évaluation des qualités descriptives et prédictives de ces modèles.

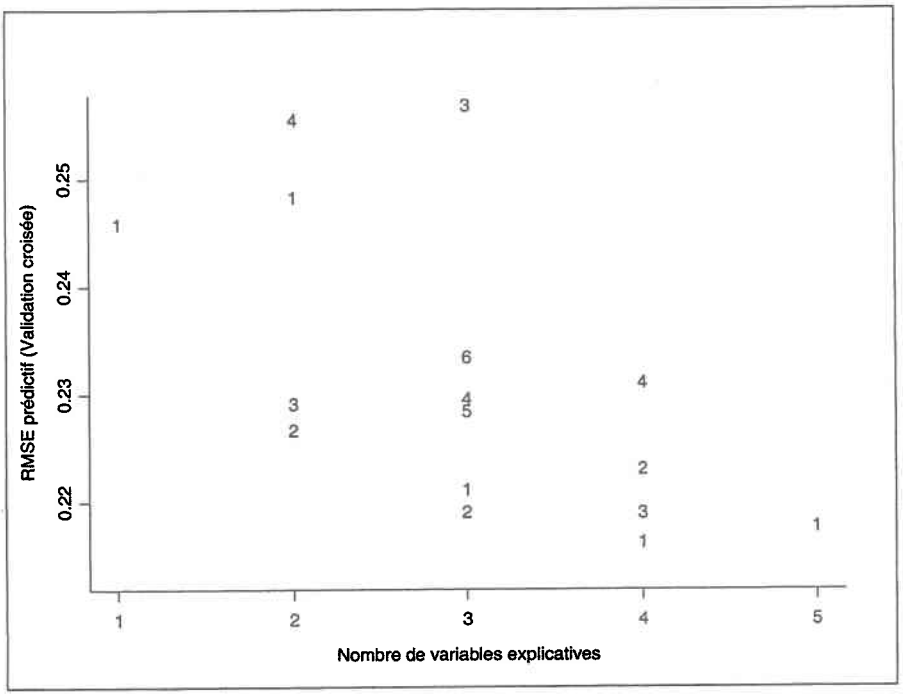

FIG. 5.2: RMSE de différents modèles de régression log-linéaire pour l'estimation de Q_{50} (voir la légende au tableau 5.3)

TAB. 5.3: Légende des différents modèles de régression

Numéro du modèle	Nombre de variables explicatives				
	1	2	3	4	5
1	A	A,S	A,S,P	A,S,P,EL	A,S,P,EL,L
2		A,P	A,S,EL	A,S,P,L	
3		A,EL	A,S,L	A,S,EL,L	
4		A,L	A,P,EL	A,P,EL,L	
5			A,P,L		
6			A,EL,L		

TAB. 5.4: Les paramètres estimés des équations de régression

T	$\hat{\beta}_0$	$\hat{\beta}_A$	$\hat{\beta}_S$	$\hat{\beta}_P$	$\hat{\beta}_{EL}$	$\hat{\beta}_L$
2	0,053	0,548	0,294	1,339	-0,263	0,262
5	0,862	0,531	0,321	1,051	-0,236	0,211
10	1,265	0,535	0,340	0,919	-0,230	0,172
25	1,675	0,594	0,317	0,817	-0,194	-
50	2,013	0,587	0,352	0,697	-0,205	-

TAB. 5.5: RMSE et DRM des modèles de régression log-linéaire

Quantile modélisé	RMSE ($\log_{10}$)		DRM (%)	
	Estimation	Prédiction	Estimation	Prédiction
Q_2	0,164	0,178	30,0	33,1
Q_5	0,161	0,173	30,8	33,7
Q_{10}	0,166	0,179	33,0	36,1
Q_{25}	0,184	0,197	38,8	42,2
Q_{50}	0,202	0,216	43,8	47,7

5.3.2 La modélisation par régression par région d'influence

Le modèle de régression par région d'influence étudié est :

$$\log(Q_T) = S_{k,W}(\log(A), \log(S), \log(P), \log(EL), \log(L)) \tag{5.5}$$

où $S_{k,W}$ est un lisseur de surface (cf. 2.2.8) de paramètres de lissage k et W. Plus précisément, $S_{k,W}$ constitue une généralisation de l'approche du lissage par une droite mobile, présentée à la section 2.2.4, et permet l'estimation, au site i (i.e. au point $\mathbf{x}^i = (A^i, S^i, P^i, EL^i, L^i)$), en ajustant une droite de régression multiple en ce point à l'aide des k plus proches sites voisins, selon la distance de "Mahalanobis" de l'équation 2.24, et en utilisant une fonction de pondération des observations (fonction noyau) W.

TAB. 5.6: RMSE prédictif des fonctions noyau rectangulaire et triplement cubique

Quantile modélisé	Fonction noyau	
	Rectangulaire	Triplement cubique
Q_2	0,168	0,166
Q_5	0,153	0,150
Q_{10}	0,153	0,149
Q_{25}	0,165	0,162
Q_{50}	0,181	0,182

La procédure de calibration du modèle de l'équation 5.5 permet de déterminer (1) les variables prédictives à inclure dans le modèle et (2) le nombre optimal k de stations dans la région d'influence et ce, pour une fonction de pondération W donnée. Nous avons donc voulu, dans un premier temps, comparer l'utilisation de la fonction de pondération triplement cubique (cf. tab. 2.1),

qui est la fonction utilisée par la procédure **loess** de S-Plus, à l'utilisation de la fonction de pondération rectangulaire employée par Tasker et Slade (1994) et par Tasker et al. (1996). Le tableau 5.6 présente les résultats du RMSE prédictif obtenus à l'aide de ces deux fonctions noyau. En raison de l'avantage de la fonction de pondération triplement cubique, suggéré par les résultats du tableau 5.6, nous avons retenu l'emploi de cette fonction de pondération pour la calibration des différents modèles de régression par région d'influence de ce chapitre.

La calibration du modèle de régression par région d'influence s'effectue en minimisant le critère de RMSE par validation-croisée à l'aide de la procédure **locfit** (cf. 5.2.3.1). La figure 5.3 illustre, pour la modélisation du quantile Q_{50}, les valeurs du critère de RMSE pour les différentes valeurs possibles du paramètre k (le nombre de stations) et ce, pour des modèles à une, deux, trois, quatre et cinq variables explicatives. Le choix des variables explicatives de ces différents modèles pour représenter Q_{50} a été effectué à l'aide des résultats de la régression log-linéaire de la figure 5.2 où nous pouvons observer, à l'aide de la légende du tableau 5.3, que le meilleur modèle à deux variables explicatives contient les variables A et P (numéro 2), le meilleur modèle à trois variables contient les variables A,S et EL (numéro 2) et le meilleur modèle à quatre variables contient les variables A,S,P et EL (numéro 1). Nous avons donc retenu ces modèles pour le calcul des RMSE de la figure 5.3. On constate, à la figure 5.3, que le meilleur modèle possède 5 variables explicatives et que le nombre optimal de stations dans la région d'influence pour ce modèle est $k = 60$. La même démarche a été effectuée pour représenter les débits de période de retour T=2,5,10 et 25. Les graphiques du RMSE pour ces quantiles sont présentés aux figures B.5 (T=2), B.6 (T=5), B.7 (T=10) et B.8 (T=25) de l'annexe B. Le tableau 5.7 décrit les différents modèles calibrés. Mentionnons que les valeurs du nombre optimal de stations du tableau 5.7 ont été obtenues à l'aide de la procédure **locfit** et non en visualisant les différents graphiques du RMSE. Le tableau 5.8 présente l'évaluation des qualités prédictives de ces modèles calibrés.

Au chapitre 1, nous avons mentionné que bien que l'approche de la régression régionale par région d'influence soit intéressante, elle implique une grande part de **subjectivité** pour le choix du nombre k de stations à inclure dans la région d'influence. Pour la modélisation de Q_{50} au Texas, Tasker et Slade (1994) ont utilisé k=50 en n'indiquant aucunement leur motivation. Pour la modélisation de Q_{50} en Arkansas, Tasker et al. (1996) motivent leur choix de k=34 par :

> "For this method to work, the value of k should be large enough to have enough degrees of freedom in the regression to estimate two or three parameters. For this study, k was chosen to be 34."

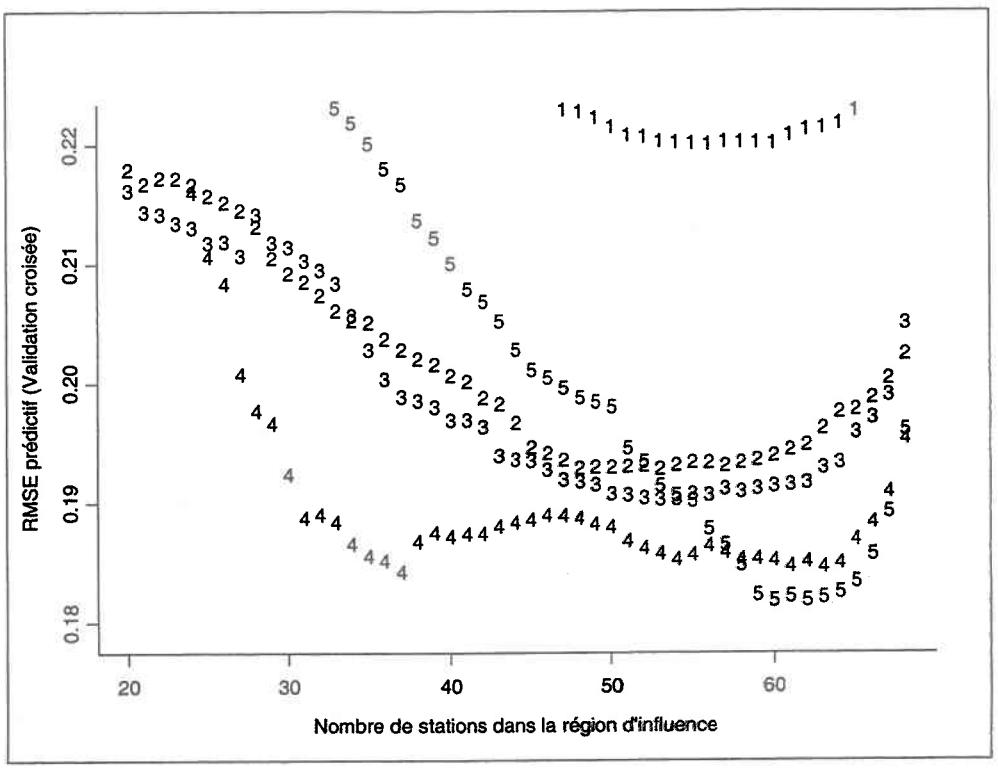

FIG. 5.3: Nombre optimal de stations dans la région d'influence pour Q_{50}

TAB. 5.7: Les modèles calibrés de régression par région d'influence

T	Variables explicatives	Nombre optimal de stations k
2	A,S,P,EL	62
5	A,P,S,EL,L	64
10	A,P,S,EL,L	62
25	A,P,S,EL,L	60
50	A,P,S,EL,L	60

TAB. 5.8: RMSE et DRM des modèles de régression par région d'influence

Quantile modélisé	RMSE ($\log_{10}$)		DRM (%)	
	Estimation	Prédiction	Estimation	Prédiction
Q_2	0,144	0,166	25,1	29,4
Q_5	0,129	0,150	22,9	27,0
Q_{10}	0,128	0,149	23,1	27,5
Q_{25}	0,135	0,162	25,7	31,5
Q_{50}	0,150	0,182	29,0	35,9

Puisque notre approche de calibration permet de choisir k de manière **objective** et **optimale** (au sens du RMSE prédictif), nous avons voulu quantifier l'amélioration pouvant être apportée par notre approche de calibration. Nous avons donc comparé les valeurs du RMSE prédictif obtenues avec notre nombre optimal de stations k^* à celles obtenues en utilisant k=34, le nombre de stations choisi subjectivement par Tasker et al. (1996) et k=50, employé par Tasker et Slade (1994). Le tableau 5.9 présente ces résultats. On peut constater une amélioration moyenne des capacités prédictives de l'ordre de 6,3% pour k=50 et de 16,8% pour k=34.

TAB. 5.9: RMSE prédictif pour différentes valeurs de k

Quantile modélisé	Nombre de stations dans la région d'influence			Amélioration (%)	
	$k = 34$	$k = 50$	k^*	$k^*/34$	$k^*/50$
Q_2	0,185	0,172	0,166	10,5	3,4
Q_5	0,183	0,159	0,150	18,1	6,0
Q_{10}	0,183	0,159	0,149	18,6	6,5
Q_{25}	0,199	0,175	0,162	18,6	7,5
Q_{50}	0,222	0,198	0,182	18,1	8,2
			Moyenne	16,8	6,3

5.3.3 La modélisation additive par polynômes locaux

Nous présentons ici les résultats de la calibration du modèle additif de polynômes locaux :

$$\log(Q_T) = S_{\lambda_1}(\log(A)) + S_{\lambda_2}(\log(S)) + S_{\lambda_3}(\log(P)) + S_{\lambda_4}(\log(EL)) + S_{\lambda_5}(\log(L)) \quad (5.6)$$

de paramètres de lissage $\lambda_{i\,(i=1...5)} \equiv (h_i, p_i, W_i)$. Rappelons que la calibration de ce modèle consiste à rechercher le vecteur de paramètres optimal $\Lambda^* = (\lambda_1^*, \lambda_2^*, \lambda_3^*, \lambda_4^*, \lambda_5^*)$ à l'aide de la procédure de sélection pas-à-pas **step.gam** modifiée pour permettre l'utilisation des critères de validation croisée $\mathrm{VCG}^*(c = 1, \Lambda)$ et $\mathrm{VCG}^*(c = 2, \Lambda)$ (cf. éq. 3.26). Cette procédure de calibration, décrite à la section 3.4.2, requiert que l'on détermine a priori le degré p de chacun des polynômes locaux. Il est aussi préférable de déterminer a priori la fonction noyau (pondération) W employée. La procédure de calibration permet alors d'obtenir un vecteur de paramètres de voisinage optimal $\mathbf{H}^* = (h_1^*, h_2^*, h_3^*, h_4^*, h_5^*)$ où dans ce chapitre, le paramètre de voisinage h_i employé est un paramètre de ***span***. Le *span* représente la proportion des observations (sites) utilisées pour l'estimation. Mentionnons que rechercher les paramètres de *span* optimaux est équivalent à rechercher les tailles optimales des voisinages k_i puisque pour la méthode des k plus proches voisins, on a $h_i = k_i/n$ où n est le nombre total d'observations.

La procédure de calibration pas-à-pas demande en entrée un vecteur de paramètres initiaux $\mathbf{H}^{(0)}=(h_1^{(0)}, h_2^{(0)}, h_3^{(0)}, h_4^{(0)}, h_5^{(0)})$. Pour la modélisation de Q_T, nous avons débuté la procédure de sélection pas-à-pas avec des voisinages de 50% des observations pour chacune des variables explicatives du modèle soit $\mathbf{H}^{(0)}=(0.5, 0.5, 0.5, 0.5, 0.5)$. La recherche des paramètres optimaux s'effectue par la suite par pas de 0.1 pour chacune des variables explicatives. Mentionnons que lorsqu'un des h_i atteint 0, la variable explicative correspondante est enlevée du modèle alors que lorsqu'un des h_i atteint 1, le polynôme local est plutôt remplacé par un terme paramétrique linéaire (cf. 3.4.2). Rappelons aussi qu'avec la procédure de sélection pas-à-pas employée, la convergence vers un optimum global n'est pas assurée. Le choix de la grandeur du pas, ici 0.1, influence cette convergence. En effet, nous avons remarqué que plus le pas employé est petit, plus la procédure a tendance à converger vers un optimum local plutôt que global. De plus, le choix d'un petit pas augmente le temps de calcul puisqu'un plus grand nombre de modèles peuvent potentiellement être considérés. À l'inverse, avec un grand pas, puisqu'un plus petit nombre de modèles peuvent être considérés, il est fort possible que le modèle ne soit optimal que localement. L'utilisation d'un pas de 0.1 nous a semblé être un compromis acceptable pour l'obtention de l'optimum. De plus, en pratique, rien n'empêche de réappliquer la procédure de sélection à partir de l'optimum ainsi obtenu avec un pas plus petit, par exemple 0.01 ou 0.02, pour se rapprocher davantage de l'optimum.

Pour la modélisation de Q_T, nous avons utilisé des polynômes locaux de degré 1 et avons par la suite examiné les différents lissages afin de vérifier si le niveau d'ajustement semblait adéquat. Nous avons aussi voulu vérifier l'influence du choix de certains paramètres de la modélisation sur les qualités prédictives des modèles additifs. Nous avons donc comparé les capacités prédictives de différents modèles additifs de polynômes locaux en fonction :

1. du critère de validation-croisée employé
   (a) $\mathrm{VCG}^*(c=1,\Lambda)$
   (b) $\mathrm{VCG}^*(c=2,\Lambda)$
2. de la fonction noyau utilisée
   (a) triplement cubique
   (b) Gaussienne
   (c) rectangulaire
3. de la fonction de pondération des sites
   (a) aucune pondération
   (b) pondération proportionnelle au nombre d'observations de DMA

Nous avons effectué ces comparaisons pour la modélisation du quantile de crue Q_{50}. Les résultats de cette comparaison sont présentés au tableau 5.10 alors que les graphiques des différents modèles calibrés se retrouvent aux figures 5.4 et 5.5 où nous présentons l'influence du critère de validation croisée, de la fonction noyau et de la fonction de pondération sur le lissage de chacune des variables explicatives des différents modèles additifs. Constatons d'abord, au tableau 5.10, que les meilleurs résultats ont été obtenus par le modèle calibré avec une fonction de pénalité $c = 1$ en utilisant une fonction noyau triplement cubique à l'intérieur de la fenêtre de lissage et en appliquant une fonction de pondération à chacun des sites jaugés. Remarquons ensuite, à la figure 5.5, la très grande similitude existant entre les lissages des différents modèles calibrés à l'aide du critère de validation-croisée de paramètre $c = 1$. Cette similitude nous démontre à quel point il est en pratique difficile de choisir un modèle particulier à l'aide d'un seul examen visuel de la calibration. Ainsi, puisqu'il ne semble pas y avoir de différences notables (pour une valeur du paramètre c fixée) au niveau du lissage entre l'utilisation d'une fonction noyau triplement cubique, Gaussienne ou rectangulaire et entre l'utilisation ou non d'une fonction de pondération des sites, nous nous en remettrons aux résultats du tableau 5.10 pour le choix de ces paramètres de modélisation. Nous opterons donc pour l'utilisation du modèle avec pondération et fonction noyau triplement cubique.

TAB. 5.10: Modélisation additive du quantile Q_{50} : Influence des paramètres de la modélisation sur les capacités prédictives

Paramètres de la modélisation			Paramètres h* du modèle calibré					RMSE ($\log_{10}$)	DRM (%)
VCG	Noyau	Pondération	A	S	P	EL	L		
c=1	tcub	oui	0.4	1	1	0.4	0.5	0,160	30,8
c=1	gauss	oui	0.4	1	1	0.4	0.6	0,162	31,3
c=1	tcub	non	0.3	1	0.9	0.4	0.5	0,164	32,3
c=2	tcub	oui	0.8	1	1	0.5	-	0,168	32,7
c=1	rect	oui	0.2	0.5	0.7	0.4	0.3	0,176	35,1

Alors que pour une valeur de c fixée, il y a peu de distinctions entre les différents lissages, on peut cependant observer, à la figure 5.4, que l'utilisation du critère de validation-croisée VCG*$(c = 2, \Lambda)$ conduit à une estimation plus lisse, notamment pour la variable A dans l'intervalle $[2, 2.5]$ (en unités logarithmiques), qu'avec le critère de paramètre $c = 1$. Mentionnons que l'on devait s'attendre à un tel résultat puisque le paramètre c pénalise le nombre de paramètres effectifs d'un modèle. Or, puisqu'il est possible de montrer qu'une estimation plus lisse

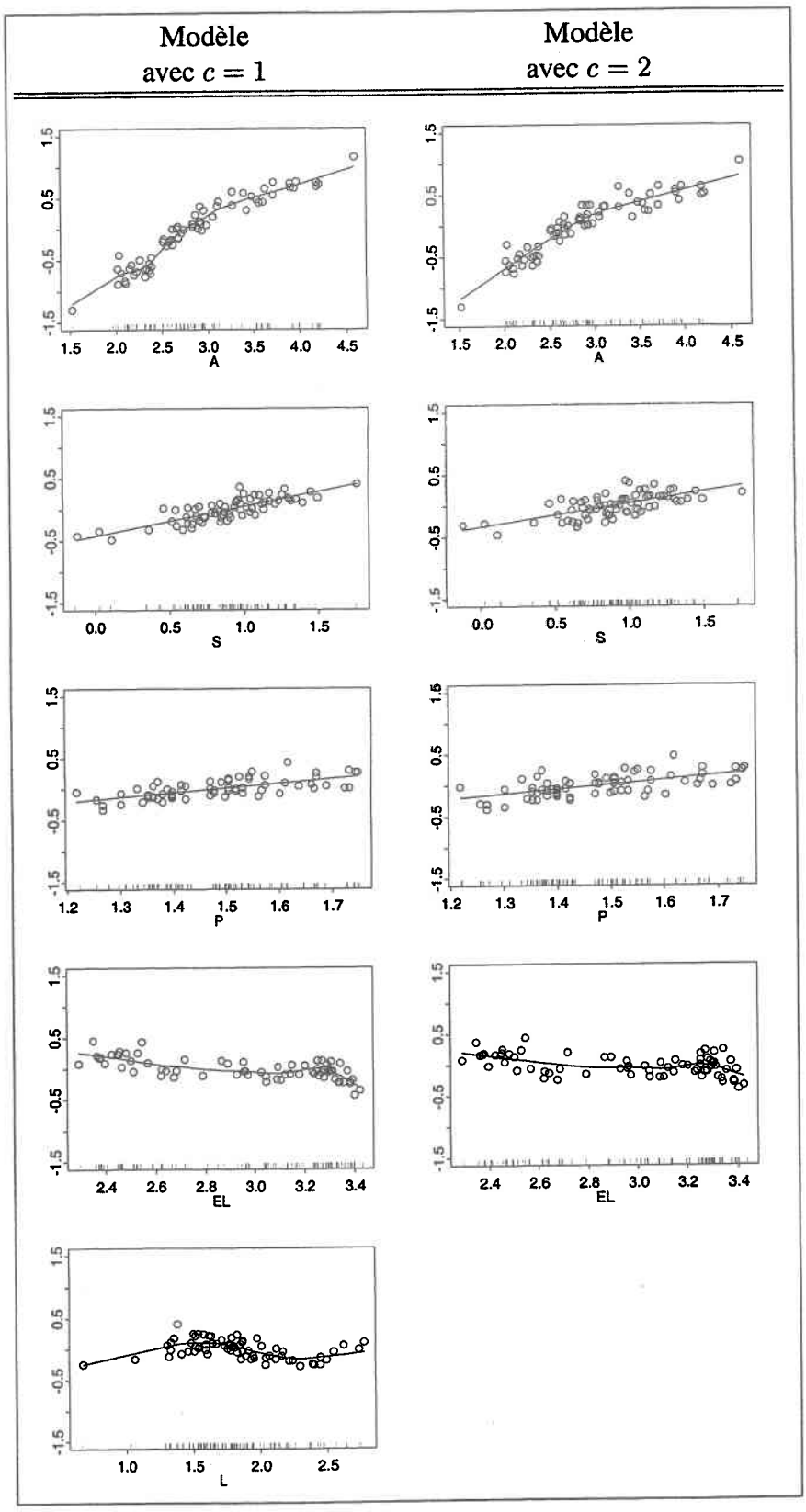

FIG. 5.4: Modélisation additive de Q_{50} : Effet du critère de validation-croisée sur le lissage

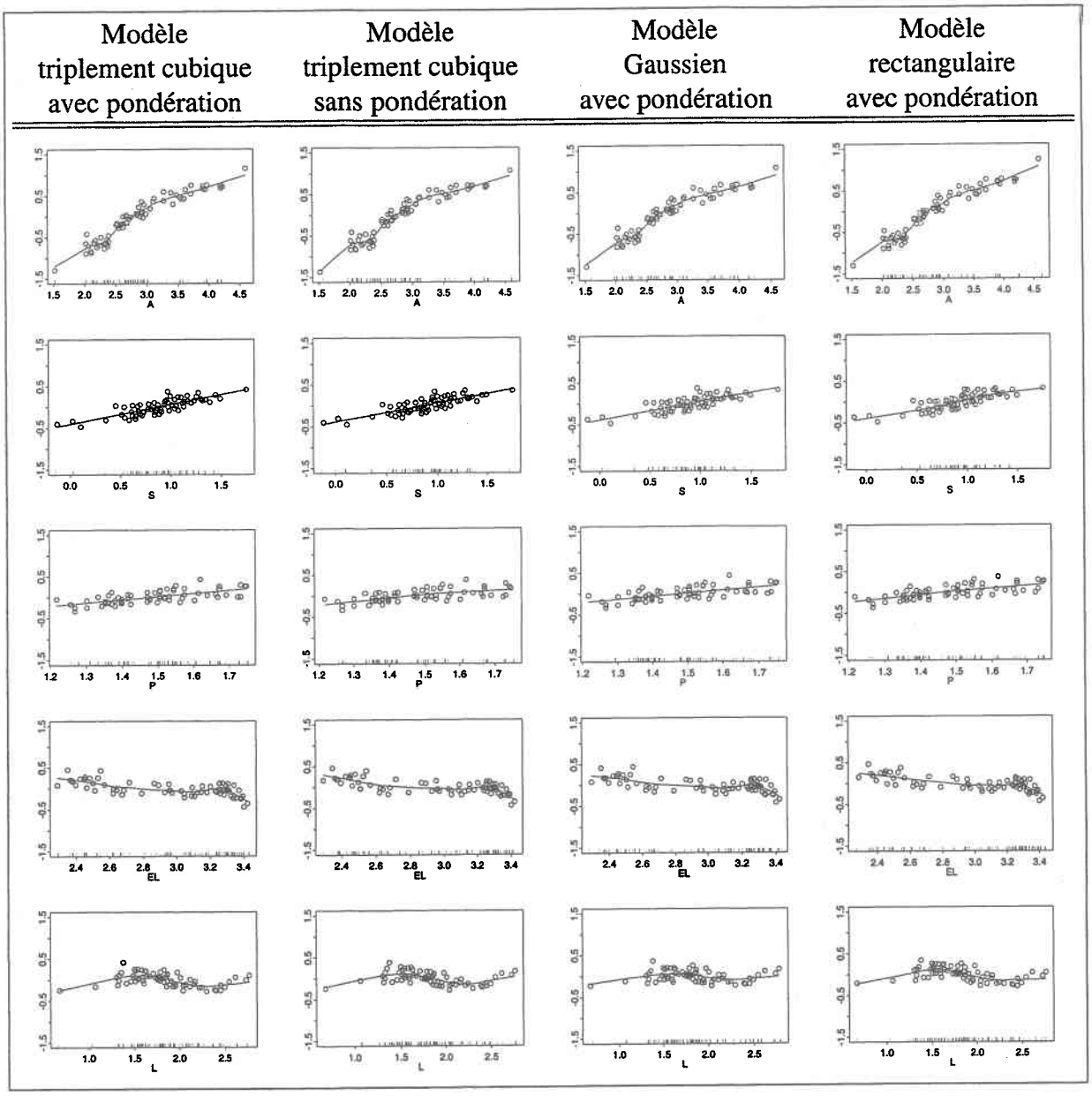

FIG. 5.5: Modélisation additive de Q_{50} : Effet de la fonction noyau et de l'utilisation d'une fonction de pondération sur le lissage ($c = 1$)

nécessite généralement moins de paramètres, l'utilisation de $c = 2$ plutôt que $c = 1$ favorise les modèles plus lisses. Pour montrer qu'en augmentant le niveau de lissage d'une courbe on en diminue généralement le nombre de paramètres effectifs, rappelons qu'en augmentant le niveau de lissage pour une variable explicative, la quantité $\mathrm{tr}(\mathbf{S}_\lambda \mathbf{S}_\lambda^T)$ a tendance à diminuer (réf. section 2.3.4.4). Or, la quantité $\mathrm{tr}(\mathbf{S}_\lambda \mathbf{S}_\lambda^T)$ n'est autre chose que la définition (ν_1) du nombre de paramètres effectifs associé à cette variable explicative.

Les différents graphiques des figures 5.4 et 5.5 représentent l'effet, conditionnellement à la présence des autres variables explicatives dans le modèle, d'une variable explicative particulière sur la variable de réponse. Par exemple, le premier graphique en haut à gauche de la figure 5.4 illustre la contribution apportée par $\log(A)$ à $\log(Q_{50})$ et ce, pour différentes valeurs de $\log(A)$. Cette contribution, représentée par la fonction $S_{\lambda_1}(\log(A))$, correspond au premier terme de l'équation 5.6. On peut aussi constater que la calibration du modèle additif optimal a fait en sorte que les deuxième et troisième termes de l'équation 5.6 ont été remplacés par des termes linéaires $\beta_S \log(S)$ et $\beta_P \log(P)$ (un cas particulier de lissage unidimensionnel) puisqu'on a alors obtenu que $h_2^* = h_3^* = 1$. Remarquons que pour les graphiques correspondant à ces variables, la valeur des paramètres β_S et β_P est représentée par la pente des relations linéaires. Par analogie au modèle linéaire où plus le paramètre β_i est élevé, plus la variable explicative correspondante est significative, il est possible d'observer graphiquement la signification relative de chacune des variables explicatives du modèle additif. Il est par exemple possible de remarquer, à la figure 5.4, que la variable explicative L a été la variable exclue du modèle (à droite) probablement parce que des 5 fonctions (d'après les graphiques de gauche), la fonction $S_{\lambda_5}(\log(L))$ semble être la moins significative (elle s'apparente à une droite horizontale de pente nulle). À l'inverse, comme on devait d'ailleurs s'y attendre, l'aire du bassin versant (A) constitue la variable explicative la plus significative.

Puisque d'une part, nous avons obtenu les meilleurs résultats de prédiction avec le modèle calibré à l'aide d'une fonction de pénalité $c = 1$ en utilisant une fonction noyau triplement cubique à l'intérieur de la fenêtre de lissage et en appliquant une fonction de pondération à chacun des sites jaugés, nous avons retenu l'emploi de ces paramètres pour la modélisation additive des quantiles de crue. Cependant, au niveau descriptif, il semble que le niveau de lissage obtenu par l'emploi de $c = 2$ plutôt que $c = 1$ soit plus approprié. En effet, si nous voulions interpréter les résultats de la figure 5.4, il nous serait difficile d'expliquer, pour $c = 1$, la forme irrégulière de la relation pour l'aire du bassin (A) dans l'intervalle $[2, 2.5]$. Il nous serait aussi difficile de justifier l'influence de la longueur du cours d'eau sur la valeur du quantile. Nous avons donc aussi retenu l'emploi du paramètre $c = 2$ pour les différentes modélisations additives de cette thèse. Les tableaux 5.11 et 5.12 présentent respectivement, pour $c = 1$ et $c = 2$, les résultats de l'application de ces approches de modélisation pour l'estimation des différents quantiles étudiés alors qu'aux figures 5.6 et 5.7, nous présentons les lissages obtenus par ces modèles pour l'estimation des quantiles Q_2, Q_5, Q_{10} et Q_{25}.

TAB. 5.11: RMSE et DRM des modèles additifs de polynômes locaux : c=1

Quantile modélisé	Paramètres optimaux h*					RMSE ($\log_{10}$)		DRM (%)	
	A	S	P	EL	L	EST	PRED	EST	PRED
Q_2	1	1	0.5	0.5	0.9	0,139	0,165	26,2	30,9
Q_5	1	0.6	0.7	0.3	0.7	0,111	0,146	20,4	26,7
Q_{10}	1	0.4	0.8	0.3	0.6	0,104	0,144	18,4	25,2
Q_{25}	0.4	1	0.9	0.3	0.6	0,102	0,143	18,6	26,5
Q_{50}	0.4	1	1	0.4	0.5	0,115	0,160	21,8	30,8

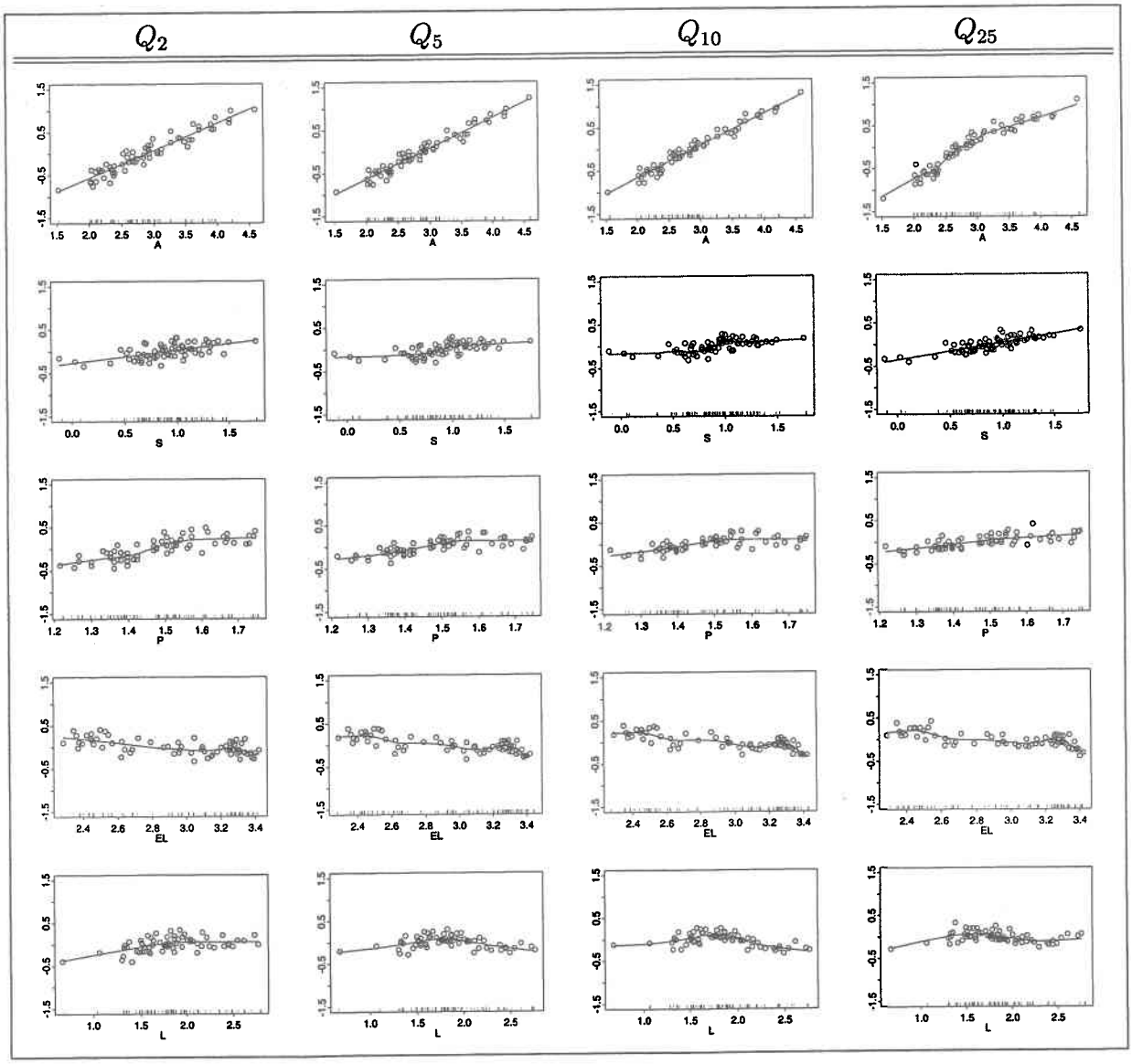

FIG. 5.6: Modélisation additive de Q_2, Q_5, Q_{10} et Q_{25} : $c = 1$

TAB. 5.12: RMSE et DRM des modèles additifs de polynômes locaux : c=2

Quantile modélisé	Paramètres optimaux h*					RMSE ($\log_{10}$)		DRM (%)	
	A	S	P	EL	L	EST	PRED	EST	PRED
Q_2	0.9	-	0.9	-	-	0,160	0,170	29,9	30,9
Q_5	1	1	1	0.5	0.8	0,131	0,157	24,6	28,4
Q_{10}	0.8	1	1	0.5	-	0,128	0,147	23,9	26,8
Q_{25}	0.8	1	1	0.5	-	0,134	0,153	25,9	28,6
Q_{50}	0.8	1	1	0.5	-	0,147	0,168	29,1	32,7

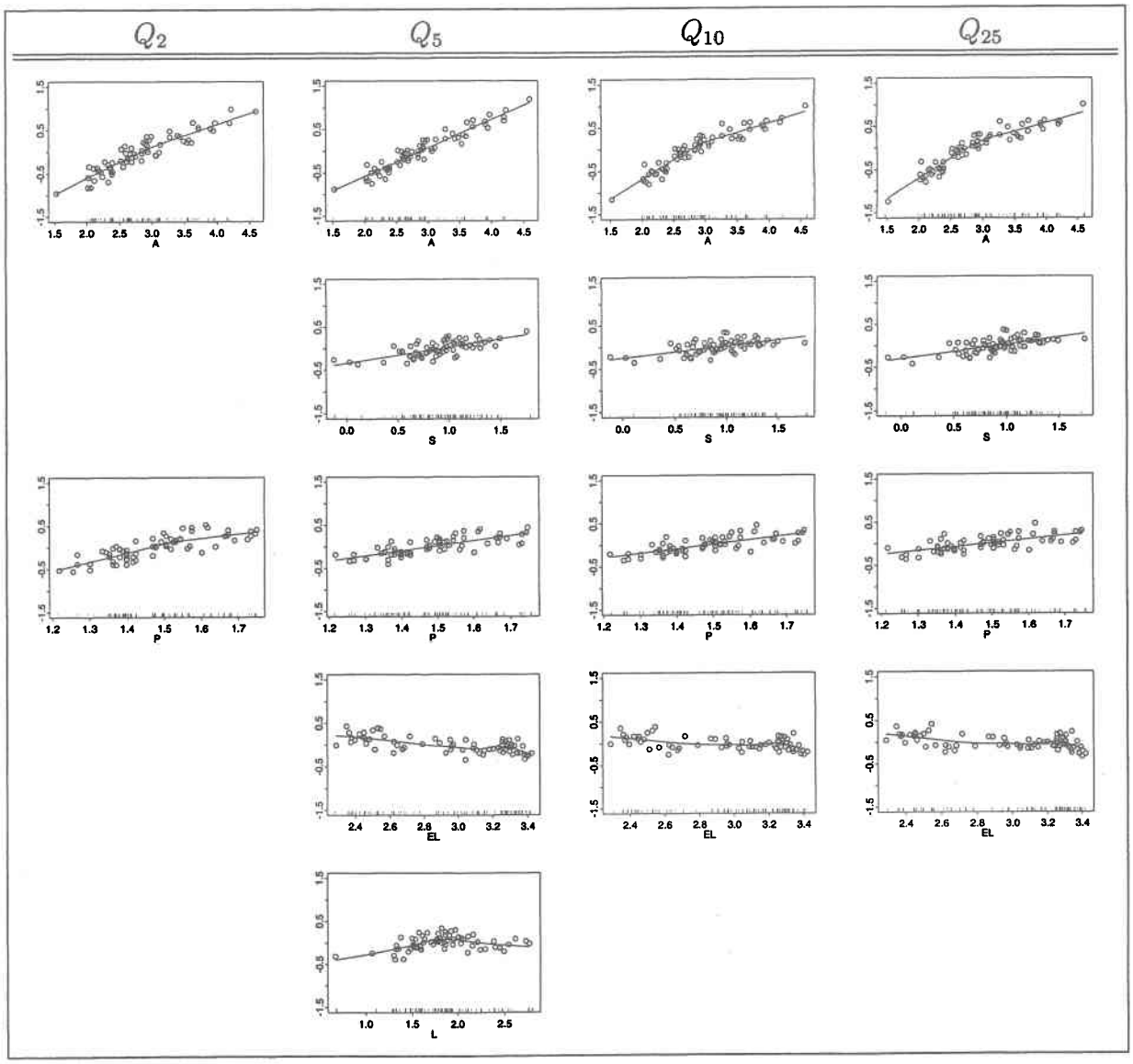

FIG. 5.7: Modélisation additive de Q_2, Q_5, Q_{10} et Q_{25} : $c = 2$

5.3.4 Synthèse et comparaison des résultats

Aux sections précédentes, nous avons procédé, pour la région du Texas, à l'estimation régionale des quantiles de crue Q_2, Q_5, Q_{10}, Q_{25} et Q_{50} à l'aide, respectivement,

1. du modèle de régression log-linéaire (ML) de Benson (1962),

2. du modèle de régression par région d'influence (RI) de Tasker et al. (1996), et

3. du modèle additif de polynômes locaux (MA) proposé dans cette thèse.

Nous avons utilisé, pour le modèle de régression log-linéaire, noté (ML), une approche de calibration permettant de déterminer de manière optimale, au sens du critère de RMSE prédictif de l'équation (5.2), les variables explicatives du ML. Pour le modèle de régression par région d'influence, noté (RI), nous avons développé une approche de calibration permettant (1) de choisir les variables explicatives à inclure dans le modèle et (2) de déterminer de manière optimale (au sens du RMSE) la taille de la région d'influence du modèle RI. Nous avons enfin procédé à une brève étude comparative des différents paramètres de calibration du modèle additif. À partir des résultats de cette étude, nous avons choisi de retenir les modèles additifs, notés MA1 et MA2, calibrés respectivement à l'aide des critères de validation-croisée $\mathrm{VCG}^*(c = 1, \Lambda)$ et $\mathrm{VCG}^*(c = 2, \Lambda)$. Mentionnons de plus qu'une fonction noyau triplement cubique de même qu'une fonction de pondération proportionnelle au nombre d'observations de DMA des divers sites jaugés ont été utilisées pour la calibration de MA1 et de MA2.

Les tableaux 5.13 et 5.14 présentent respectivement les valeurs de RMSE et de DRM obtenues par les différentes approches pour les diverses période de retour étudiées. Ces tableaux indiquent aussi l'écart entre les résultats obtenus par le meilleur modèle additif MA et les résultats obtenus par l'approche de la régression par région d'influence RI. On peut constater que l'approche du MA procure des meilleurs résultats que les autres approches sauf pour le critère de DRM pour la modélisation de Q_2. On remarque aussi que l'écart entre MA et RI augmente avec la période de retour T. Pour $T = 25$ et $T = 50$, on note des diminutions du RMSE de l'ordre d'environ 12%. Ceci peut s'expliquer par le fait que la relation entre le logarithme de la superficie du bassin versant, la variable la plus significative des différents modèles, et le logarithme des quantiles est de moins en moins linéaire à mesure que la période de retour T augmente (réf. graphiques du haut des figures 5.4 et 5.6).

TAB. 5.13: Comparaison des erreurs de prédiction : RMSE - Texas

Quantile modélisé	Modèle				Modèle optimal	Δ(MA,RI)=100(MA/RI-1) (%)
	ML	RI	MA1	MA2		
Q_2	0,178	0,166	0,165	0,170	MA1	-0,6
Q_5	0,173	0,150	0,146	0,157	MA1	-2,7
Q_{10}	0,179	0,149	0,144	0,147	MA1	-3,2
Q_{25}	0,197	0,162	0,143	0,153	MA1	-11,6
Q_{50}	0,216	0,182	0,160	0,168	MA1	-11,7

TAB. 5.14: Comparaison des erreurs de prédiction : DRM - Texas

Quantile modélisé	Modèle				Modèle optimal	Δ(MA,RI)=MA-RI (%)
	ML	RI	MA1	MA2		
Q_2	33,1	29,4	30,9	30,9	RI	1,5
Q_5	33,7	27,0	26,7	28,4	MA1	-0,3
Q_{10}	36,1	27,5	25,2	26,8	MA1	-2,4
Q_{25}	42,2	31,5	26,5	28,6	MA1	-4,9
Q_{50}	47,7	35,9	30,8	32,7	MA1	-5,2

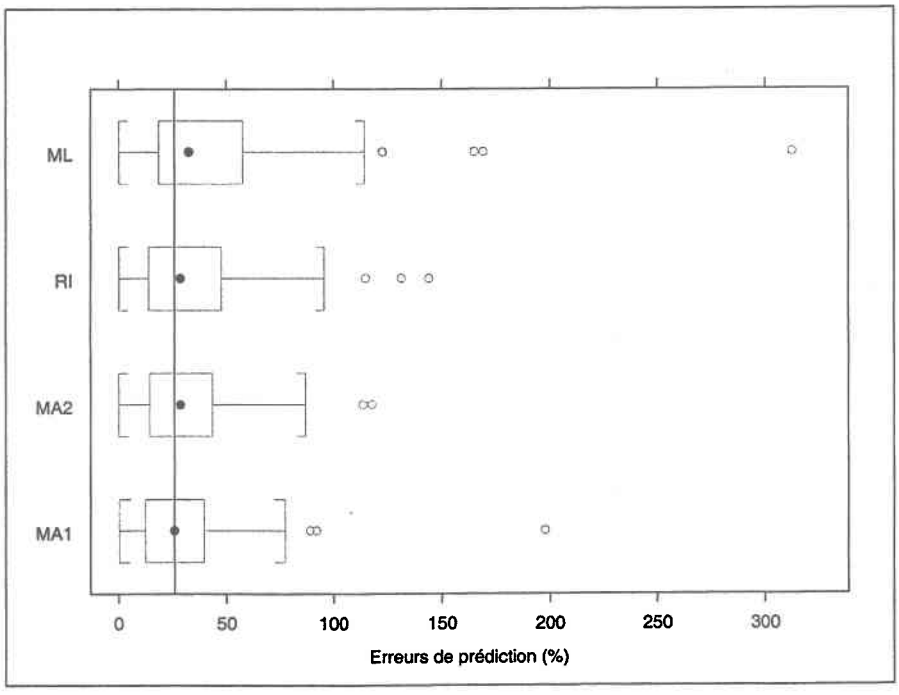

FIG. 5.8: Modélisation de Q_{50} : graphique en boîte des erreurs (relatives) de prédiction (%)

Les critères de RMSE et de DRM constituent des mesures des erreurs de prédiction **moyennes**. Il est cependant intéressant d'avoir en notre possession une représentation de la dispersion de ces erreurs. Le graphique en boîte de la figure 5.8 présente les différentes erreurs de prédiction de Q_{50} dont la moyenne est le critère de DRM. On remarque, sur la figure 5.8, la présence de deux erreurs très élevées de l'ordre d'environ 300% pour ML et de 200% pour MA1. Un examen des erreurs de prédiction nous a permis de constater que ces erreurs proviennent du bassin versant dont la superficie (A) est la plus petite. En examinant les graphiques du haut de la figure 5.4, il est possible d'évaluer la difficulté d'effectuer l'extrapolation permettant l'estimation de Q_{50} pour la plus petite valeur de $\log(A) \approx 1,5$. Puisque d'une part, l'objectif principal de l'estimation régionale n'est généralement pas de procéder à de l'extrapolation et que, d'autre part, certaines grandes erreurs d'extrapolation sont susceptible d'affecter à tort la performance d'une approche particulière, nous avons recalculé, aux tableaux 5.15 et 5.16, les critères de RMSE et de DRM en enlevant les erreurs d'extrapolation. On constate alors une augmentation de l'écart entre MA et RI.

TAB. 5.15: Comparaison des erreurs de prédiction (sans extrapolation) : RMSE - Texas

Quantile modélisé	Modèle				Modèle optimal	Δ(MA,RI)=100(MA/RI-1) (%)
	ML	RI	MA1	MA2		
Q_2	0,185	0,172	0,166	0,180	MA1	-3,4
Q_5	0,179	0,156	0,145	0,150	MA1	-7,2
Q_{10}	0,182	0,154	0,142	0,146	MA1	-7,4
Q_{25}	0,195	0,162	0,139	0,151	MA1	-14,1
Q_{50}	0,210	0,176	0,148	0,164	MA1	-15,8

TAB. 5.16: Comparaison des erreurs de prédiction (sans extrapolation) : DRM - Texas

Quantile modélisé	Modèle				Modèle optimal	Δ(MA,RI)=MA-RI (%)
	ML	RI	MA1	MA2		
Q_2	34,5	30,4	30,7	33,5	RI	0,3
Q_5	34,9	28,2	26,1	27,1	MA1	-2,1
Q_{10}	36,6	28,2	24,7	26,0	MA1	-3,5
Q_{25}	41,0	30,9	25,3	27,6	MA1	-5,6
Q_{50}	45,3	34,1	28,2	31,7	MA1	-5,9

5.4 Application à la région de la Nouvelle-Angleterre

Dans cette section, nous appliquons les différentes approches de modélisation ML, RI, MA1 et MA2 aux données de la région de la Nouvelle-Angleterre qui correspond à la région 01 du USGS (cf. tab. 5.1). Puisqu'à la section précédente, nous avons présenté en détail les approches de calibration des différents modèles, nous ne présentons ici que les principaux résultats de la calibration et du calcul des critères de comparaison. La région 01 du USGS comporte 71 sites et un seul site a du être omis en raison de variables explicatives manquantes. Les données physiographiques et hydrologiques, c'est-à-dire les estimateurs locaux par GEV/PWM des différents quantiles étudiés, sont présentés à l'annexe C (cf. tab. C.1). Mentionnons de plus que nous avons retenu l'emploi des mêmes variables explicatives qu'à la section précédente.

5.4.1 Le modèle de régression log-linéaire

Après avoir calculé le critère de RMSE des différents modèles potentiels du tableau 5.3, nous avons été en mesure de détecter le modèle pour lequel le RMSE était optimal. Les graphiques du RMSE des différents modèles de régression log-linéaire correspondant aux périodes de retour (T=2,5,10,25 et 50) se retrouvent respectivement aux figures C.1,C.2,C.3,C.4 et C.5 de l'annexe C. Le tableau 5.17 présente les estimateurs des paramètres des différents modèles alors que le tableau 5.18 présente les critères de RMSE et de DRM calculés pour ces modèles.

TAB. 5.17: Les paramètres estimés des équations de régression

T	$\hat{\beta}_0$	$\hat{\beta}_A$	$\hat{\beta}_S$	$\hat{\beta}_P$	$\hat{\beta}_{EL}$	$\hat{\beta}_L$
2	0,409	0,907	0,385	-	-	0,367
5	-0,950	0,884	0,365	0,959	-	0,384
10	-1,209	0,861	0,371	1,175	-	0,412
25	-1,637	0,828	0,375	1,504	-	0,453
50	-2,010	0,802	0,378	1,776	-	0,487

TAB. 5.18: RMSE et DRM des modèles de régression log-linéaire

Quantile modélisé	RMSE ($\log_{10}$)		DRM (%)	
	Estimation	Prédiction	Estimation	Prédiction
Q_2	0,111	0,119	20,5	22,0
Q_5	0,111	0,120	20,3	22,2
Q_{10}	0,113	0,123	21,4	23,3
Q_{25}	0,124	0,134	24,0	26,1
Q_{50}	0,138	0,148	27,0	29,4

5.4.2 Le modèle de régression par région d'influence

Après avoir calculé le critère de RMSE par validation-croisée pour les différentes valeurs possibles du paramètre k (le nombre de stations dans la région d'influence) et ce, pour les meilleurs modèles comprenant de une à cinq variables explicatives, nous avons été en mesure de déterminer les paramètres des modèles optimaux. Les graphiques du critère de RMSE en fonction du nombre de stations k sont présentés aux figures C.6 (T=2), C.7 (T=5), C.8 (T=10), C.9 (T=25) et C.10 (T=50) de l'annexe C. Le tableau 5.19 décrit les différents modèles calibrés alors que le tableau 5.20 présente l'évaluation de leurs capacités descriptive et prédictive en terme de RMSE et de DRM. Remarquons que pour les périodes de retour T de 2,5,10 et 25 années, le nombre optimal de stations k correspond au nombre maximal de stations pouvant être incluses dans la région d'influence. Tous les sites, à part le site cible, sont donc utilisés pour la calibration du modèle prédictif. La seule distinction entre ces modèles et les modèles de régression log-linéaire résident donc dans l'utilisation d'une fonction noyau triplement cubique pour l'approche de régression par région d'influence.

TAB. 5.19: Les modèles calibrés de régression par région d'influence

T	Variables explicatives	Nombre optimal de stations k
2	A,S,L	26
5	A,P,S,L	69
10	A,P,S,L	69
25	A,P,S,L	69
50	A,P,S,L	69

TAB. 5.20: RMSE et DRM des modèles de régression par région d'influence

Quantile modélisé	RMSE ($\log_{10}$)		DRM (%)	
	Estimation	Prédiction	Estimation	Prédiction
Q_2	0,086	0,115	15,3	22,2
Q_5	0,106	0,118	19,4	21,8
Q_{10}	0,109	0,122	20,3	22,8
Q_{25}	0,119	0,133	22,6	25,3
Q_{50}	0,133	0,147	25,6	28,7

5.4.3 Le modèle additif par polynômes locaux

Nous avons, dans un premier temps, tenté d'effectuer la calibration des différents modèles pour les données de la région 01 à l'aide de la procédure **locfit**. Nous avons alors rencontré des difficultés au niveau de la convergence de l'algorithme de *backfitting*. Le nombre d'itérations par défaut (25) du logiciel S-Plus ne permettait pas l'atteinte de la convergence des différents modèles. Nous avons donc augmenté ce nombre d'itérations à 50 ce qui a permis la convergence de l'algorithme mais en augmentant considérablement le temps de calcul. En pratique, ces problèmes de convergence sont généralement causés par une trop forte corrélation entre les différentes variables prédictives, comme l'indiquent d'ailleurs Hastie et Tibshirani (1990) :

> "When a smoothing-spline or running-line smoother is used for several predictors, practical experience has shown that if the predictors are correlated, many iterations may be required to get the correct average slope of the functions"

Or, ce problème de corrélation est bien connu en régression log-linéaire des quantiles de crue et Roy et al. (1989) recommandent d'ailleurs l'utilisation de la régression ridge (Hoerl et Kennard, 1970) afin d'estimer de manière plus efficace les paramètres du modèle de régression des quantiles. Le tableau 5.21 illustre les résultats des corrélations obtenues par la procédure **lm** de S-Plus pour les régions du Texas et de la Nouvelle-Angleterre. On constate la présence de corrélations élevées (-0,91 et -0,73) pour la Nouvelle-Angleterre, comparativement à (-0,81 et -0,51) au Texas, ce qui pourrait expliquer les problèmes de vitesse de convergence rencontrés.

TAB. 5.21: Matrices de corrélation des paramètres des modèles de régression log-linéaire

	Texas					Nouvelle-Angleterre				
	β_0	β_A	β_S	β_P	β_{EL}	β_0	β_A	β_S	β_P	β_{EL}
β_A	-0.20					0.04				
β_S	-0.14	-0.03				0.32	-0.14			
β_P	-0.96	0.25	0.06			-0.98	0.04	-0.34		
β_{EL}	-0.71	0.09	-0.51	0.67		-0.10	0.01	-0.73	0.01	
β_L	0.09	-0.81	0.50	-0.21	-0.36	-0.03	-0.91	0.45	-0.05	-0.29

Alors qu'en régression paramétrique, la régression ridge permet de faire face au problème de corrélation entre les variables prédictives (pour plus de détails, voir Roy et al. (1989)), en modélisation additive, Buja et al. (1989) ont développé un algorithme amélioré de *backfitting* (à ne pas confondre avec l'algorithme modifié de Hastie et Tibshirani (1990) de la figure 3.3) permettant de s'attaquer à ce problème de vitesse de convergence. Cet algorithme consiste à séparer chacun des lisseurs en une partie linéaire paramétrique et en une partie non linéaire.

Les termes paramétriques sont ensuite regroupés afin de procéder à leur estimation en une seule étape. L'algorithme régulier de *backfitting* est quant à lui utilisé pour l'estimation de la partie non linéaire des différents lisseurs. Cet algorithme est utilisé par S-Plus lorsque le lisseur employé est le **loess**. Puisque les procédures **loess** et **locfit** sont équivalentes lorsque la fonction noyau utilisée est la fonction triplement cubique, nous avons décidé d'opter pour la procédure **loess** pour la calibration des différents modèles. Ce changement de procédure nous a permis de faire passer le temps moyen de calcul d'un modèle additif de 7,5 secondes à 0,6 seconde (sur un ordinateur IBM Pentium II - 350 Mhz) et donc 12,5 fois plus rapide que la procédure **locfit**.

Les tableaux 5.22 et 5.23 présentent respectivement les résultats de la calibration des modèles MA1 et MA2 ainsi que les critères RMSE et DRM associés à ces modèles. La figure 5.9 illustre les lissages obtenus par MA1 et MA2 pour la modélisation de Q_{50}. Les lissages des quantiles Q_2, Q_5, Q_{10} et Q_{25} se retrouvent respectivement aux figures C.11, C.12, C.13 et C.14 de l'annexe C. Remarquons que le modèle MA2 a convergé vers une relation linéaire pour la variable A alors que pour MA1, bien que $h_A = 0.6$, on observe toutefois la présence d'une relation assez linéaire. Mentionnons enfin que l'approche MA2 a convergé, pour Q_{10} et Q_{25}, vers le même modèle qu'avec l'approche RI alors que pour Q_{50}, seul le lisseur de la variable S n'a pas convergé vers un terme paramétrique.

TAB. 5.22: RMSE et DRM des modèles additifs de polynômes locaux : c=1

Quantile modélisé	Paramètres optimaux h*					RMSE ($\log_{10}$)		DRM (%)	
	A	S	P	EL	L	EST	PRED	EST	PRED
Q_2	0.5	0.9	-	0.9	0.6	0,095	0,118	17,3	21,9
Q_5	0.5	0.9	0.5	-	0.6	0,096	0,126	17,5	22,8
Q_{10}	0.5	0.9	0.5	-	0.6	0,099	0,130	19,0	24,6
Q_{25}	0.6	0.8	0.5	-	0.6	0,111	0,143	21,8	27,6
Q_{50}	0.6	0.9	0.4	0.9	0.9	0,123	0,158	23,9	30,0

TAB. 5.23: RMSE et DRM des modèles additifs de polynômes locaux : c=2

Quantile modélisé	Paramètres optimaux h*					RMSE ($\log_{10}$)		DRM (%)	
	A	S	P	EL	L	EST	PRED	EST	PRED
Q_2	1	1	-	0.9	0.9	0,103	0,115	18,7	21,7
Q_5	1	1	-	1	-	0,113	0,121	21,0	22,5
Q_{10}	1	1	1	-	1	0,113	0,123	21,4	23,3
Q_{25}	1	1	1	-	1	0,124	0,134	24,0	26,1
Q_{50}	1	0.9	1	-	1	0,134	0,146	26,1	28,3

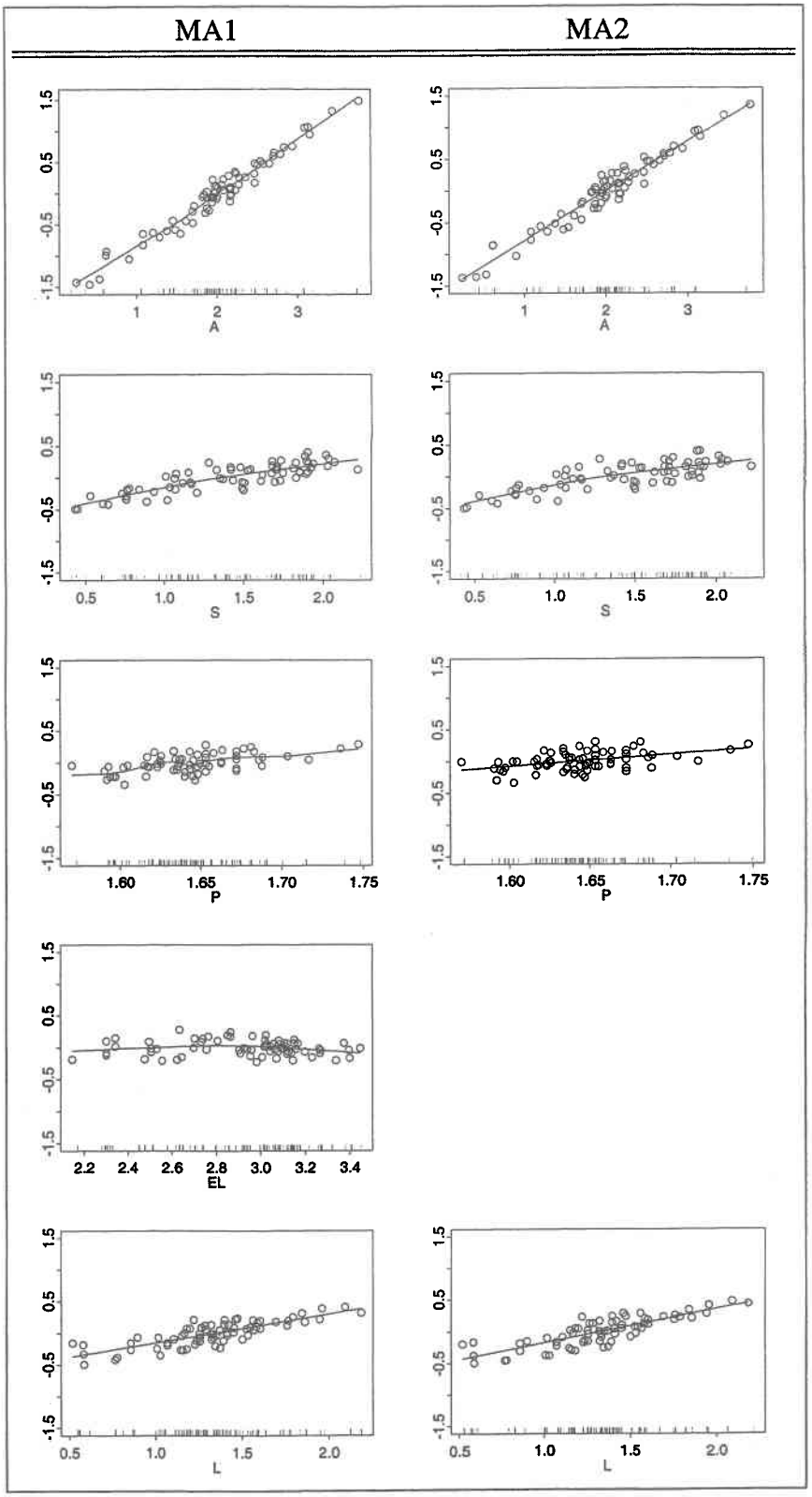

FIG. 5.9: Modélisation additive de Q_{50}

5.4.4 Comparaison des résultats

Les tableaux 5.24 et 5.25 présentent respectivement les valeurs de RMSE et de DRM obtenues par les différentes approches pour les diverses période de retour étudiées. Ces tableaux indiquent aussi l'écart entre les résultats obtenus par le meilleur modèle additif MA et le modèle de régression par région d'influence RI. On constate que les différentes approches procurent sensiblement les mêmes résultats avec au plus un écart entre MA et RI de 2%. Ces faibles différences étaient prévisibles à partir du moment où la calibration par les différentes approches a convergé vers des modèles similaires.

TAB. 5.24: Comparaison des erreurs de prédiction : RMSE - Nouvelle-Angleterre

Quantile modélisé	Modèle				Modèle optimal	Δ(MA,RI)=100(MA/RI-1) (%)
	ML	RI	MA1	MA2		
Q_2	0,119	0,115	0,118	0,115	RI	0,5
Q_5	0,120	0,118	0,126	0,121	RI	2,0
Q_{10}	0,123	0,122	0,130	0,123	RI	1,0
Q_{25}	0,134	0,133	0,143	0,134	RI	0,5
Q_{50}	0,148	0,147	0,158	0,146	MA2	-1,3

TAB. 5.25: Comparaison des erreurs de prédiction : DRM - Nouvelle-Angleterre

Quantile modélisé	Modèle				Modèle optimal	Δ(MA,RI)=MA-RI (%)
	ML	RI	MA1	MA2		
Q_2	22,0	22,2	21,9	21,7	MA2	-0,5
Q_5	22,2	21,8	22,8	22,5	RI	0,7
Q_{10}	23,3	22,8	24,6	23,3	RI	0,5
Q_{25}	26,1	25,3	27,6	26,1	RI	0,8
Q_{50}	29,4	28,7	30,0	28,3	MA2	-0,4

5.5 Application à la région de l'Arkansas

Dans cette section, nous appliquons les différentes approches de modélisation aux données de l'Arkansas. Il s'agit du même ensemble de données que celui utilisé par Tasker et al. (1996) pour la modélisation du quantile Q_{50} par la méthode de la régression par région d'influence. Les données hydrologiques et physiographiques des 204 stations utilisées en Arkansas se retrouvent au tableau D.1 de l'annexe D. Les données de quantiles utilisées proviennent du USGS (Hodge

et Tasker, 1995) et ont été estimés par l'ajustement d'une loi Log-Pearson type III à l'aide de la procédure recommandée par le Hydrology Subcommittee of the Interagency Advisory Committee on Water Data (1982). Mentionnons aussi que les variables explicatives disponibles sont les mêmes que précédemment (A, S, P, EL et L) mais que pour la modélisation, Tasker et al. (1996) ont utilisé la variable **facteur de forme du bassin** (*Basin Shape Factor*) (SH) calculée en divisant la superficie du bassin versant (A) par le carré de la longueur du cours d'eau principal (L) plutôt que la variable L. De manière à reproduire leurs résultats, nous utiliserons donc comme variables prédictives potentielles : A, S, P, EL et SH.

5.5.1 Le modèle de régression log-linéaire

Après avoir calculé le critère de RMSE des différents modèles potentiels du tableau 5.3, nous avons été en mesure de détecter le modèle pour lequel le RMSE était optimal. Les graphiques du RMSE des différents modèles de régression log-linéaire correspondant aux périodes de retour (T=2,5,10,25 et 50) se retrouvent respectivement aux figures D.1,D.2,D.3,D.4 et D.5 de l'annexe D. Le tableau 5.26 présente les estimateurs des paramètres des différents modèles alors que le tableau 5.27 présente les critères de RMSE et de DRM calculés pour ces modèles.

TAB. 5.26: Les paramètres estimés des équations de régression

T	$\hat{\beta}_0$	$\hat{\beta}_A$	$\hat{\beta}_S$	$\hat{\beta}_P$	$\hat{\beta}_{EL}$	$\hat{\beta}_{SH}$
2	-5,199	0,724	0,155	2,162	0,484	0,127
5	-5,173	0,739	0,179	2,210	0,553	0,217
10	-4,874	0,745	0,192	2,115	0,570	0,258
25	-4,588	0,752	0,203	2,015	0,599	0,305
50	-4,229	0,756	0,215	1,884	0,602	0,338

TAB. 5.27: RMSE et DRM des modèles de régression log-linéaire

Quantile modélisé	RMSE ($\log_{10}$)		DRM (%)	
	Estimation	Prédiction	Estimation	Prédiction
Q_2	0,226	0,234	48,6	51,1
Q_5	0,200	0,209	40,9	43,3
Q_{10}	0,198	0,207	40,0	42,6
Q_{25}	0,203	0,213	40,7	43,5
Q_{50}	0,209	0,219	42,0	45,1

5.5.2 Le modèle de régression par région d'influence

Après avoir calculé le critère de RMSE par validation-croisée pour les différentes valeurs possibles du paramètre k, nous avons été en mesure de déterminer les paramètres des modèles optimaux. Les graphiques du critère de RMSE en fonction du nombre de stations k sont présentés aux figures D.6 (T=2), D.7 (T=5), D.8 (T=10), D.9 (T=25) et D.10 (T=50) de l'annexe D. Le tableau 5.28 décrit les différents modèles calibrés alors que le tableau 5.29 présente l'évaluation de leurs capacités descriptive et prédictive en terme de RMSE et de DRM. On constate, au tableau 5.28, que comme on devait s'y attendre, plus le modèle optimal de régression par région d'influence comporte de variables explicatives (plus de paramètres à estimer) plus la taille optimale de la région d'influence augmente (plus de données sont nécessaires pour estimer les paramètres). Remarquons enfin que la taille optimale k ne représente environ que 10% à 15% du nombre total d'observations.

TAB. 5.28: Les modèles calibrés de régression par région d'influence

T	Variables explicatives	Nombre optimal de stations k
2	A,S,P,EL	22
5	A,S,P,EL,SH	26
10	A,S,EL	19
25	A,S,P,EL,SH	31
50	A,S,P,EL,SH	28

TAB. 5.29: RMSE et DRM des modèles de régression par région d'influence

Quantile modélisé	RMSE ($\log_{10}$)		DRM (%)	
	Estimation	Prédiction	Estimation	Prédiction
Q_2	0,100	0,190	17,4	35,6
Q_5	0,080	0,176	13,8	31,9
Q_{10}	0,107	0,160	19,1	30,8
Q_{25}	0,094	0,178	16,3	31,3
Q_{50}	0,092	0,185	15,8	33,2

5.5.3 Le modèle additif par polynômes locaux

La procédure de calibration pas-à-pas des modèles additifs demande, en entrée, un vecteur de paramètres initiaux $\mathbf{H}^{(0)}=(h_1^{(0)}, h_2^{(0)}, h_3^{(0)}, h_4^{(0)}, h_5^{(0)})$. Pour la modélisation de Q_T au Texas et en Nouvelle-Angleterre, nous avons débuté la procédure de sélection pas-à-pas avec des voisinages de 50% des observations pour chacune des variables explicatives du modèle soit $\mathbf{H}^{(0)}=(0.5, 0.5, 0.5, 0.5, 0.5)$. Avec les données de l'Arkansas, nous avons observé que pour l'approche de modélisation MA2, il était préférable de choisir $\mathbf{H}^{(0)}=(0.9, 0.9, 0.9, 0.9, 0.9)$ comme vecteur de paramètres initiaux. En effet, lorsque $c = 2$, le nombre de paramètres effectifs du modèle est davantage pénalisé, il est alors probable qu'une ou plusieurs variables explicatives soient (1) représentées par un terme linéaire ou encore (2) exclues du modèles. Or, en débutant avec $h_i^{(0)} = 0.9$, les modèles comportant des termes linéaires sont évalués dès le départ ce qui d'une part, permet d'augmenter la vitesse d'exécution de la procédure de sélection pas-à-pas, et d'autre part, empêche qu'un optimum local soit atteint avant d'avoir pu évaluer les modèles comportant des termes linéaires. Pour l'approche de modélisation MA1, l'utilisation de $h_i^{(0)} = 0.5$ a toujours procuré les meilleurs résultats.

Les tableaux 5.30 et 5.31 présentent respectivement les résultats de la calibration des modèles MA1 et MA2 ainsi que les critères de RMSE et de DRM associés à ces modèles. La figure 5.10 illustre les lissages obtenus par MA2 pour la modélisation de Q_{10}, Q_{25} et Q_{50}. Les lissages correspondant pour MA1 se retrouvent à la figure D.13 de l'annexe D alors que les lissages obtenus par MA1 et MA2 pour Q_2 et Q_5 sont présentés respectivement aux figures D.11 et D.12 de l'annexe D. On observe à la figure 5.10, et ce pour les différentes périodes de retour, la présence d'une relation non linéaire (non log-linéaire) pour la variable explicative S (la pente du bassin versant). Cette non-linéarité de la variable S pourrait être l'élément important expliquant la bonne performance, lors de l'étude de Tasker et al. (1996), de la méthode de la régression par région d'influence en Arkansas.

TAB. 5.30: RMSE et DRM des modèles additifs de polynômes locaux : c=1

Quantile modélisé	Paramètres optimaux h*					RMSE ($\log_{10}$)		DRM (%)	
	A	S	P	EL	SH	EST	PRED	EST	PRED
Q_2	0.2	0.7	0.3	0.5		0,166	0,189	31,9	36,3
Q_5	0.2	0.5	0.3	0.4		0,136	0,156	25,9	29,8
Q_{10}	0.9	0.6	0.3	0.5	0.6	0,140	0,158	26,8	29,7
Q_{25}	0.9	0.5	0.3	0.8	0.6	0,145	0,164	27,8	30,5
Q_{50}	0.9	0.5	0.3	0.8	0.5	0,152	0,172	29,0	32,1

TAB. 5.31: RMSE et DRM des modèles additifs de polynômes locaux : c=2

Quantile modélisé	Paramètres optimaux h*					RMSE ($\log_{10}$)		DRM (%)	
	A	S	P	EL	SH	EST	PRED	EST	PRED
Q_2	0.9	0.8	1	0.6		0,184	0,193	36,9	37,5
Q_5	0.9	0.7	1	0.6		0,150	0,159	29,7	30,4
Q_{10}	0.9	0.7	1	0.9		0,147	0,157	28,8	29,5
Q_{25}	0.8	0.6	1	0.9		0,150	0,161	29,1	29,8
Q_{50}	0.9	0.6	1	0.9		0,159	0,170	30,5	31,3

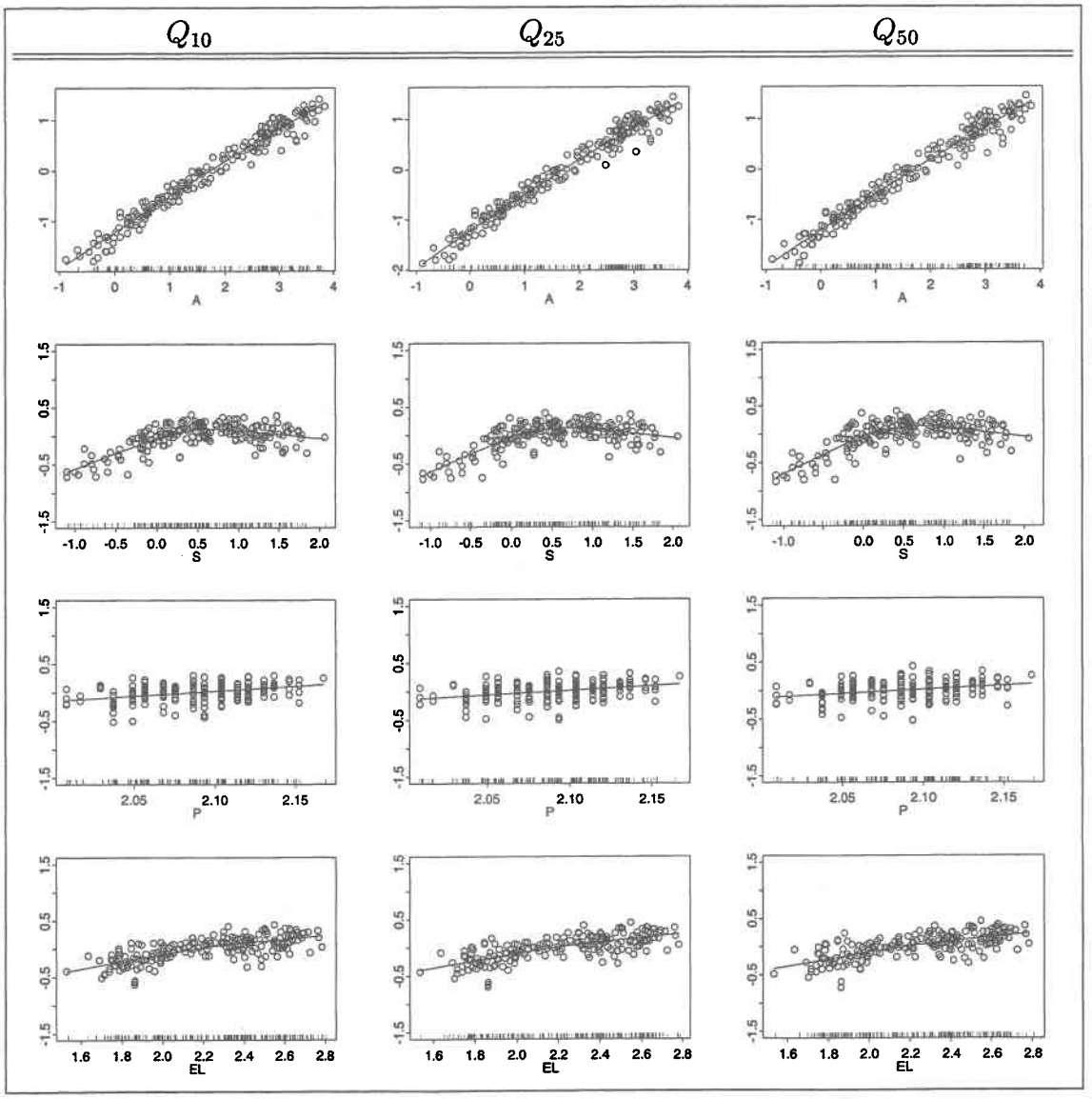

FIG. 5.10: Modélisation additive ($c = 2$) de Q_{10}, Q_{25} et Q_{50}

5.5.4 Comparaison des résultats

Les tableaux 5.32 et 5.33 présentent respectivement les résultats de RMSE et de DRM obtenus par les différentes approches pour les diverses périodes de retour étudiées. Ces tableaux indiquent aussi l'écart entre les résultats obtenus par le meilleur modèle additif MA et le modèle de régression par région d'influence RI. On peut constater que l'approche du MA conduit aux meilleurs résultats sauf pour le critère DRM pour la modélisation de Q_2 (cf. tab. 5.33). On remarque aussi que la modélisation de Q_{50} à l'aide de MA2 permet une diminution du RMSE d'environ 8% par rapport à l'approche de régression par région d'influence. Cependant, de manière à comparer les performances des approches RI, MA1 et MA2 aux approches utilisées par Tasker et al. (1996), nous avons complété à partir du tableau 5.33 la figure 3 de Tasker et al. (1996) afin d'y inclure les critères de DRM obtenus pour Q_{50}. La figure 5.11 illustre ces résultats. Rappelons que l'étude de Tasker et al. (1996) consistait à comparer diverses procédures de régression régionale selon la procédure de regroupement de sites en régions utilisée : les méthodes A,B et C consistaient à séparer respectivement les 204 stations de l'Arkansas en 1, 2 et 4 régions géographiques, la méthode D consistait à regrouper les sites en régions à l'aide d'une analyse de regroupement (*cluster analysis*) basée sur les caractéristiques physiographiques des bassins versants alors que la méthode E consistait plutôt à déterminer une région d'influence à chacun des sites cibles. Cette figure montre la bonne performance des méthodes présentées dans cette thèse par rapport aux approches considérées par Tasker et al. (1996).

TAB. 5.32: Comparaison des erreurs de prédiction : RMSE - Arkansas

Quantile modélisé	Modèle				Modèle optimal	Δ(MA,RI)=100(MA/RI-1) (%)
	ML	RI	MA1	MA2		
Q_2	0,234	0,190	0,189	0,193	MA1	-0,3
Q_5	0,209	0,176	0,156	0,159	MA1	-11,4
Q_{10}	0,207	0,160	0,158	0,157	MA2	-1,7
Q_{25}	0,213	0,178	0,164	0,161	MA2	-9,5
Q_{50}	0,219	0,185	0,172	0,170	MA2	-7,9

TAB. 5.33: Comparaison des erreurs de prédiction : DRM - Arkansas

Quantile modélisé	Modèle				Modèle optimal	Δ(MA,RI)=MA-RI (%)
	ML	RI	MA1	MA2		
Q_2	51,1	35,6	36,3	37,5	RI	0,8
Q_5	43,3	31,9	29,8	30,4	MA1	-2,1
Q_{10}	42,6	30,8	29,7	29,5	MA2	-1,3
Q_{25}	43,5	31,3	30,5	29,8	MA2	-1,4
Q_{50}	45,1	33,2	32,1	31,3	MA2	-1,9

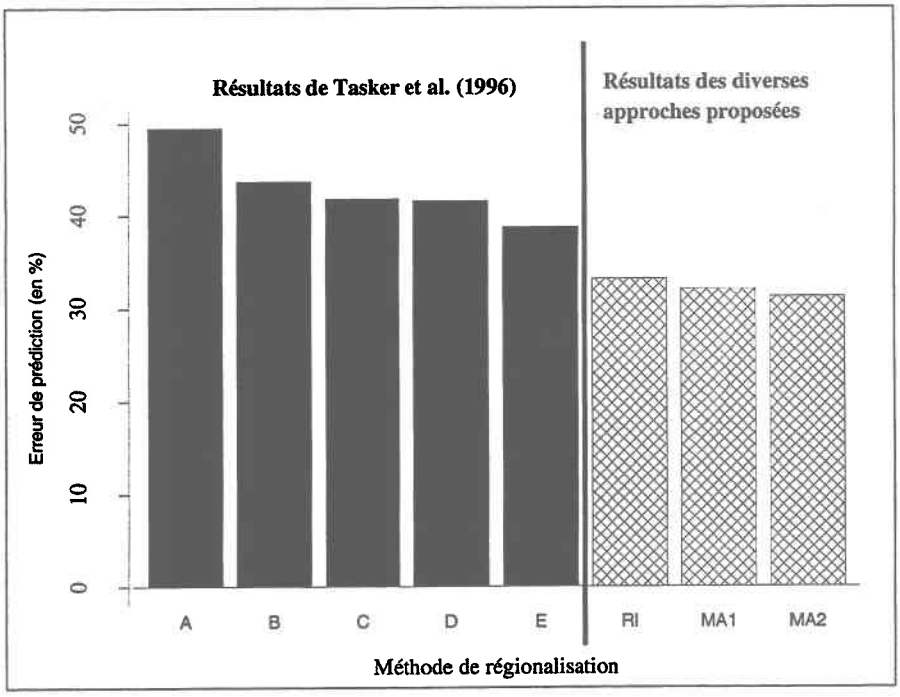

FIG. 5.11: Comparaison des erreurs de prédiction (DRM) en Arkansas (adapté de Tasker et al. (1996))

6. CONCLUSION

6.1 Motivation de l'étude

Pour la planification, incluant la conception, des ouvrages de contrôle des inondations ou de tout autre ouvrage soumis au risque de défaillance par les eaux, les hydrologues sont souvent amenés à produire des estimations de la magnitude Q_T de la crue ayant une période de retour T préalablement définie. En raison des grandes étendues territoriales et du coût associé à l'installation et au maintien de stations de mesures, il arrive fréquemment que l'estimation de Q_T soit requise en un site où l'on ne dispose d'aucune information hydrométrique (site non jaugé). Dans cette situation, il est alors possible d'utiliser des procédures dites de régionalisation pour transférer l'information disponible en des sites jaugés vers le site non jaugé où l'on désire produire une estimation (site cible).

De manière générale, les procédures de régionalisation comportent deux étapes distinctes qui consistent à (1) choisir les sites jaugés à partir desquels s'effectuera le transfert d'information et (2) appliquer aux sites choisis un modèle de transfert d'information régionale. Les hydrologues ont à leur disposition une abondance de méthodes leur permettant d'effectuer l'une ou l'autre de ces étapes. Aux États-Unis, Tasker et al. (1996), des hydrologues du USGS, l'organisme impliqué dans le développement des procédures de régionalisation pour chacun des États américains, ont proposé récemment l'emploi de la **méthode de la régression régionale par région d'influence**, une procédure dont le modèle de transfert utilisé est le modèle classique de régression log-linéaire des quantiles (Benson, 1962) et pour laquelle le choix des sites s'effectue à l'aide de la méthode de la région d'influence (Burn, 1990a). Lors d'études récentes (Tasker et Slade (1994), Tasker et al. (1996)), cette approche a permis d'obtenir de bien meilleurs résultats que l'approche traditionnelle du USGS consistant à séparer l'État en régions hydrologiques géographiques à l'aide d'une analyse du signe des résidus des équations de régression et à appliquer un modèle de régression différent pour chacune des régions ainsi formées. L'État de l'Arkansas a d'ailleurs inclus cette méthode dans son dernier rapport sur l'estimation de la magnitude et de la fréquence des crues (e.g Hodge et Tasker (1995)) avec certaines réserves cependant :

> "The region of influence method is still being improved and is to be considered only as a second alternative to the regional regression equations. The regional regression equations are the recommended procedure."

La motivation première de ce travail a été l'amélioration de la procédure de régression par région d'influence. Réalisant, dans un premier temps, que la méthode de la régression par région d'influence appartient à une famille de modèles de régression non paramétrique connue en statistique sous le nom de **modèles de régression locale**, nous avons d'abord voulu utiliser les concepts de la régression locale afin d'optimiser l'estimation de la méthode de la régression par région d'influence, notamment au niveau du choix du nombre optimal de stations à inclure dans la région d'influence. Puis, réalisant que de par son appartenance à la famille des modèles de régression locale, le modèle de régression régionale par région d'influence se heurte au problème, bien connu en statistique, de la raréfaction des données dans un espace à grande dimension, nous avons plutôt proposé une nouvelle approche de modélisation non paramétrique par région d'influence qui ne soit pas influencée par ce problème de dimensionalité : **l'approche de la modélisation additive par polynômes locaux**.

6.2 Objectifs

L'objectif principal de cette recherche était d'évaluer l'utilisation de l'approche de modélisation additive par polynômes locaux comme méthode alternative d'estimation régionale des débits de crue Q_T en des sites non jaugés. Afin d'atteindre cet objectif, deux étapes principales ont été identifiées. La première étape consistait à présenter une description détaillée des diverses composantes de base d'une approche de modélisation non paramétrique et plus particulièrement de l'approche de modélisation additive par régression locale polynomiale. Plus spécifiquement, il s'agissait de discuter des éléments à prendre en considération lors du choix (1) des variables explicatives à inclure dans le modèle, (2) du niveau de lissage associé à chacune de ces variables explicatives, (3) du degré des différents polynômes locaux, et (4) de la fonction noyau à utiliser. La deuxième étape consistait quant à elle à évaluer les qualités prédictives de l'approche de modélisation proposée. Dans les sections 6.3 et 6.4, nous décrivons comment ces objectifs ont été atteints.

6.3 Démarche

Pour atteindre les objectifs visés, nous avons dans un premier temps rassemblé et résumé à partir de plusieurs documents, les éléments nécessaires à la compréhension de notre travail. Ainsi, aux chapitres 2 et 3, nous avons respectivement présenté les concepts importants de l'approche de modélisation par régression locale et de l'approche de modélisation additive. Nous avons ainsi voulu rendre plus accessibles ces notions à l'ingénieur praticien.

Au chapitre 4, nous avons présenté certains concepts de base liés au problème de la régionalisation des quantiles de crue et avons effectué une revue bibliographique des approches de modélisations jugées pertinentes pour notre travail. Nous avons constaté qu'en ce qui concerne le choix des sites, on reproche aux différentes méthodes leur grande subjectivité et le fait d'être développées pour l'analyse de la similarité hydrologique en utilisant des critères qui ne soient pas directement reliés aux objectifs de l'estimation des crues, soit de produire une estimation de Q_T qui soit la plus précise possible. Par exemple, Nguyen et Pandey (1996) mentionnent :

> "Recent techniques, such as discriminant analysis (Wiltshire, 1986a), region of influence (Burn, 1990a), and discordancy measure (Hosking et Wallis, 1993) also involved a great deal of subjectivity in determining the grouping of homogeneous basins. Further, all previous classification techniques were developed for the assessment of watershed similarity using criteria that are not directly related to the purpose of estimation of floods. The accuracy of flood estimates for an ungauged site based on these techniques is thus rather limited."

Et dans un même ordre d'idée, Roy (1993) indique à propos du caractère subjectif de la méthode de la région d'influence :

> "Bien que Burn (1990b) indique qu'il est préférable pour des considérations de robustesse de retenir un grand nombre de stations, il est très souhaitable de disposer d'une méthode objective pour déterminer le nombre de stations à inclure dans le voisinage"

En ce qui concerne le modèle de transfert d'information régionale, nous avons constaté qu'au cours des dernières années, de nouveaux modèles régionaux sont apparus dans la littérature afin de combler certaines lacunes de leurs prédécesseurs, notamment au niveau de la forme des modèles utilisés qui était inconsistante avec ce qui est généralement observé empiriquement. De plus, suite à de nombreuses critiques (e.g. Potter (1987), NRC (1988), Bobée et Rasmussen (1995)) à l'effet que les différents modèles de régionalisation sont trop "statistiques", on a vu apparaître de nouvelles approches accordant plus d'importance à la physique des phénomènes. Fill et Stedinger (1998) indiquent à ce sujet qu'il est grand temps de mettre de côté les méthodes d'estimation régionales qui ne font pas intervenir la dépendance observée empiriquement des quantiles de crue avec l'aire des bassins versants et les autres caractéristiques physiographiques importantes.

La revue bibliographique du chapitre 4 nous a permis de mettre en évidence les lacunes des méthodes existantes et d'identifier la direction empruntée par les nouvelles approches. Nous avons ainsi pu proposer une nouvelle approche de modélisation faisant intervenir, comme le

suggèrent d'ailleurs Fill et Stedinger (1998), la dépendance entre les quantiles de crue et les caractéristiques physiographiques/climatologiques importantes. La forme de cette dépendance, contrairement aux approches d'invariance d'échelle (Gupta et al., 1994) ou des quantiles normalisés (Fill et Stedinger, 1998), n'est cependant pas déterminée a priori et ce sont plutôt les données qui la déterminent. Nous avons enfin mis en évidence le fait que dans l'approche de modélisation proposée, l'étape du choix des sites se trouve intégrée à la procédure d'estimation du modèle de transfert ce qui, d'un point de vue opérationnel, simplifie la procédure de modélisation régionale et permet de plus une estimation objective des paramètres directement reliée à l'objectif principal des procédures de régionalisation : l'estimation la plus précise possible des quantiles de crue.

Au chapitre 5, nous avons appliqué l'approche de modélisation additive par polynômes locaux et l'avons comparé à un modèle de régression log-linéaire de même qu'à un modèle de régression régionale par région d'influence. De plus, nous avons présenté une approche de calibration du modèle de régression régionale par région d'influence permettant de déterminer de manière optimale et objective le nombre de sites à inclure dans la région d'influence. Nous avons évalué les qualités prédictives des trois modèles pour trois régions des États-Unis.

6.4 Résumé des résultats

Un de nos sous-objectifs était d'évaluer les qualités prédictives de l'approche de modélisation additive par polynômes locaux proposée. Nous avons utilisé des données provenant de 3 régions des États-Unis : le Texas, la Nouvelle-Angleterre et l'Arkansas. Ces trois régions possédaient des caractéristiques particulières dont l'influence sur l'approche proposée se devait d'être évaluée. Les données de la région du Texas semblaient indiquer la présence d'une relation non linéaire entre le logarithme du quantile de crue Q_{50} et le logarithme de la superficie du bassin versant alors que les données de la Nouvelle-Angleterre semblaient plutôt indiquer la présence d'une relation linéaire. La région de l'Arkansas se distinguait quant à elle des autres régions par une forte densité spatiale des stations ce qui implique l'existence d'une corrélation spatiale élevée entre les mesures de débit des sites de la région.

Afin d'évaluer les qualités prédictives de l'approche proposée, nous avons utilisé la procédure du *Jackknife*, une technique de rééchantillonnage permettant d'obtenir une estimation des erreurs de prédiction. Nous avons retenu l'utilisation de deux critères de comparaison, la racine carrée des erreurs quadratiques moyennes de prédiction (RMSE) et la déviation relative moyenne

(DRM), exprimée en pourcentage, des erreurs de prédiction. Nous avons calculé ces critères pour l'estimation (prédiction), en des sites non jaugés, des quantiles de crue Q_2, Q_5, Q_{10}, Q_{25} et Q_{50}. Nous avons comparé les résultats obtenus par l'approche de modélisation additive aux résultats obtenus par un modèle de régression par région d'influence et par le modèle classique de régression log-linéaire.

Avant d'appliquer les différents modèles, nous avons procédé à leur calibration. Partant du fait que la méthode de régression par région d'influence, telle que définie par Tasker et al. (1996), constitue un cas particulier de régression locale multivariée, nous avons proposé l'utilisation d'une approche de calibration permettant de choisir de manière optimale la taille de la région d'influence. Nous avons alors constaté que cette façon de faire permettait de diminuer de manière appréciable les erreurs de prédiction de la méthode. Nous avons aussi modifié quelque peu l'approche de Tasker et al. (1996) en ajoutant une fonction noyau à la méthode. Encore une fois, l'ajout d'une fonction noyau triplement cubique a permis l'obtention de meilleurs résultats. Ces modifications à l'approche initiale de Tasker et al. (1996) ont par exemple permis, dans le cas de l'Arkansas, de réduire l'erreur de prédiction de 38% à 33% (cf. fig. 5.11). Dans le chapitre 5, nous avons donc comparé l'approche de modélisation additive à une approche de régression par région d'influence plus performante que l'approche initiale de Tasker et al. (1996).

Pour la calibration des modèles additifs, nous avons retenu deux approches particulières employant toutefois toutes deux une fonction noyau triplement cubique. La première approche, notée MA1, pénalise peu le nombre de paramètres effectifs (ou degrés de liberté) du modèle final alors que la seconde approche, notée MA2, le pénalise davantage. De MA1 ou MA2, il n'a pas été possible de déterminer a priori l'approche permettant d'obtenir les meilleurs résultats en terme d'erreur de prédiction. Ainsi, au Texas, l'approche MA1 a toujours obtenu les meilleurs résultats alors qu'en Nouvelle-Angleterre et en Arkansas, l'approche MA2 a généralement été la meilleure. Par contre, du point de vue du lissage des diverses relations, il semble que l'approche MA2 produise les estimations dont le niveau de lissage soit le plus adéquat.

En ce qui concerne la comparaison des diverses approches, nous avons constaté que l'approche de modélisation additive (MA) a été la meilleure au Texas et en Arkansas, sauf pour le critère de DRM et la période de retour de 2 ans ou l'approche RI a été la meilleure. De manière générale, on observe que l'écart entre les approches s'accentue lorsque la période de retour augmente. Pour la modélisation de Q_{25} et Q_{50}, on observe des diminutions du RMSE de 11% à 12% au Texas et de 8% à 10% en Arkansas en employant l'approche de modélisation additive plutôt que l'approche de régression par région d'influence. En Nouvelle-Angleterre, région où l'hypothèse

de linéarité semblait respectée, nous avons constaté que les diverses approches de modélisation ont obtenu des résultats voisins. En fait, les approches de modélisation additive et de régression par région d'influence ont toutes deux convergé vers le modèle de régression log-linéaire. Ce résultat nous indique que même lorsque les hypothèses de linéarité émises par le modèle classique de régression multiple semblent respectées, il y a peu de risque à avoir recours à une approche de modélisation additive puisque celle-ci convergera vers le modèle paramétrique linéaire classique.

D'après les résultats obtenus en Arkansas, il semble que la présence de corrélation spatiale entre les mesures de débit de crue ne soit pas un facteur limitant l'applicabilité du modèle additif. Par contre, nous avons remarqué que le problème de multicollinéarité (i.e. de corrélation entre les variables prédictives) fréquemment rencontré dans les études de régionalisation des quantiles de crue pouvait avoir un impact sur la vitesse de calibration des modèles additifs. Nous avons toutefois présenté un algorithme permettant de résoudre ce problème et l'avons appliqué avec succès.

L'approche de modélisation additive proposée produit des estimations des quantiles plus précises (précision prédictive) que les estimations obtenues par la méthode de la régression par région d'influence. Cette approche a en outre un avantage majeur sur l'approche de régression par région d'influence ; elle permet d'observer graphiquement l'effet (conditionnel) des variables explicatives sur la valeur des quantiles modélisés. Par exemple, pour la modélisation de Q_{50}, ces graphiques nous ont permis d'identifier la présence d'une relation non-linéaire (non log-linéaire) pour (1) la variable explicative S (la pente du bassin versant) en Arkansas (cf. fig. 5.10) et (2) la variable explicative A (l'aire du bassin versant) au Texas (cf. fig. 5.4). Ces graphiques permettent d'expliquer la bonne performance, dans ces régions, de la méthode de la région d'influence et de l'approche de modélisation additive par rapport au modèle classique de régression log-linéaire.

6.5 Contribution

Comme on l'a noté dans les sections 6.3 et 6.4, nous avons atteint les objectifs que nous nous sommes fixés pour ce travail. Notre contribution à la régionalisation des quantiles de crue se situe à différents niveaux :

- une revue de la littérature résumant la problématique de la régionalisation des quantiles de crue et permettant de mettre en évidence les lacunes des méthodes existantes et d'identifier la direction empruntée par les nouvelles approches ;

- une synthèse des concepts de base de la régression non paramétrique et, plus particulièrement, des approches de modélisation additive et de lissage par régression locale ;
- la mise en évidence de certaines lacunes de l'approche de la régression par région d'influence de Tasker et al. (1996) ;
- le développement d'une approche de calibration permettant de déterminer de manière optimale la taille de la région d'influence du modèle de Tasker et al. (1996) ;
- un modification mineure apportée à la méthode de Tasker et al. (1996) afin d'en améliorer les capacités prédictives ;
- la présentation détaillée d'une nouvelle approche de modélisation régionale des quantiles de crue ;
- l'application et l'évaluation de la méthode proposée pour trois régions des États-Unis.

6.6 Travaux futurs

Plusieurs voies de recherche se présentent à la suite de ce travail. D'abord, en régression régionale des quantiles de crue, il est bien connu (Stedinger et Tasker, 1985, 1986) qu'un modèle de moindres carrés pondérés ou généralisés est plus représentatif de la problématique réelle de l'estimation régionale des quantiles en raison respectivement de la variance inégale des estimateurs locaux $Q_{T,i}$ (basés sur des tailles d'échantillon différentes) et de la corrélation spatiale annuelle pouvant exister entre des DMA de sites géographiquement voisins.

En ce qui concerne le problème des variances inégales, nous avons vu que le modèle de régression locale ne permet l'utilisation que d'une approche particulière d'estimation par moindres carrés pondérés et que cette estimation requiert une connaissance a priori de la pondération accordée à chacune des observations (chacun des sites). Il serait intéressant de disposer d'une méthode objective et optimale pour déterminer la fonction de pondération à employer. Cette fonction pourrait être obtenue en optimisant un critère de performance et devrait probablement tenir compte de la forme paramétrique de l'expression de la variance des estimateurs locaux des quantiles.

En ce qui concerne le problème de la corrélation spatiale des erreurs, Opsomer et al. (1999), dans une revue sur le problème de la régression non paramétrique en présence d'erreurs corrélées, mentionnent que ce n'est que récemment qu'un certain nombre d'auteurs ont commencé à s'intéresser (pour la régression non paramétrique) au problème de la présence d'une struc-

ture dans la matrice de corrélation des erreurs. Il serait donc intéressant d'effectuer une analyse des différentes méthodes disponibles afin de déterminer celle qui est la plus appropriée pour le problème de la régionalisation des quantiles de crue. À cet égard, l'article de Opsomer et al. (1999) semble être un point de départ intéressant. On y présente d'ailleurs un exemple de régression non paramétrique multidimensionnelle à propos de mesures de l'acidité de lacs et on modélise la corrélation spatiale des erreurs à l'aide de la structure exponentielle suivante : $\mathrm{Cov}(\epsilon_i, \epsilon_j) = \sigma^2 \exp(-d_{ij}/\rho)$ où d_{ij} est la distance euclidienne entre deux lacs i et j. Cette approche pourrait être adaptée au cas de la modélisation additive.

Dans cette thèse, les diverses calibrations des modèles additifs ont été effectuées à l'aide de la version 4.0 du logiciel S-Plus pour Windows. Nous avons ainsi du retenir une procédure de calibration qui est facilement disponible avec S-Plus. D'autres procédures de calibration (e.g. méthode d'Opsomer (Opsomer, 1995), BRUTO (Hastie et Tibshirani, 1990)) sont disponibles et mériteraient d'être programmées et étudiées. Dans un même ordre d'idées, il serait aussi utile de disposer d'un programme indépendant de S-Plus permettant d'effectuer la procédure de modélisation additive proposée. Cette procédure pourrait aussi être ajoutée à certains logiciels hydrologiques, tel que le logiciel HYFRAN.

L'approche de modélisation additive pourrait aussi être appliquée à d'autres domaines de l'hydrologie statistique et plus particulièrement aux domaines où la régression non paramétrique a déjà permis d'obtenir des résultats intéressants. Enfin, cette approche de modélisation n'a été comparée qu'à la méthode de la régression par région d'influence. D'autres études de comparaison sont nécessaires. Il serait intéressant de comparer, par exemple, l'approche de modélisation additive à une approche de régression pour laquelle une analyse des corrélations canoniques a permis de déterminer les voisinages (cf. 4.3.2.4).

Bibliographie

Abi-Zeid, I. (1997). *La modélisation stochastique des étiages et de leurs durées en vue de l'analyse du risque.* Thèse de Doctorat es Sciences (Eau), INRS-Eau, Université du Québec, Ste-Foy (Québec).

Adamowski, K. (1985). Nonparametric kernel estimation of flood frequencies. *Water Resources Research, 21*(11), 1585-1590.

Adamowski, K. (1989). A Monte Carlo comparison of parametric and nonparametric estimation of flood frequencies. *Journal of Hydrology, 108*, 295-308.

Adamowski, K. (1996). Nonparametric estimation of low-flow frequencies. *Journal of Hydraulic Engineering, 122*, 46-49.

Adamowski, K., et Feluch, W. (1991). Application of nonparametric regression to groundwater level prediction. *Canadian Journal of Civil Engineering, 18*, 600-606.

Adamowski, K., Gingras, D., et Pilon, P. J. (1994). *Regional flood frequency analysis by nonparametric and L-Moment methods for Ontario and Quebec* (Report to NSERC Strategic Grant). Ottawa : University of Ottawa, Faculty of Engineering.

Bardsley, W. E. (1989). A simple parameter-free flood ,magnitude estimator. *Journal of Hydrology, 108*, 249-255.

Becker, R. A., Chambers, J. M., et Wilks, A. R. (1988). *The New S Language.* Pacific Grove, CA : Wadsworth & Brooks/Cole.

Bellman, R. E. (1961). *Adaptive Control Processes.* Princeton University Press.

Benson, M. A. (1962). *Evolution of methods for evaluating the occurence of floods* (USGS Water-Supply Paper 1580-A). Washington, DC : US Govt. Printing Office.

Benson, M. A. (1968). Uniform flood frequency estimating methods for federal agencies. *Water Resources Research, 4*(5), 891-908.

Bernier, J. (1990). Les incertitudes hydrologiques dans les problèmes de dimensionnement d'ouvrages. *Revue des sciences de l'eau, 3*(1), 37-53.

Bernier, J. (1993). *Sur les utilisations des L-moments en hydrologie statistique.* Rapport Interne No 128, Institut national de la recherche scientifique, Ste-Foy, Québec.

Bernier, J. (1997). *Risque et décisions en gestion de l'eau : Essai d'analyse de la rationalité du dialogue entre hydrologue et gestionnaire* [Papier présenté au Séminaire Jacques Cartier sur l'Analyse de Décision et du Risque en Hydrologie, 27-29 octobre 1997].

Bobée, B. (1975). The log-Pearson type 3 distribution and its application in hydrology. *Water Resources Research, 11*(5), 681-689.

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

Bobée, B., Cavadias, G., Ashkar, F., Bernier, J., et Rasmussen, P. F. (1993). Towards a systematic approach to comparing distributions used in flood frequency analysis. *Journal of Hydrology, 142*, 121-136.

Bobée, B., et Rasmussen, P. F. (1994). Statistical analysis of annual flood series. *Trends in Hydrology, 1*, 117-135.

Bobée, B., et Rasmussen, P. F. (1995). Recent advances in flood frequency analysis. *U.S. National Report to International Union of Geodesy and Geophysics 1991-1994, Reviews of Geophysics, 33 suppl.*

Bowers, N. L., Gerber, H. U., Hickman, J. C., Jones, D. A., et Nesbitt, C. J. (1986). *Actuarial Mathematics.* Itasca, Illinois : The Society of Actuaries.

Breiman, L., et Friedman, J. H. (1985). Estimating optimal transformation for multiple regression and correlation (with discussion). *Journal of the American Statistical Association, 80*, 580-619.

Buja, A., Hastie, T. J., et Tibshirani, R. J. (1989). Linear smoothers and additive models (with discussion). *Annals of Statistics, 17*(2), 453-555.

Burden, R. L., et Faires, J. D. (1989). *Numerical Analysis.* Boston : PWS-KENT Publishing Company.

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

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

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

Cavadias, G. S. (1990). The canonical correlation approach to regional flood estimation. In *Regionalization in hydrology* (Proceedings of the Ljubljana Symposium, April 1990, p. 171-178). IAHS Publications no. 191.

Chambers, J. M., et Hastie, T. J. (1992). *Statistical Models in S.* Pacific Grove, CA : Wadsworth and Brooks/Cole.

Chow, V. T. (1964). *Handbook of Applied Hydrology.* New York : McGraw-Hill.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association, 74,* 829-836.

Cleveland, W. S., et Devlin, S. J. (1988). Locally weighted regression : An approach to regression analysis by local fitting. *Journal of the American Statistical Association, 83,* 596-610.

Cleveland, W. S., et Loader, C. L. (1994). *Smoothing by local fitting : Various issues stimulated by a paper of Fan and Marron* (Research Report). Murray Hill, New Jersey : AT&T Bell Laboratories.

Cleveland, W. S., et Loader, C. L. (1996a). Smoothing by local regression : Principles and methods. In W. Härdle et M. G. Schimek (Eds.), *Statistical Theory and Computational Aspects of Smoothing* (p. 10-49). New York : Springer.

Cleveland, W. S., et Loader, C. L. (1996b). *Réplique aux discussions de l'article "Smoothing by local regression : Principles and methods" par W. S. Cleveland and C. L. Loader* (Research Report). Murray Hill, New Jersey : AT&T Bell Laboratories.

Craven, P., et Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik, 31,* 377-403.

Cunnane, C. (1978). Unbiaised plotting positions - a review. *Journal of Hydrology, 37,* 205-222.

Cunnane, C. (1987). Review of statistical models for flood frequency estimation. In V. P. Singh (Ed.), *Hydrologic Frequency Modeling* (Proceedings of the International Symposium on Flood Frequency and Risk Analysis, Baton Rouge, LA, p. 49-95). D. Reidel Publishing Company, Boston.

Dalrymple, T. (1960). *Flood-frequency analyses* (Manual of Hydrology : Part 3. Flood-flow techniques). Washington, DC : U.S. Govt. Printing Office.

Dawdy, D. R. (1961). *Variation of flood ratios with size of drainage area* (USGS Research Paper C36). Reston, VA.

DeCoursey, D. G. (1973). Objective regionalization of peak flow rates. In E. F. Koelzer, V. A. Koelzer, et K. Mahmood (Eds.), *Floods and Droughts* (Proceedings of the 2nd International Symposium in Hydrology, September 11-13, 1972, Fort Collins, Colorado, p. 395-405). Fort Collins, Colorado : Water Resources Publications.

Duckstein, L., et Parent, E. (1997). *Cadre général d'analyse et de gestion du risque et applications* [Conférence présentée au Séminaire Jacques Cartier sur l'analyse de décision et du risque en hydrologie, 27-29 octobre 1997].

Efron, B., et Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* New York : Chapman and Hall.

Fill, H. D. (1994). *Improving flood quantile estimates using regional information.* PhD Dissertation, Cornell University, Ithaca, New York.

Fill, H. D., et Stedinger, J. R. (1998). Using regional regression within index flood procedures and an empirical Bayesian estimator. *Journal of Hydrology, 210*, 128-145.

Fisher, R. A., et Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the smallest and largest member of a sample. *Proceedings of the Cambridge Philosophia Society, 24*, 180-190.

Fortin, V. (1994). *Une méthode rationnelle de comparaison des distributions de crue.* Mémoire de maîtrise es Sciences (Eau), INRS-Eau, Université du Québec, Ste-Foy (Québec).

Fortin, V. (1997). *Estimation de la valeur de l'information hydrologique à l'aide de probabilités imprécises.* Thèse de Doctorat es Sciences (Eau), INRS-Eau, Université du Québec, Ste-Foy (Québec).

Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Annals of Statistics, 19*(1), 1-141.

Friedman, J. H., et Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association, 76*, 817-823.

Gingras, D., et Adamowski, K. (1993). Homogeneous region delineation based on annual flood generation mechanisms. *Hydrological Sciences Journal, 38*(2), 103-121.

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

Green, P. J., et Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models.* New York : Chapman and Hall.

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

GREHYS. (1996a). Presentation and review of some methods for regional flood frequency analysis. *Journal of Hydrology, 186*, 63-84.

GREHYS. (1996b). Inter-comparaison of regional flood frequency procedures for Canadian rivers. *Journal of Hydrology, 186*, 85-103.

Greis, N. P. (1983). Flood frequency analysis : A review of 1979-1982. *Reviews of Geophysics and Space Phys., 21*(3), 699-706.

Greis, N. P., et Wood, E. F. (1981). Regional flood frequency estimation and network design. *Water Resources Research*, *17*(4), 1167-1177.

Gu, C., et Wahba, G. (1988). *Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method* (Technical Report 847). University of Wisconsin-Madison.

Gupta, V. K., et Dawdy, D. R. (1995). Physical interpretation of regional variations in the scaling exponents of flood quantiles. *Hydrological processes*, *9*, 347-361.

Gupta, V. K., Mesa, O. J., et Dawdy, D. R. (1994). Multiscaling theory of flood peaks : Regional quantile analysis. *Water Resources Research*, *30*(12), 3405-3421.

Gupta, V. K., et Waymire, E. (1997). Scale invariance and regionalization of floods. In G. Sposito (Ed.), *Scale Invariance and Scale Dependence in Hydrology.* Cambridge University Press.

Halphen, E. (1941). Sur un nouveau type de courbe de fréquence. *Comptes rendus de l'Académie des Sciences, Tome 213*, 633-635.

Härdle, W. (1989). *Applied Nonparametric Regression.* Cambridge : Cambridge University Press.

Hastie, T. J., et Loader, C. L. (1993). Local regression : Automatic kernel carpentry (with discussion). *Statistical Science*, *8*, 120-143.

Hastie, T. J., et Tibshirani, R. J. (1987). Generalized additive models : Some applications. *Journal of the American Statistical Association*, *82*(398), 371-386.

Hastie, T. J., et Tibshirani, R. J. (1990). *Generalized Additive Models.* New York : Chapman and Hall.

Hastie, T. J., Tibshirani, R. J., et Buja, A. (1993). *Flexible discriminant analysis by optimal scoring* (Research Report). Murray Hill, New Jersey : AT&T Bell Laboratories.

Hazen, A. (1914a). Storage to be provided in impounding reservoirs for municipal water supply. *Transactions of the American Society of Civil Engineers*, *77*, 1547-1550.

Hazen, A. (1914b). Discussion de l'article "Flood flows" par W. E. Fuller. *Transactions of the American Society of Civil Engineers*, *77*, 626-632.

Hodge, S. A., et Tasker, G. D. (1995). *Magnitude and frequency of floods in Arkansas* (U.S. Geological Survey Water-Resources Investigations Report 95-4224). Little Rock, Arkansas.

Hoerl, A. E., et Kennard, R. W. (1970). Ridge regression : biased estimation for nonorthogonal problems. *Technometrics*, *12*, 55-67.

Horton, R. E. (1913). Frequency of recurrence of Hudson river floods. *U.S. Weather Bureau Bulletin*, *2*, 109-112.

Hosking, J. R. M. (1986). *The theory of probability weighted moments* (Research Report RC 13412). New York : IBM Research Division.

Hosking, J. R. M. (1990). L-moments : Analysis and estimation of distributions using linear combination of order statistics. *Journal of the Royal Statistical Society, 52*(1), 105-124.

Hosking, J. R. M., et Wallis, J. R. (1993). Some statistics useful in regional frequency analysis. *Water Resources Research, 29*(2), 271-281.

Hosking, J. R. M., Wallis, J. R., et Wood, E. F. (1985a). An appraisal of the regional flood frequency procedure in the UK flood studies report. *Hydrological Sciences Journal, 30*(1), 85-109.

Hosking, J. R. M., Wallis, J. R., et Wood, E. F. (1985b). Estimation of the generalized extreme value distribution by the method of probability weighted moments. *Technometrics, 27*, 251-261.

Hydrology Subcommittee of the Interagency Advisory Committee on Water Data . (1982). *Guidelines for determining flood frequency* (U.S. Geological Survey Bulletin 17B, Office of Water Data Colection). Reston, Virginia.

IEA. (1987). *Australian rainfall and runoff flood frequency analysis and design.* Canberra : Institute of Engineers, Australia.

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

Jennings, M. E., Thomas Jr., W. O., et Riggs, H. C. (1994). *Nationwide summary of U.S. Geological Survey regional regression equations for estimating magnitude and frequency of floods for ungauged sites, 1993* (U.S. Geological Survey Water-Resources Investigation Report 94-4002). Reston, Virginia.

Kinnison, H. B., et Colby, B. R. (1945). Flood formulas based on drainage-basin characteristics. *Transactions of the American Society of Civil Engineers, 110*, 849-904.

Kuczera, G. (1982). Robust flood frequency models. *Water Resources Research, 18*(2), 315-324.

Lall, U. (1995). Nonparametric function estimation : Recent hydrologic applications. *U.S. National Report to International Union of Geodesy and Geophysics 1991-1994, Reviews of Geophysics, 33 suppl.*

Landwehr, J. M., Tasker, G. D., et Jarrett, R. D. (1987). Discussion de l'article "Relative accuracy of log-Pearson III procedures" par J. R. Wallis et E. F. Wood. *Journal of Hydraulic Engineering, 113*, 1206-1210.

Lettenmaier, D. P. (1985). *Regionalisation in flood frequency analysis : Is it the answer ?* [Paper presented at US-China Bilateral Symposium on the Analysis of Extraordinary Flood Events, Nanjing, October 1985].

Lettenmaier, D. P., et Potter, K. W. (1985). Testing flood frequency estimation methods using a regional flood generation model. *Water Resources Research, 21*(12), 1903-1914.

Linsley, R. K. (1986). Flood estimates : How good are they ? *Water Resources Research, 22*(9), 159S-164S.

Loader, C. L. (1999). *Local Regression and Likelihood.* New York, NY : Springer-Verlag.

Mallows, C. (1973). Some comments on C_p. *Technometrics, 15*, 661-675.

Mimikou, M. (1990). Regional analysis of hydrological variables in Greece. In *Regionalization in hydrology* (Proceedings of the Ljubljana Symposium, April 1990, p. 195-201). IAHS Publications no. 191.

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

Morlat, G. (1956). Les lois de probabilité de Halphen. *Revue de Statistique Appliquée, 3*, 21-43.

Mosley, M. P. (1981). Delimitation of New Zealand hydrologic regions. *Journal of Hydrology, 49*, 173-192.

Moss, M. E. (1979). Space, time, and the third dimension (model error). *Water Resources Research, 15*(6), 1797-1800.

Müller, H. G. (1987). Weighted local regression and kernel methods for nonparametric curve fitting. *Journal of the American Statistical Association, 82*, 231-238.

NERC. (1975). *Flood Studies Report* (Vol. 1). London : National Environment Research Council.

Nguyen, V.-T.-V., et Pandey, G. R. (1994). *Regional flood estimation using regression methods : a comparative study* (Water Resources Management and Engineering Series , Research Report no WRME94/1). Montreal : McGill University.

Nguyen, V.-T.-V., et Pandey, G. R. (1996). A new approach to regional estimation of floods in Quebec. In C. E. Delisle et M. A. Bouchard (Eds.), *Compte rendu / Proceedings de la 49ième Conférence annuelle de l'ACRH (CWRA), Québec, 26-28 juin 1996* (Vol. II No. 6 hors série, p. 587-596). Collection Environnement de l'Université de Montréal.

NRC. (1988). *Estimating probabilities of extreme floods : Methods and recommended research.* Washington D.C. : National Academy Press.

Opsomer, J.-D. (1995). *Optimal bandwidth selection for fitting an additive model by local polynomial regression.* PhD Dissertation, Cornell University, Ithaca, New York.

Opsomer, J.-D., Wang, Y., et Yang, Y. (1999). *Nonparametric regression with correlated errors* (Version du 1er février 1999, soumis pour publication). Iowa State University.

Ouarda, T. B., Boucher, G., Rasmussen, P. F., et Bobée, B. (1997). Regionalization of floods by canonical correlation analysis. In J. C. Refsgaard et E. A. Karalis (Eds.), *Operational Water Management* (Proceedings of the European Water Resources Association Conference, Copenhagen, Denmark, 3-6 september 1997, p. 297-302). Rotterdam : A.A. Balkema.

Ouarda, T. B., Haché, M., et Bobée, B. (1998). *Projet C5 - Régionalisation des événements hydrologiques extrêmes* (Rapport de recherche No R-534). Ste-Foy (Québec) : INRS-Eau.

Ouarda, T. B., Lang, M., Bobée, B., Bernier, J., et Bois, P. (1999). Synthèse de modèles régionaux d'estimation de crue utilisés en France et au Québec. *Revue des sciences de l'eau, 12*(1), 155-182.

Owen, A. (1991). Discussion de l'article "Multivariate adaptive regression splines" par J. H. Friedman. *Annals of Statistics, 19*(1), 102-112.

Perreault, L., Bobée, B., et Rasmussen, P. F. (1999a). Halphen distribution system. I : Mathematical and statistical properties. *Journal of Hydrologic Engineering, 4*(3), 189-199.

Perreault, L., Bobée, B., et Rasmussen, P. F. (1999b). Halphen distribution system. II : Parameter and quantile estimation. *Journal of Hydrologic Engineering, 4*(3), 200-208.

Potter, K. W. (1987). Research on flood frequency analysis : 1983-1986. *Reviews of Geophysics, 25*(2), 113-118.

Quenouille, M. H. (1956). Notes on biais in estimation. *Biometrika, 43*, 353-360.

Reinsch, C. (1967). Smoothing by spline functions. *Numerical Mathematics, 10*, 177-183.

Ribeiro-Corréa, J., Cavadias, G. S., Clément, B., et Rousselle, J. (1995). Identification of hydrological neighborhoods using canonical correlation analysis. *Journal of Hydrology, 173*, 71-89.

Riggs, H. (1990). Estimating flow characteristics at ungauged sites. In *Regionalization in hydrology* (Proceedings of the Ljubljana Symposium, April 1990, p. 159-169). IAHS Publications no. 191.

Rossi, F., et Villani, P. (1994a). Regional flood estimation methods. In G. Rossi, N. Harmancioglu, et V. Yevjevich (Eds.), *Coping with Floods* (p. 135-169). Netherlands : Kluwer Academic Publishers.

Rossi, F., et Villani, P. (1994b). A project for regional analsysis of floods in Italy. In G. Rossi, N. Harmancioglu, et V. Yevjevich (Eds.), *Coping with Floods* (p. 193-217). Netherlands : Kluwer Academic Publishers.

Roy, R. (1993). *Régionalisation des caractéristiques de crue - utilisation d'une méthode combinant les approches déterministes et stochastiques.* Thèse de Doctorat es Sciences (Eau), INRS-Eau, Université du Québec, Ste-Foy (Québec).

Roy, R., Bobée, B., Ashkar, F., et Roberge, F. (1989). Regional flood frequency using ridge regression. In *New Directions for Surface-Water Modeling* (Proceedings of the Baltimore Symposium, May 1989, p. 293-300). IAHS Publications no. 181.

Silverman, B. W. (1984). Spline smoothing : The equivalent variable kernel method. *Annals of Statistics, 12*, 898-916.

Slack, J. R., et Landwehr, J. M. (1992). *HCDN : A U.S. Geological Survey streamflow data set for the United States for the study of climate variations, 1874-1988* (U.S. Geological Survey Open-File Report 92-129). Reston, Virginia.

Slack, J. R., Lumb, A. M., et Landwehr, J. M. (1993). *Hydro-Climatic Data Network (HCDN) : Streamflow data set, 1874-1988* (U.S. Geological Survey Water-Resources Investigations Report 93-4076). Reston, Virginia.

Sockett, E. B., Daneman, D., Clarson, C., et Ehrich, R. M. (1987). Factors affecting and patterns of residual insulin secretion during the first year of type I (insulin dependent) diabetes mellitus in children. *Diabet., 30*, 453-459.

Stedinger, J. R., et Cohn, T. A. (1986). Flood frequency analysis with historical and paleoflood information. *Water Resources Research, 22*(5), 785-793.

Stedinger, J. R., et Tasker, G. D. (1985). Regional hydrologic analysis. 1. Ordinary, weighted and generalized least squares compared. *Water Resources Research, 21*(9), 1421-1432.

Stedinger, J. R., et Tasker, G. D. (1986). Regional hydrologic analysis. 2. Model-error estimators, estimation of sigma and log Pearson type 3 distributions. *Water Resources Research, 22*(10), 1487-1499.

Stedinger, J. R., Vogel, R. M., et Foufoula-Georgiou, E. (1993). Frequency analysis of extreme events. In *Handbook of Applied Hydrology* (p. 18.1-18.66).

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics, 10*, 1040-1053.

Stone, C. J. (1985). Additive regression and other nonparametric models. *Annals of Statistics, 13*, 689-705.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society*, *36*(B), 111-147.

Tasker, G. D. (1982). Comparing methods of hydrologic regionalization. *Water Resources Research*, *18*(6), 965-970.

Tasker, G. D., Hodge, S. A., et Barks, C. S. (1996). Region of influence regression for estimating the 50-year flood at ungauged sites. *Water Resources Bulletin*, *32*(1), 163-170.

Tasker, G. D., et Slade, R. M. (1994). An interactive regional regression approach to estimating flood quantiles. In D. G. Fontane et H. N. Tuvel (Eds.), *Water Resources Policy and Management : Solving the Problems* (Proceedings of the 21st Annual Conference on Water Resources Policy and Management, p. 782-785). New York : ASCE.

Tasker, G. D., et Stedinger, J. R. (1987). Regional regression of flood characteristics employing historical information. *Journal of Hydrology*, *96*, 255-264.

Thomas Jr., W. O. (1994). History and overview of flood regionalization methods. In *Nationwide Summary of U.S. Geological Survey Regional Regression Equations for Estimating Magnitude and Frequency of Floods for Ungauged Sites, 1993* (U.S. Geological Survey Water-Resources Investigation Report 94-4002, p. 3-5). Reston, Virginia.

USWRC. (1981). *Guidelines for determining flood flow frequency.* Washington, DC : United States Water Resources Council, Hydrology Committee.

Vieu, P. (1996). Régression non paramétrique multidimensionnelle. In *Recueil des Résumés des Communications des XXVIIIe Journées de Statistique (ASU 96)* (p. 85-87). Sainte-Foy, Université Laval, 27-30 mai 1996.

Wand, M. P., et Jones, M. C. (1990). *Kernel Smoothing.* London : Chapman and Hall.

Weber, J. E., Kisiel, C. C., et Duckstein, L. (1973). On the mismatch between data and models of hydrologic and water resource systems. *Water Resources Bulletin*, *9*(6), 1075-1088.

Wegman, E. J., et Wright, I. W. (1983). Splines in statistic. *Journal of the American Statistical Association*, *78*(382), 351-365.

Weibull, W. (1939). A statistical theory of the strength of materials. In (The Royal Swedish Institute for Engineering Research Proceedings). Ingeniors Vetenskaps Akademiens-Handlingar, no. 151.

White, E. L. (1975). Factor analysis of drainage basin properties : classification of flood behavior in terms of basin geomorphology. *Water Resources Bulletin*, *11*(4), 676-686.

Wiltshire, S. E. (1986a). Regional flood frequency analysis I : Homogeneity statistics. *Hydrological Sciences Journal*, *31*(3), 321-333.

Wiltshire, S. E. (1986b). Regional flood frequency analysis II : Multivariate classification of drainage basins in Britain. *Hydrological Sciences Journal*, *31*(3), 335-346.

Wu, K., et Woo, M. K. (1989). Estimating annual flood probabilities using Fourier series method. *Water Resources Bulletin*, *25*(4), 743-750.

Yakowitz, S. J. (1985). Nonparametric density estimation, prediction, and regression for markov sequences. *Journal of the American Statistical Association*, *80*(389), 205-221.

Yevyevich, V. (1974). Determinism and stochasticity in hydrology. *Journal of Hydrology*, *22*, 225-238.

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

A. Estimation de Q_T par GEV/PWM

Nous présentons dans cette annexe la procédure d'estimation locale de Q_T à partir d'un échantillon de taille n de débits maxima annuels (DMA) $(x_1, x_2, ..., x_n)$ à l'aide de la loi généralisée des valeurs extrêmes (GEV) et de la méthode des moments pondérés par probabilités (PWM).

La loi généralisée des valeurs extrêmes (GEV)

La théorie des valeurs extrêmes origine des travaux de Fisher et Tippett (1928) qui ont démontré l'existence de trois formes asymptotiques possibles pour la distribution de la plus grande (ou plus petite) valeur d'un échantillon de taille n. Jenkinson (1955) a par la suite montré que les trois distributions de valeurs extrêmes pouvaient être exprimées par la forme générale suivante :

$$F(x) = \exp\left\{-(1 - k\frac{(x-\xi)}{\alpha})^{1/k}\right\} \tag{A.1}$$

où α est un paramètre d'échelle, ξ un paramètre de position et k un paramètre de forme.

L'estimation par moments pondérés par probabilités (PWM)

Greenwood et al. (1979) ont défini formellement les PWMs d'une variable aléatoire X de fonction de répartition F par :

$$M_{p,r,s} = E\{x(F)^p F^r [1-F]^s\} \tag{A.2}$$

$$= \int_0^1 [x(F)]^p F^r [1-F]^s dF \tag{A.3}$$

où i, j et k sont des nombres réels et $x(F)$ est la fonction de répartition inverse de F. Pour l'estimation de la loi GEV, nous utilisons un cas particulier de PWM, les β_r définis par :

$$\beta_r = M_{1,r,0} = \int_0^1 x(F) F^r dF \tag{A.4}$$

L'estimation par PWM est similaire à l'estimation à l'aide de la méthode classique des moments en ce sens où nous associons les PWMs estimés à l'aide des observations aux PWMs empiriques de la loi à estimer puis résolvons le système d'équations ainsi formé. Pour l'estimation des PWMs (β_r) empiriques, nous utilisons, tel que le recommandent Hosking, Wallis, et Wood

(1985b), les estimateurs biaisés mais consistants suivants :

$$\hat{\beta}_r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{i-0.35}{n}\right)^r x_{(i)} \tag{A.5}$$

où $x_{(i)}$ représente la i ième plus petite valeur de l'échantillon des DMAs.

Application des PWMs à la loi GEV

En utilisant l'équation (A.1), il est possible d'obtenir pour $x(F)$ l'expression suivante :

$$x(F) = \xi + \frac{\alpha}{k}[1-(-\log(F))^k] \tag{A.6}$$

En insérant ce résultat dans l'équation (A.4), il est possible d'obtenir l'expression générale suivante pour les β_r empiriques :

$$\beta_r = \frac{\xi + \frac{\alpha}{k}[1-(r+1)^{-k}\,\Gamma(1+k)]}{r+1} \tag{A.7}$$

où $\Gamma(.)$ est la fonction gamma définie par $\Gamma(x) = \int_0^\infty t^{x-1}e^{-1}dt$.

Puisque la loi GEV est une loi à trois paramètres, trois PWMs sont requis pour l'estimation. En utilisant β_0, β_1 et β_2, il est alors possible d'obtenir le système d'équations suivant à résoudre :

$$\beta_0 = \xi + \frac{\alpha}{k}[1-\Gamma(1+k)] \tag{A.8}$$

$$2\beta_1 - \beta_0 = \frac{\alpha}{k}\Gamma(1+k)(1-2^{-k}) \tag{A.9}$$

et

$$\frac{3\beta_2-\beta_0}{2\beta_1-\beta_0} = \frac{1-3^{-k}}{1-2^{-k}} \tag{A.10}$$

Remarquons qu'à l'équation (A.10), seul le paramètre de forme k est inconnu. Il n'existe cependant pas de solution explicite pour k mais une bonne approximation (Hosking et al., 1985b) est donnée par :

$$\hat{k} = 7.859c + 2.9554c^2 \tag{A.11}$$

où

$$c = \frac{2\beta_1-\beta_0}{3\beta_2-\beta_0} - \frac{\ln 2}{\ln 3}$$

Les paramètres d'échelle et de forme peuvent alors être estimés respectivement, à l'aide de (A.8) et (A.9), par :

$$\hat{\alpha} = \frac{(2\beta_1 - \beta_0)\hat{k}}{\Gamma(1+\hat{k})(1-2^{-\hat{k}})} \tag{A.12}$$

$$\hat{\xi} = \beta_0 + \frac{\hat{\alpha}}{\hat{k}}[\Gamma(1+\hat{k}) - 1] \tag{A.13}$$

Enfin, en utilisation la relation suivante entre la période de retour T et la probabilité au non dépassement F

$$F = 1 - 1/T \tag{A.14}$$

et en se basant sur l'équation (A.6), on obtient comme expression de l'estimateur du quantile de crue :

$$\hat{Q}_T = \hat{\xi} + \frac{\hat{\alpha}}{\hat{k}}\left[1 - (-\log(1-1/T))^{\hat{k}}\right] \tag{A.15}$$

où les paramètres $\hat{\xi}$, $\hat{\alpha}$ et $\hat{k}$ sont définis respectivement aux équations (A.11), (A.12) et (A.13).

B. Données et résultats du Texas

Nous présentons dans cette annexe les données physiographiques / climatologiques et hydrologiques ayant servi aux diverses modélisations de la section 5.3. Mentionnons que nous présentons les données en unités de mesures métriques alors que la modélisation a plutôt été effectuée avec les données originales, en unités de mesures anglo-saxonnes, disponibles sur internet via FTP à l'adresse internet **ftp ://ftprvares.er.usgs.gov/hcdn92/**. Nous présentons aussi les résultats, sous forme de figures, n'ayant pas été présentés à la section 5.3.

Les données physiographiques et hydrologiques

Nous présentons, dans un premier temps, l'ensemble de données composé des 69 stations du Texas utilisées pour la modélisation. Nous retrouvons dans ce tableau, pour chacune des stations, le numéro d'identification de la station (NO), la superficie (A) du bassin versant, en km^2, la pente (S) du cours d'eau principal, en m/km, la précipitation moyenne annuelle (P) survenant sur le bassin versant, en cm, l'élévation moyenne (EL) du bassin versant, en m, la longueur (L) du cours d'eau principal, en m, le nombre d'observations de débits maxima annuels (n) de même que les estimations des débits de crue (Q_T), en m^3/s, de périodes de retour T égales à 2, 5, 10, 25 et 50 années.

TAB. B.1: Données du Texas

NO	A	S	P	EL	SH	n	Q_2	Q_5	Q_{10}	Q_{25}	Q_{50}
08025500	383.3	1.69	1372	92	36.4	31	86	170	255	415	586
08029500	331.5	2.27	1346	99	33.5	36	49	90	130	205	283
08030500	24160	0.14	1422	86	867	42	1118	1653	2040	2568	2992
08032000	2966	0.43	1105	147	143	49	161	318	459	696	927
08033500	9417	0.24	1245	76	407	56	412	719	941	1246	1492
08033900	409.2	1.38	1168	128	34.4	24	68	117	151	196	231
08041000	20590	0.2	1410	85	563	45	834	1383	1807	2421	2942
08041500	2227	0.73	1372	59	115	52	236	486	726	1154	1595
08042800	1769	0.78	749	324	94	32	66	152	248	446	681
08053500	1036	1.3	813	279	95.6	39	79	176	263	405	542
08055500	6369	1.15	889	237	164	47	357	710	1051	1660	2290

Suite page suivante

TAB. B.1: Données du Texas (suite)

NO	A	S	P	EL	SH	n	Q_2	Q_5	Q_{10}	Q_{25}	Q_{50}
08064800	536.1	1.32	1016	127	41.2	26	29	58	87	139	194
08066200	365.2	1.42	1194	81	34.6	25	95	176	244	350	447
08070000	841.8	0.9	1194	87	68.4	49	127	280	429	696	974
08070500	272	1.54	1181	88	46.3	44	52	98	137	200	258
08080500	22780	1.41	584	737	282	59	212	414	597	904	1204
08082000	13290	1.8	572	805	264	23	246	411	527	682	803
08082500	40240	0.99	673	661	452	64	419	765	1025	1393	1699
08082700	269.4	1.16	660	432	52.6	25	11	32	56	107	169
08083100	590.5	2.09	508	667	63.1	26	21	45	66	99	129
08084800	1238	0.95	635	529	107	26	42	99	158	270	392
08085500	10330	0.86	673	554	319	64	160	333	495	778	1063
08095000	2507	1.85	800	339	136	65	262	460	589	749	867
08095300	471.4	3.6	813	273	50.8	26	88	178	255	378	493
08101000	1178	2.1	787	377	117	38	113	251	379	601	823
08103900	86.25	5.94	749	373	18.8	25	12	23	31	42	50
08109700	611.2	0.85	927	144	64.5	26	43	81	107	139	164
08109800	632	0.8	940	134	55.7	26	51	98	133	184	226
08111700	973.8	0.91	1041	97	75.1	25	298	485	597	724	811
08126500	42510	0.67	470	640	394	45	398	649	811	1010	1154
08128000	1070	2.25	419	718	55.5	58	23	76	141	286	471
08128500	6871	1.47	470	768	175	30	96	184	255	361	453
08130500	593.1	2.82	457	736	48.9	28	13	36	59	104	153
08134000	3279	1.79	508	754	113	64	52	155	275	532	845
08144500	2940	1.51	597	691	96.5	73	66	185	313	565	852
08146000	7889	1.64	660	601	232	73	162	395	638	1104	1620
08150000	4807	1.89	597	670	107	73	185	505	829	1434	2087
08150800	556.8	3.97	635	569	56.3	25	35	82	128	208	290
08151500	10870	1.67	673	596	237	49	403	855	1215	1756	2230
08153500	2334	2.42	749	549	117	49	205	488	777	1322	1913
08158000	101000	0.66	851	616	944	36	1197	2101	3013	4708	6530
08163500	279.7	1.78	1054	107	38.6	49	80	176	268	430	598

Suite page suivante

TAB. B.1: Données du Texas (suite)

NO	A	S	P	EL	SH	n	Q_2	Q_5	Q_{10}	Q_{25}	Q_{50}
08164000	2116	0.62	952	70	130	50	285	523	719	1022	1294
08164300	859.9	0.95	952	103	60.7	27	221	413	556	759	928
08164500	2139	0.55	902	68	152	39	356	667	958	1462	1968
08166000	295.3	5.36	635	625	32.5	45	12	38	69	140	231
08167000	2173	2.52	762	606	103	49	134	332	534	913	1323
08167500	3406	1.73	813	546	209	66	194	464	731	1218	1732
08171000	919.4	3.26	838	404	86.9	61	67	166	266	452	654
08171300	1067	2.58	864	388	124	32	104	206	283	393	484
08172000	2170	2.16	838	285	177	49	195	377	519	727	904
08175000	1422	10.8	813	112	7.67	33	103	278	479	901	1409
08176500	13460	0.8	889	223	666	26	513	1034	1496	2260	2993
08177500	1331	1.21	864	73	98.5	16	137	330	517	854	1204
08179000	1228	3.07	762	461	91.7	41	86	212	342	588	858
08189500	1787	1.02	813	72	118	49	124	295	478	838	1243
08190000	1909	2.8	584	570	98.1	65	101	293	512	973	1525
08190500	1797	2.61	572	616	101	43	68	207	355	643	965
08192000	4820	1.97	572	546	187	49	136	363	582	974	1379
08194200	1215	1.31	546	158	66.1	26	79	199	330	597	909
08194500	20960	1.14	610	276	454	45	246	553	837	1323	1807
08195000	1008	3.79	610	582	70	63	64	175	294	529	795
08196000	326.3	4.74	610	581	52	36	31	69	101	150	195
08198000	533.5	4.26	635	520	52.8	46	40	93	141	225	308
08198500	624.2	3.45	622	485	82.4	36	44	111	175	289	405
08202700	435.1	3.85	635	425	62.9	28	20	55	92	160	233
08205500	8881	2.18	584	337	208	73	119	307	519	962	1492
08206700	2028	1.34	559	188	144	24	104	202	280	397	501
08210000	39960	1.03	635	257	502	68	376	740	1081	1674	2270

Résultats de la modélisation log-linéaire

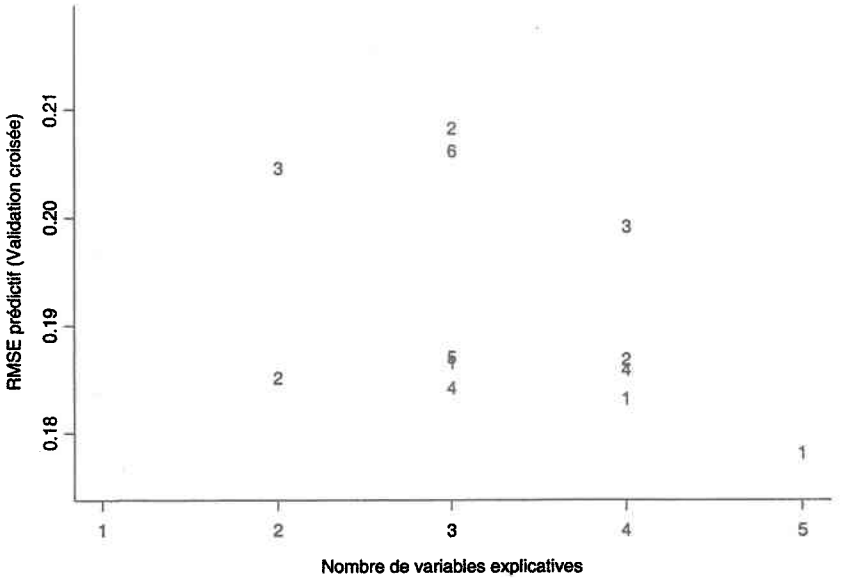

FIG. B.1: RMSE de différents modèles de régression log-linéaire pour l'estimation de Q_2 (voir la légende au tableau 5.3)

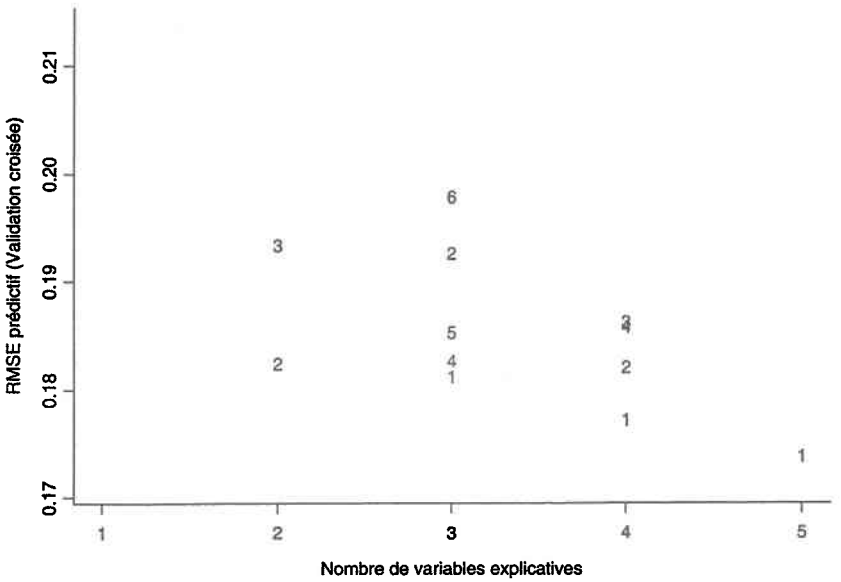

FIG. B.2: RMSE de différents modèles de régression log-linéaire pour l'estimation de Q_5 (voir la légende au tableau 5.3)

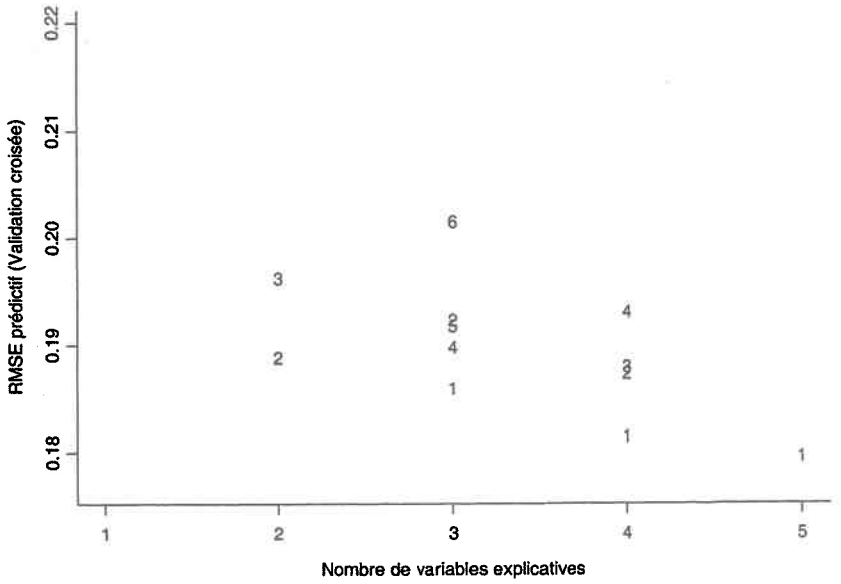

FIG. B.3: RMSE de différents modèles de régression log-linéaire pour l'estimation de Q_{10} (voir la légende au tableau 5.3)

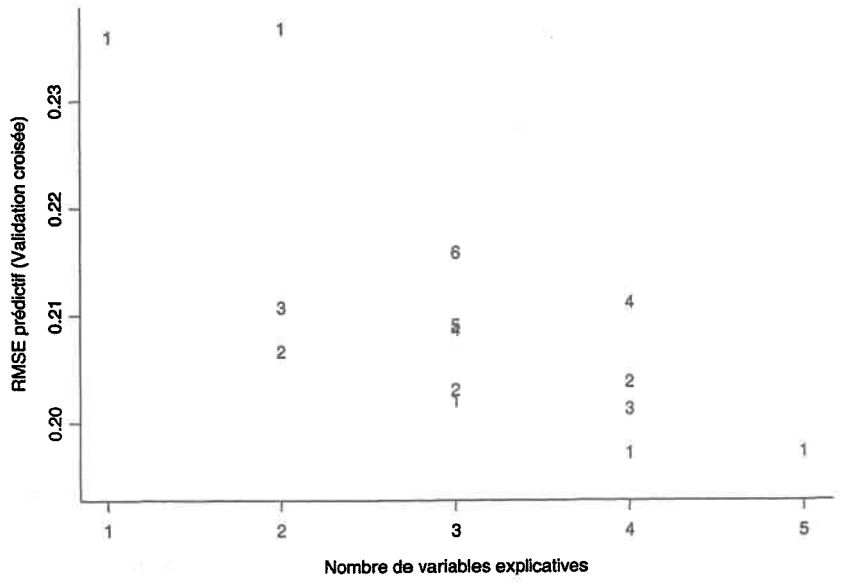

FIG. B.4: RMSE de différents modèles de régression log-linéaire pour l'estimation de Q_{25} (voir la légende au tableau 5.3)

Résultats de la modélisation par région d'influence

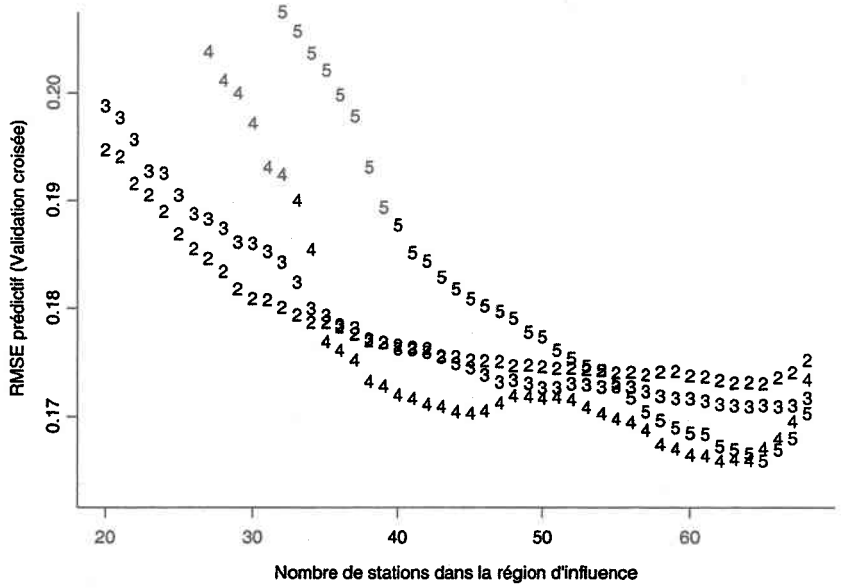

FIG. B.5: Nombre optimal de stations dans la région d'influence pour Q_2

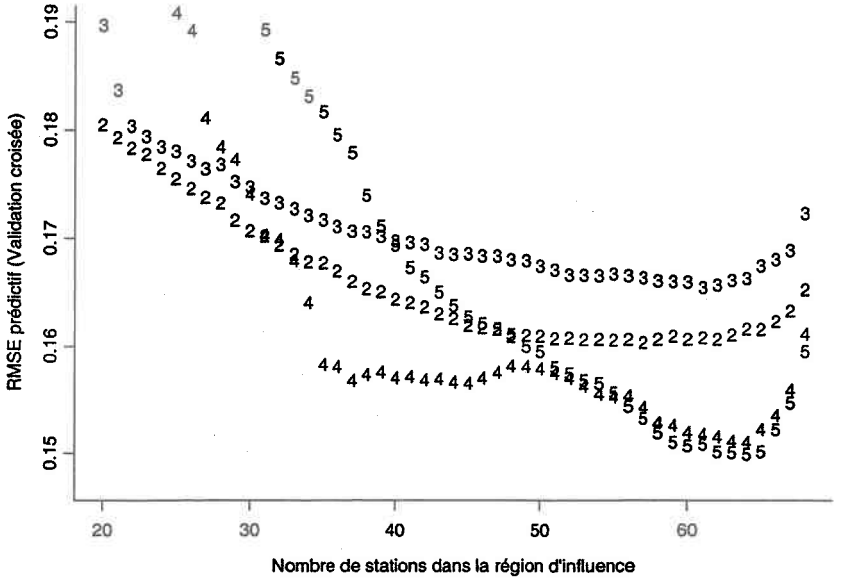

FIG. B.6: Nombre optimal de stations dans la région d'influence pour Q_5

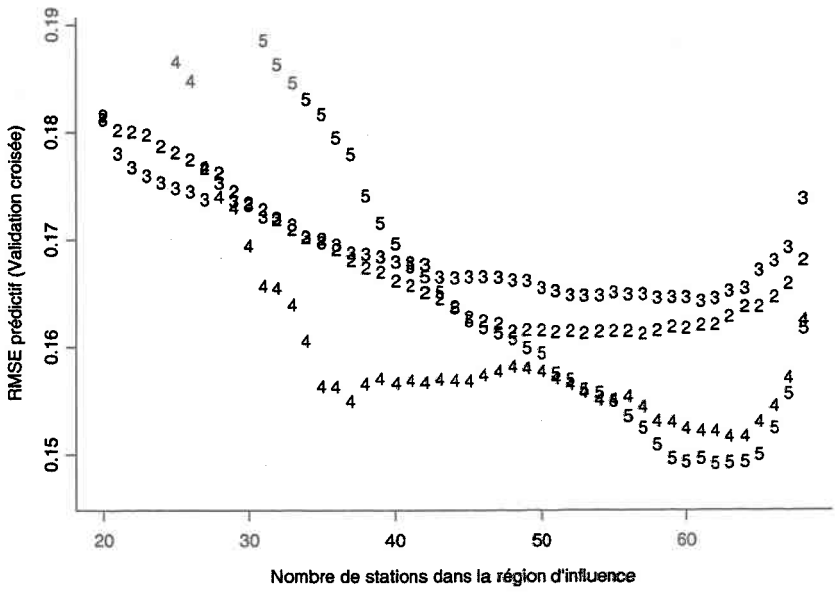

FIG. B.7: Nombre optimal de stations dans la région d'influence pour Q_{10}

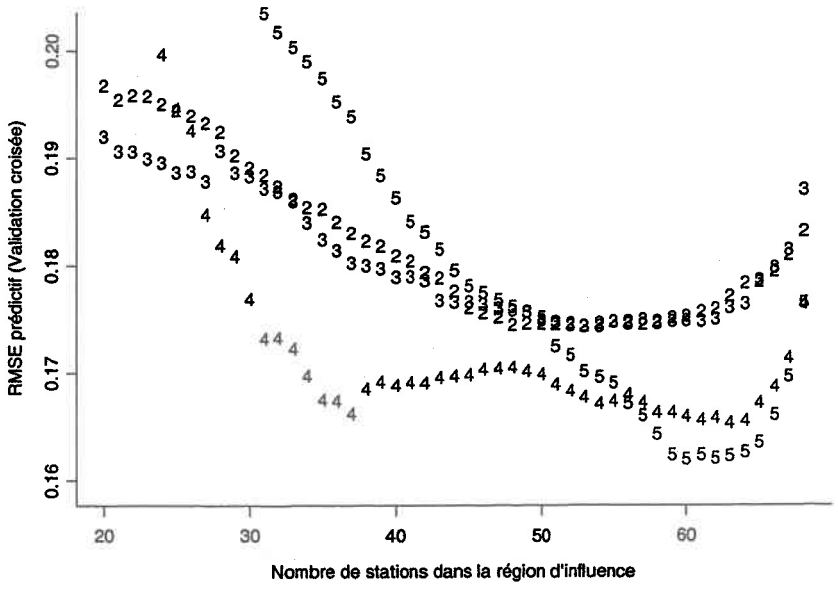

FIG. B.8: Nombre optimal de stations dans la région d'influence pour Q_{25}

C. Données et résultats de la Nouvelle-Angleterre

Nous présentons dans cette annexe les données physiographiques / climatologiques et hydrologiques ayant servi aux diverses modélisations de la section 5.4. Mentionnons que nous présentons les données en unités de mesures métriques alors que la modélisation a plutôt été effectuée avec les données originales, en unités de mesures anglo-saxonnes, disponibles sur internet via FTP à l'adresse internet **ftp ://ftprvares.er.usgs.gov/hcdn92/**. Nous présentons aussi les résultats, sous forme de figures, n'ayant pas été présentés à la section 5.4.

Les données physiographiques et hydrologiques

Nous présentons, dans un premier temps, l'ensemble de données composé des 70 stations de la Nouvelle-Angleterre utilisées pour la modélisation. Nous retrouvons dans ce tableau, pour chacune des stations, le numéro d'identification de la station (NO), la superficie (A) du bassin versant, en km^2, la pente (S) du cours d'eau principal, en m/km, la précipitation moyenne annuelle (P) survenant sur le bassin versant, en cm, l'élévation moyenne (EL) du bassin versant, en m, la longueur (L) du cours d'eau principal, en m, le nombre d'observations de débits maxima annuels (n) de même que les estimations des débits de crue (Q_T), en m^3/s, de périodes de retour T égales à 2, 5, 10, 25 et 50 années.

TAB. C.1: Données de la Nouvelle-Angleterre

NO	A	S	P	EL	SH	n	Q_2	Q_5	Q_{10}	Q_{25}	Q_{50}
01010000	3473	1.1	1022	448	113	38	665	853	955	1064	1132
01010500	6941	1.1	1016	402	199	42	1350	1811	2083	2394	2602
01011000	3183	0.64	994	369	147	57	407	558	656	778	868
01011500	1357	1.32	989	369	100	37	204	283	330	384	421
01013500	2261	0.53	1001	277	117	64	233	298	334	372	396
01014000	14670	1.08	1004	366	246	62	2301	2988	3383	3824	4114
01018000	453.2	4.99	943	177	24.5	41	90	125	147	175	195
01021200	240.6	1.02	1071	67	45.7	33	37	49	59	73	84
01021500	1184	1.09	1067	98	92.8	63	160	216	259	319	369
01022500	587.9	2.07	1066	98	61.8	40	109	153	182	219	247

Suite page suivante

TAB. C.1: Données de la Nouvelle-Angleterre (suite)

NO	A	S	P	EL	SH	n	Q_2	Q_5	Q_{10}	Q_{25}	Q_{50}
01023000	383.3	1.47	1050	134	51.2	60	45	60	70	84	95
01030500	3673	0.77	997	174	142	54	467	600	674	755	808
01031500	771.8	5.01	1073	320	58.9	86	185	277	352	469	575
01033500	839.2	6.41	1101	351	62.4	59	199	304	388	514	624
01035000	774.4	2	1017	128	35.7	64	56	76	90	110	126
01038000	375.5	2.83	1073	104	41.7	50	49	77	100	137	170
01047000	914.3	9.2	1123	390	65	67	239	366	465	609	733
01048000	1331	2.78	1102	317	94.1	51	306	444	541	672	776
01052500	393.7	7.77	1238	664	39.7	47	111	141	160	183	199
01054200	180.3	19.82	1170	719	25.6	24	99	145	173	204	226
01055000	251	16.46	1194	558	29	59	82	123	154	200	238
01055500	437.7	2.26	1062	323	40.4	47	85	130	166	219	266
01057000	196.3	9.83	1095	326	22.8	67	50	75	94	120	143
01060000	365.2	1.95	1092	67	47.9	39	96	137	166	204	234
01064500	997.2	9.66	1284	567	56.5	65	310	455	559	700	813
01073000	31.34	4.07	1052	61	12.7	53	6	10	12	15	17
01074500	269.4	20.61	1419	853	23.7	23	93	139	185	270	360
01075000	499.9	15.28	1384	759	37.8	38	164	245	314	424	527
01076000	370.4	20.28	1169	482	36	48	94	145	188	256	319
01076500	1611	7.95	1233	564	65.2	85	435	627	773	985	1162
01078000	222.2	4.28	1134	384	34.6	70	43	62	78	102	123
01086000	378.1	6.02	1126	293	38.1	39	57	79	92	108	119
01094000	442.9	5.91	1110	247	55	67	75	111	141	190	236
01106000	20.75	6.1	1092	43	9.65	38	4	5	6	7	7
01111300	41.44	6.65	1143	165	11.6	24	10	16	20	25	29
01111500	236.2	4.47	1143	152	29.1	32	42	65	83	108	128
01117500	259	0.84	1148	61	29.1	47	20	28	34	43	51
01117800	91.17	5.89	1194	110	17.2	23	12	18	21	24	27
01118000	187.5	3.07	1194	91	22.5	36	25	35	42	51	58
01118500	764	0.51	1168	61	59.9	47	64	87	105	131	151
01119500	313.4	2.71	1092	216	46.7	57	50	88	123	185	248

Suite page suivante

TAB. C.1: Données de la Nouvelle-Angleterre (suite)

NO	A	S	P	EL	SH	n	Q_2	Q_5	Q_{10}	Q_{25}	Q_{50}
01120500	10.75	15.15	1118	223	5.31	31	4	6	8	10	12
01121000	74.07	12.86	1105	195	16.7	48	17	26	33	46	57
01127500	231.3	3.64	1143	131	26.9	58	52	83	111	159	207
01134500	194.8	14.85	1097	518	27.4	41	47	64	77	93	106
01137500	226.9	13.64	1323	765	33.9	49	69	96	118	149	176
01142500	78.99	15.23	992	402	16.4	48	14	19	22	25	28
01144000	1787	2.48	1047	396	80	72	375	507	595	706	789
01145000	208.5	9.51	1109	427	23.2	39	37	50	59	68	74
01153500	266.8	10.7	1118	408	34.8	45	66	94	114	139	158
01162500	50.25	5.15	1098	338	19	36	9	13	16	20	24
01165000	130.8	9.83	1072	323	20.6	31	18	30	41	61	82
01165500	31.34	9.11	1130	265	11.6	66	6	9	11	15	18
01169000	230.5	12.42	1240	439	33.8	39	75	114	143	184	217
01169900	62.42	16	1194	354	19	22	21	30	36	45	52
01170100	107.2	13.6	1168	421	28	21	33	46	54	65	72
01174000	8.78	12.92	1130	311	9.41	34	2	3	3	4	5
01174900	6.604	31.25	1143	274	6.19	27	2	3	3	4	4
01175500	489.5	2.23	1128	244	41.8	27	45	67	88	124	161
01176000	388.5	1.63	1054	256	45.5	52	28	43	59	92	131
01180000	4.481	22.35	1118	338	6.15	28	1	1	2	3	4
01180500	136.5	14.96	1224	433	31.4	54	41	62	83	119	156
01181000	243.5	10.4	1206	433	33.8	50	78	127	169	235	295
01188000	10.62	14.62	1219	280	6.11	57	3	5	7	10	13
01193500	259	5.74	1143	154	29.8	60	53	86	115	164	211
01196500	297.8	1.14	1194	95	42.6	58	47	73	92	120	142
01198000	132.1	10.27	1123	360	24.5	20	24	32	36	40	42
01201500	174.8	2.22	1156	167	40.2	35	23	38	54	88	126
01204000	194.5	9.17	1130	223	31.1	56	42	71	98	145	192
04296000	316	1.09	1052	378	54.1	37	51	67	77	88	96

Résultats de la modélisation log-linéaire

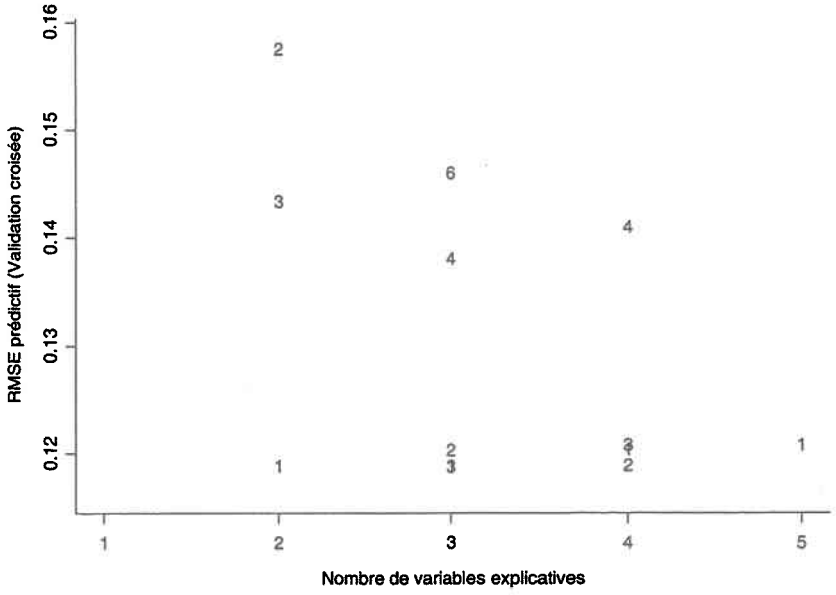

FIG. C.1: RMSE de différents modèles de régression log-linéaire pour l'estimation de Q_2 (voir la légende au tableau 5.3)

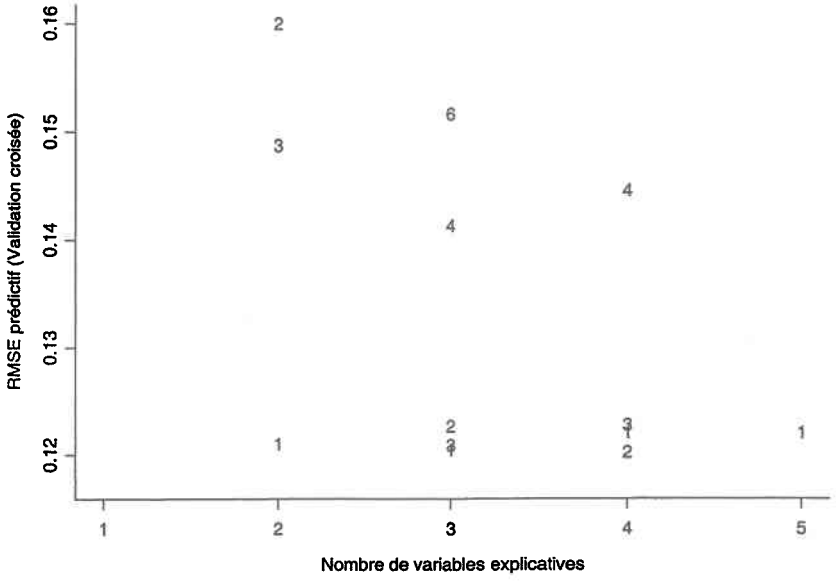

FIG. C.2: RMSE de différents modèles de régression log-linéaire pour l'estimation de Q_5 (voir la légende au tableau 5.3)

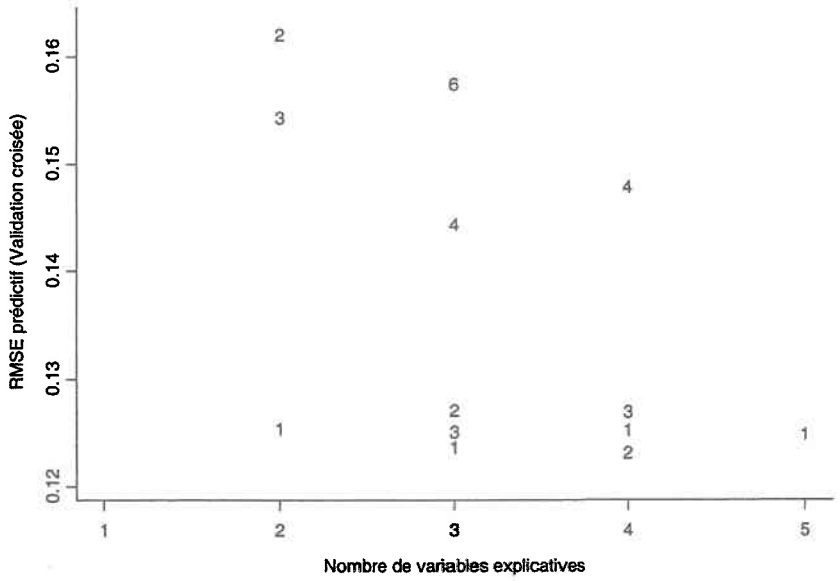

FIG. C.3: RMSE de différents modèles de régression log-linéaire pour l'estimation de Q_{10} (voir la légende au tableau 5.3)

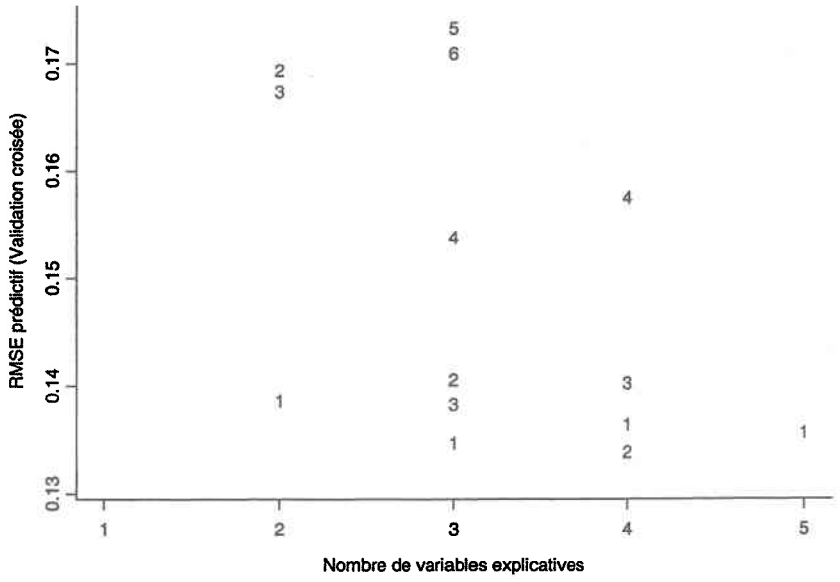

FIG. C.4: RMSE de différents modèles de régression log-linéaire pour l'estimation de Q_{25} (voir la légende au tableau 5.3)

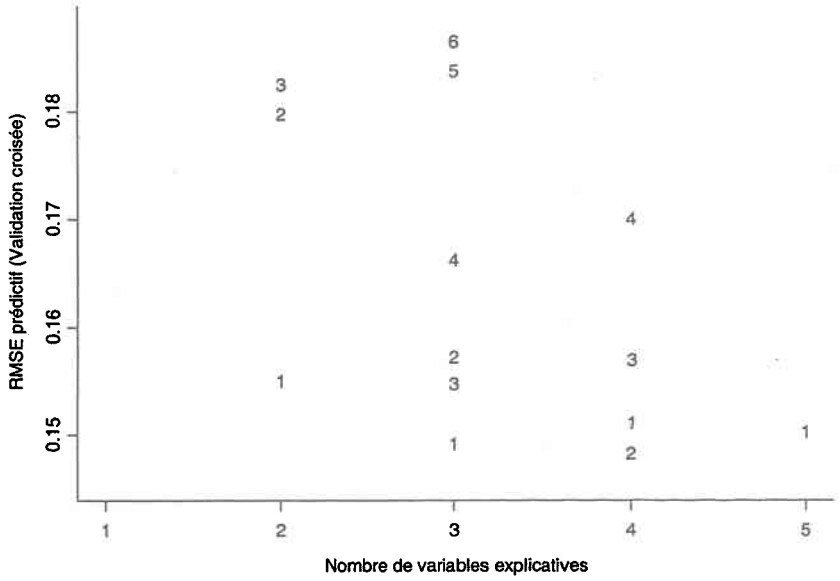

FIG. C.5: RMSE de différents modèles de régression log-linéaire pour l'estimation de Q_{50} (voir la légende au tableau 5.3)

Résultats de la modélisation par région d'influence

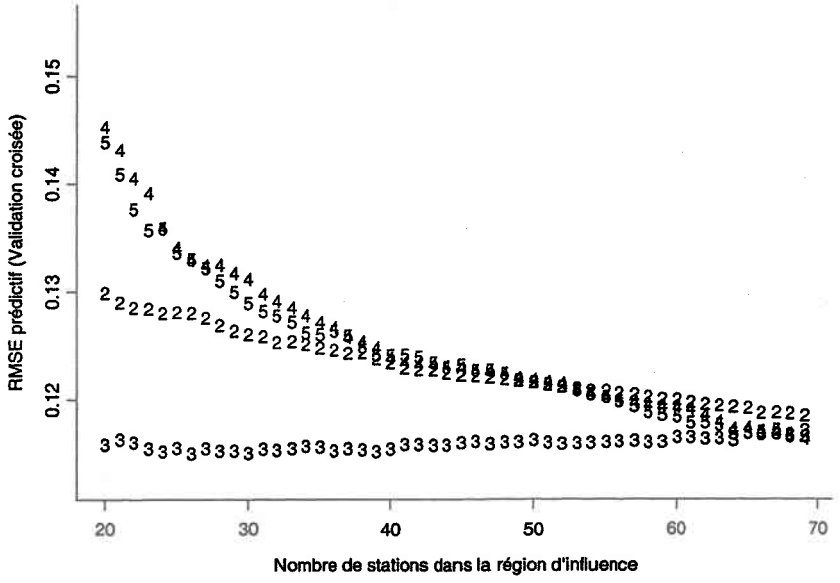

FIG. C.6: Nombre optimal de stations dans la région d'influence pour Q_2

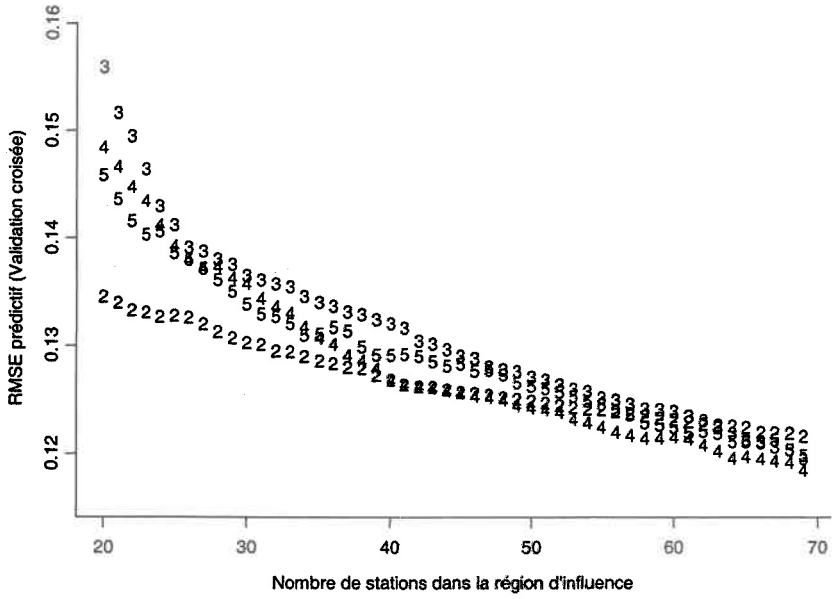

FIG. C.7: Nombre optimal de stations dans la région d'influence pour Q_5

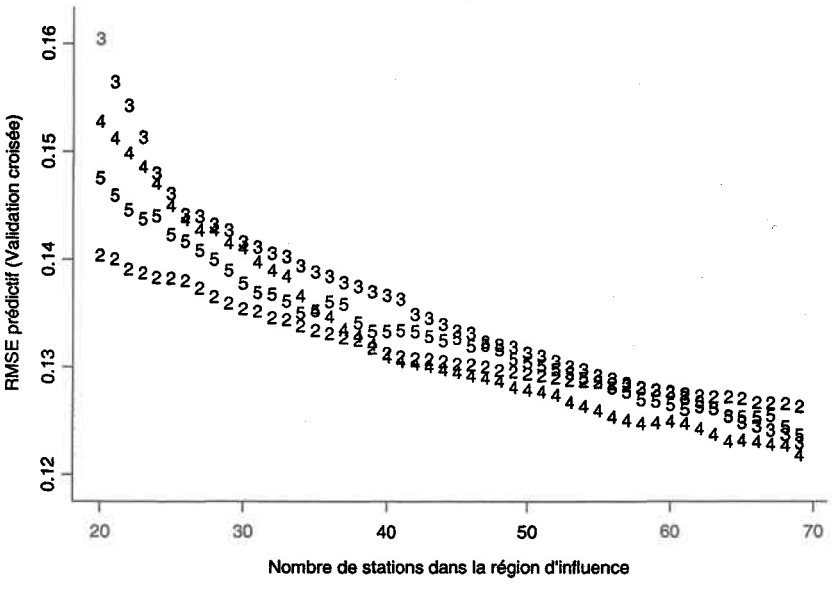

FIG. C.8: Nombre optimal de stations dans la région d'influence pour Q_{10}

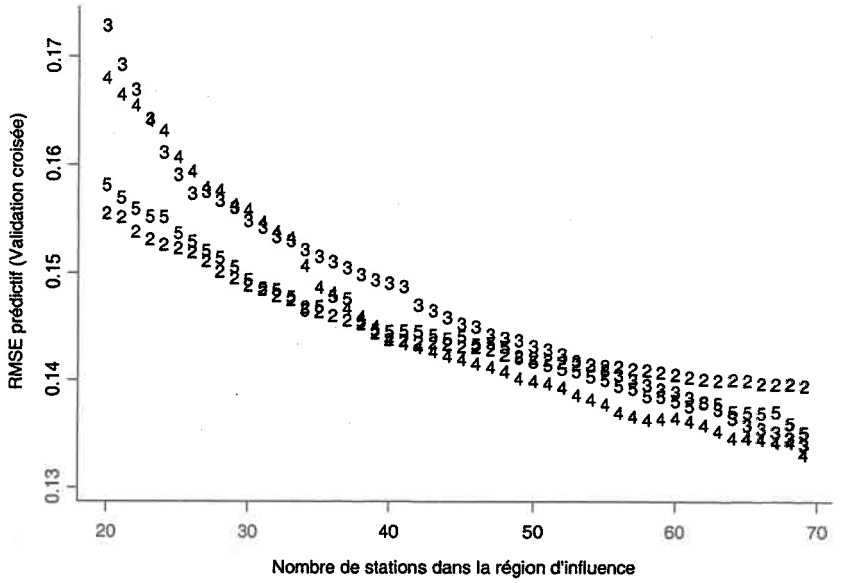

FIG. C.9: Nombre optimal de stations dans la région d'influence pour Q_{25}

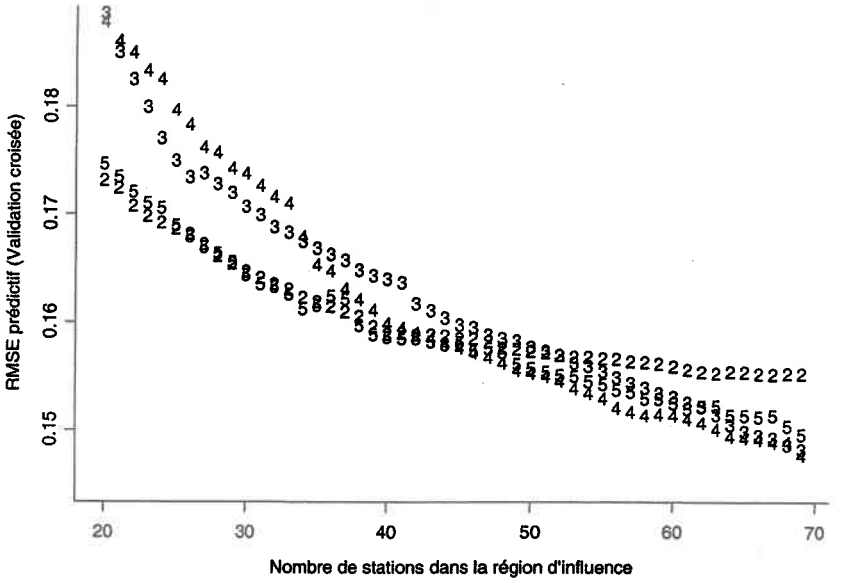

FIG. C.10: Nombre optimal de stations dans la région d'influence pour Q_{50}

Résultats de la modélisation additive

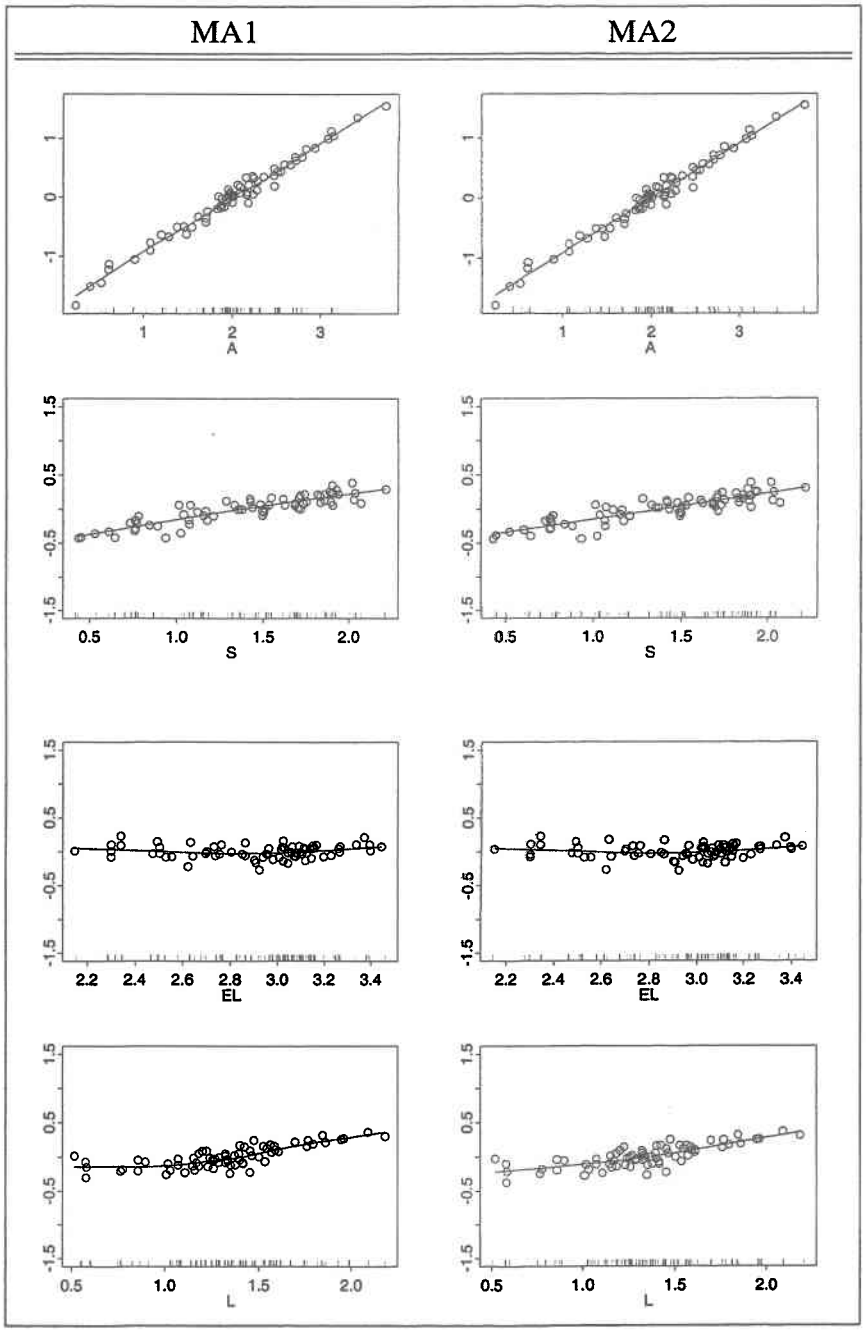

FIG. C.11: Modélisation additive de Q_2

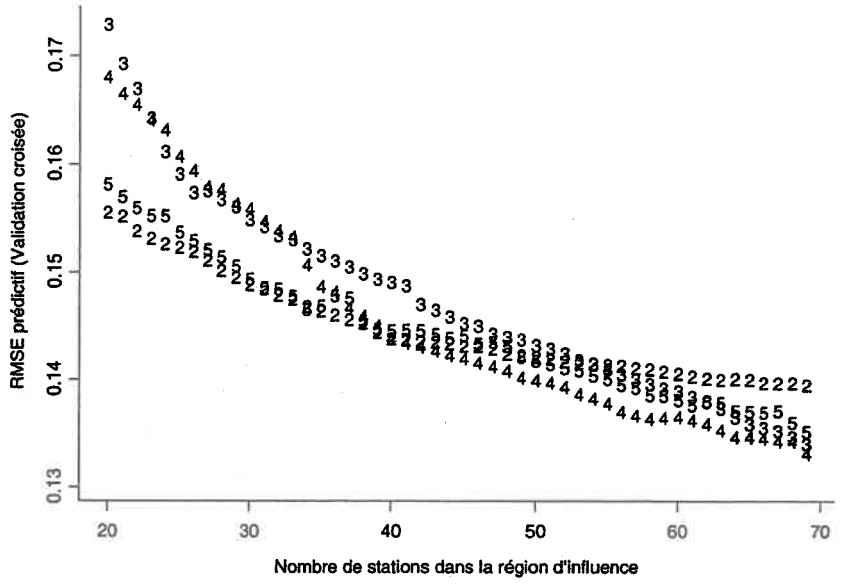

FIG. C.9: Nombre optimal de stations dans la région d'influence pour Q_{25}

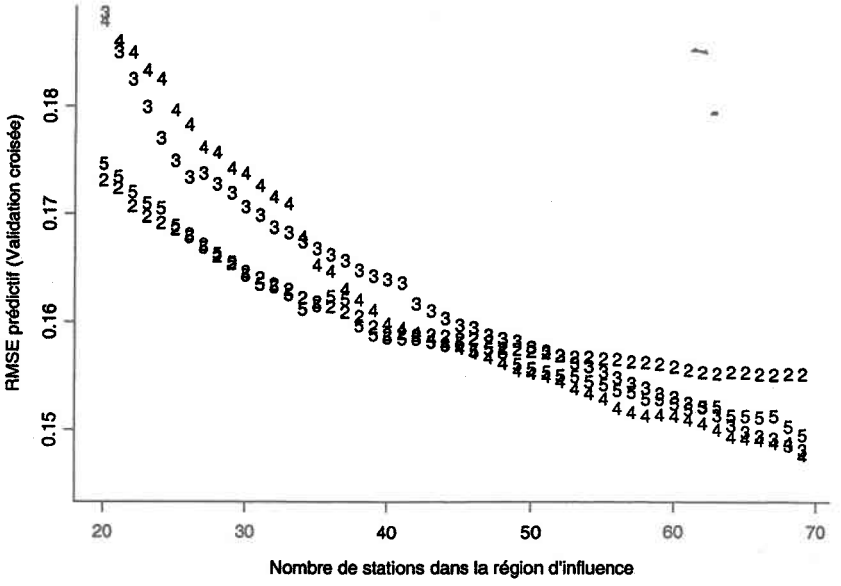

FIG. C.10: Nombre optimal de stations dans la région d'influence pour Q_{50}

Résultats de la modélisation additive

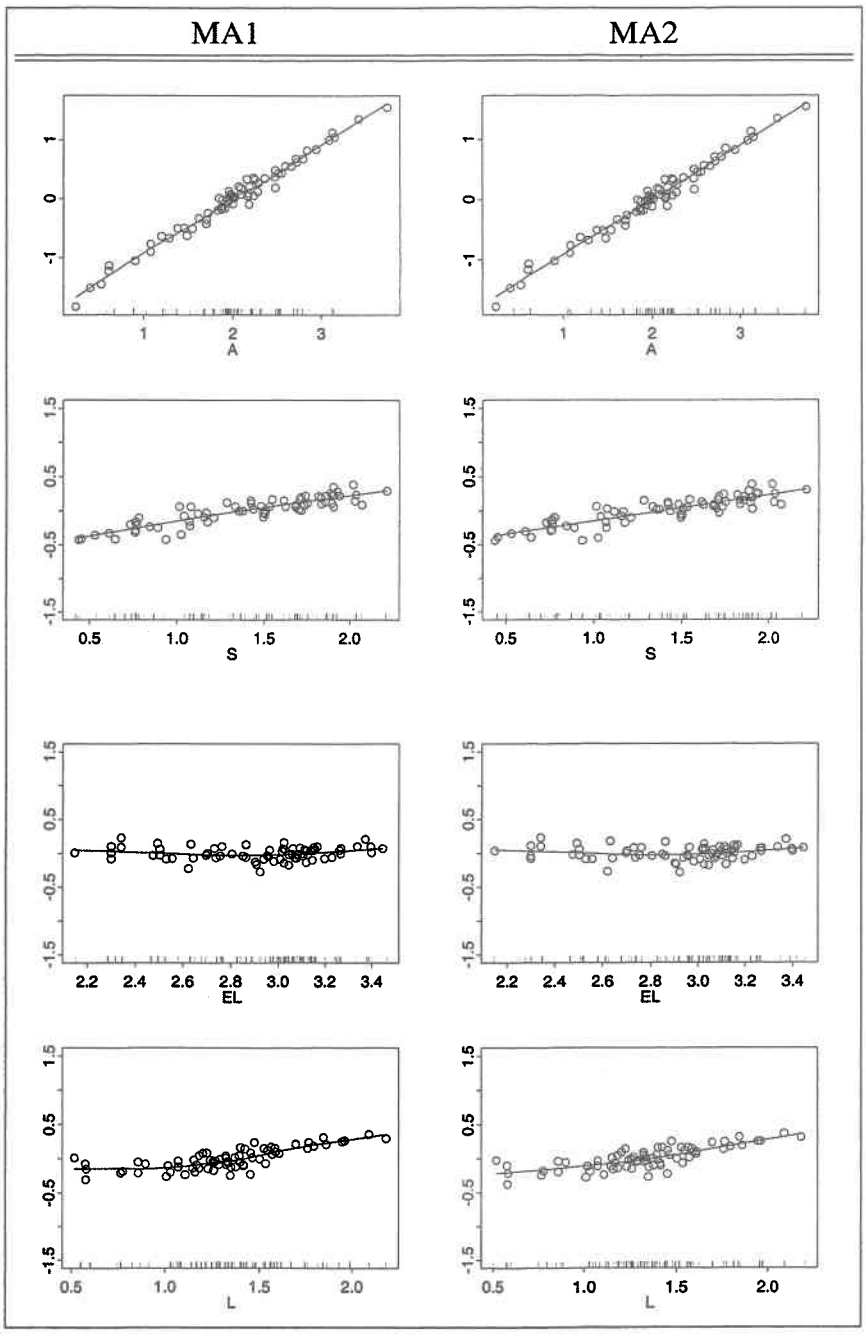

FIG. C.11: Modélisation additive de Q_2

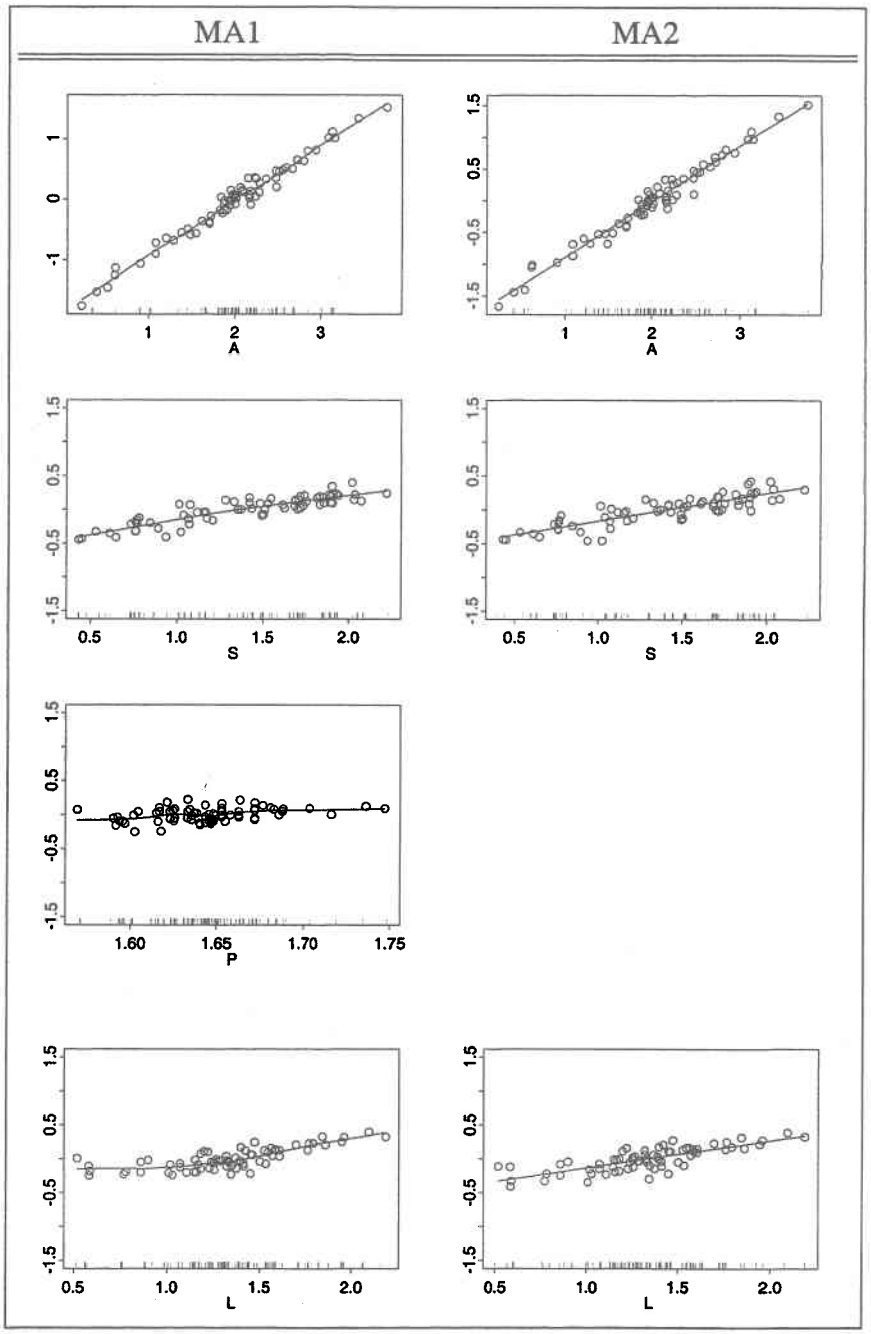

FIG. C.12: Modélisation additive de Q_5

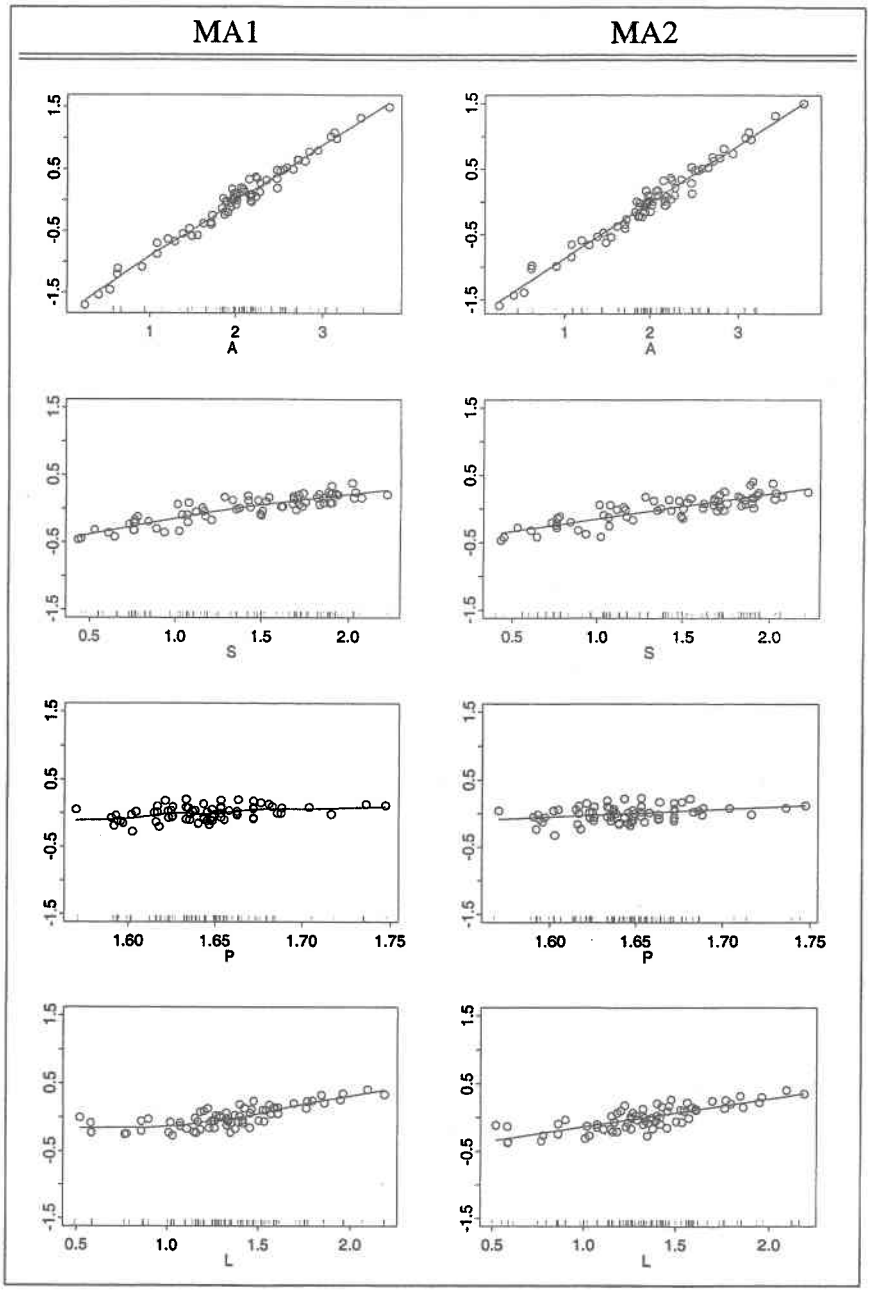

FIG. C.13: Modélisation additive de Q_{10}

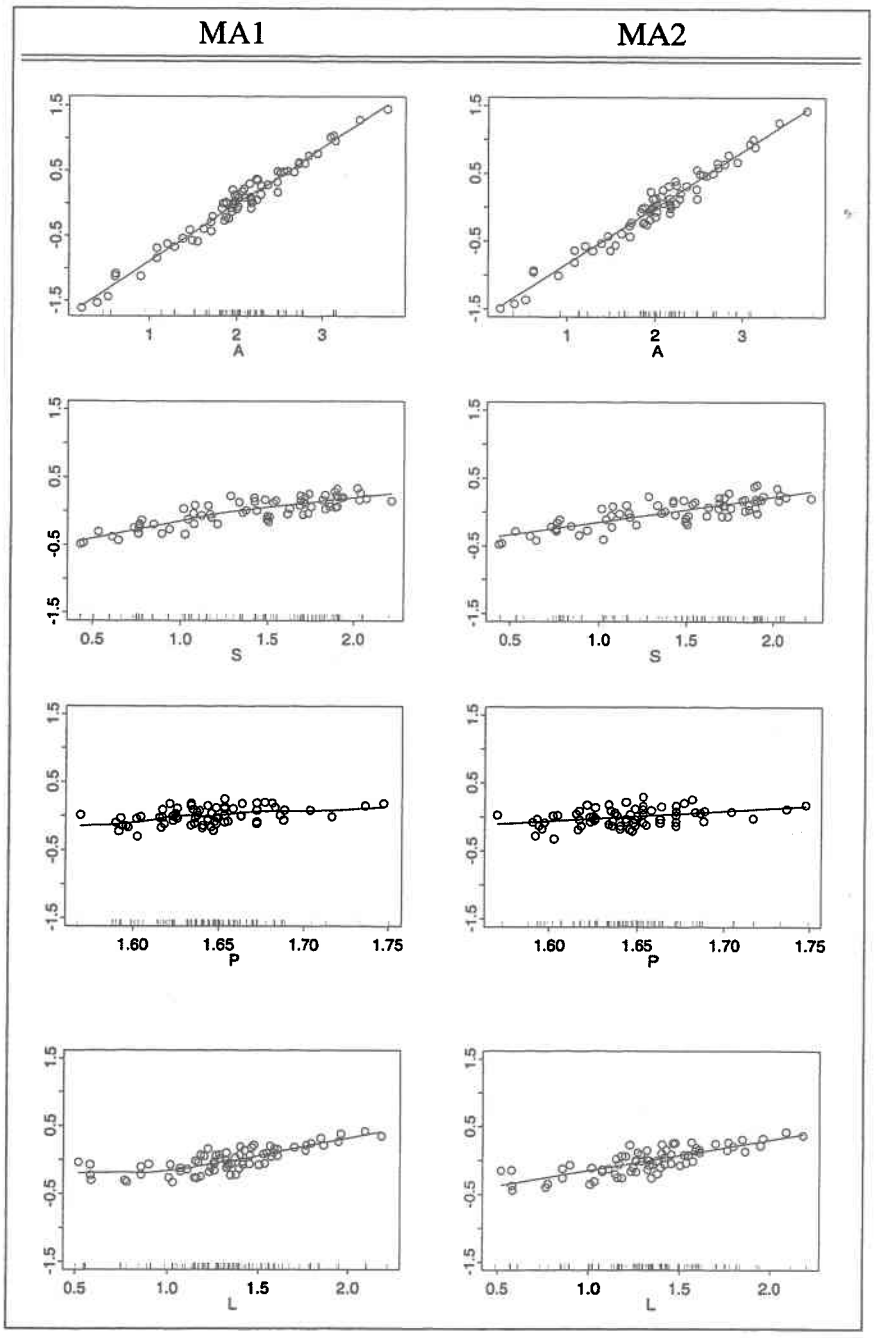

FIG. C.14: Modélisation additive de Q_{25}

D. Données et résultats de l'Arkansas

Nous présentons dans cette annexe les données physiographiques / climatologiques et hydrologiques ayant servi aux diverses modélisations de la section 5.5. Nous présentons aussi les résultats, sous forme de figures, n'ayant pas été présentés à la section 5.5.

Les données physiographiques et hydrologiques

Nous présentons, dans un premier temps, l'ensemble de données composé de 204 stations de l'Arkansas. Nous retrouvons dans ce tableau, pour chacune des stations, le numéro d'identification de la station (NO), la superficie (A) du bassin versant, en km^2, la pente (S) du cours d'eau principal, en m/km, la précipitation moyenne annuelle (P) survenant sur le bassin versant, en cm, l'élévation moyenne (EL) du bassin versant, en m, le facteur de forme (SH) du bassin, le nombre d'observations de débits maxima annuels (n) de même que les estimations des débits de crue (Q_T), en m^3/s, de périodes de retour T égales à 2, 5, 10, 25 et 50 années.

TAB. D.1: Données de l'Arkansas

NO	A	S	P	EL	SH	n	Q_2	Q_5	Q_{10}	Q_{25}	Q_{50}
07188900	2,49	20,60	109	354	0,240	21	3	8	14	22	29
07191220	344	3,79	109	366	0,275	34	96	268	444	743	1020
07194890	105	3,69	109	396	0,247	21	41	117	199	347	493
07195000	337	3,20	109	387	0,392	30	142	298	425	606	754
07195200	0,96	20,30	109	390	0,370	21	2	5	7	12	15
07195450	37,80	7,77	114	415	0,282	24	50	114	164	232	284
07195500	1640	1,61	114	424	0,321	38	527	985	1330	1790	2140
07195800	36,80	4,30	109	408	0,369	33	20	56	98	177	260
07196000	285	3,67	112	363	0,201	38	115	317	522	870	1200
07196900	119	7,62	117	402	0,342	36	196	383	517	685	806
07247000	526	1,85	114	253	0,180	30	324	565	740	973	1150
07249300	114	8,81	107	308	0,260	20	149	314	458	681	877
07249400	381	2,69	109	235	0,203	36	184	317	419	564	682
07249500	91,40	7,01	117	427	0,197	44	131	267	392	594	779

Suite page suivante

TAB. D.1: Données de l'Arkansas (suite)

NO	A	S	P	EL	SH	n	Q_2	Q_5	Q_{10}	Q_{25}	Q_{50}
07249650	21,10	13,80	117	433	0,290	20	37	70	95	130	158
07249950	0,88	35,60	117	305	0,402	32	1	3	4	7	11
07250000	1100	3,30	114	326	0,151	52	691	1210	1610	2140	2560
07252000	966	3,43	122	436	0,136	55	568	1030	1360	1780	2090
07252200	1,19	60,20	117	280	0,995	26	4	8	10	13	16
07252500	11	10,60	104	219	0,145	16	24	42	54	70	82
07254000	7,15	11,40	102	195	0,156	18	11	22	32	44	55
07254500	15,10	3,75	104	219	0,043	16	25	41	53	68	80
07255100	11,60	1,48	102	140	0,154	15	22	43	59	79	95
07256000	137	4,51	102	155	0,274	16	89	143	179	226	261
07256500	158	9,34	124	265	0,264	41	153	287	387	521	625
07257000	710	3,22	124	430	0,157	45	556	1040	1440	2020	2510
07257060	0,49	70,30	127	533	0,530	20	1	2	3	4	5
07257100	0,49	58,90	124	165	0,130	23	1	3	4	5	6
07257200	399	4,45	117	293	0,133	15	266	320	349	379	399
07257500	624	5,30	124	402	0,207	47	500	909	1250	1750	2180
07257700	18,30	15,60	124	268	0,149	26	21	61	79	122	160
07258200	2,38	11,10	109	247	0,209	30	5	9	12	16	20
07258500	624	1,86	112	204	0,205	55	333	548	696	885	1020
07260000	211	3,67	119	283	0,097	48	188	302	382	489	572
07260500	1980	0,56	117	219	0,099	30	429	834	1170	1680	2120
07260630	4,79	7,35	119	163	0,420	23	13	24	33	44	54
07260673	575	2,27	122	210	0,234	15	155	327	495	781	1060
07260679	0,23	35,60	119	113	0,780	27	1	2	2	2	3
07261000	438	1,40	127	219	0,130	38	253	377	455	550	616
07261050	0,75	16,70	127	213	0,350	23	2	5	7	10	12
07261300	6,03	54,60	119	415	0,228	22	12	28	43	67	90
07261500	1060	2,08	117	317	0,114	55	710	1260	1680	2260	2720
07261800	2,69	31,80	117	271	0,352	31	7	13	18	26	33
07263000	544	2,29	122	253	0,200	52	557	896	1140	1460	1700
07263100	3,81	25,40	124	151	0,177	32	10	16	21	27	31

Suite page suivante

TAB. D.1: Données de l'Arkansas (suite)

NO	A	S	P	EL	SH	n	Q_2	Q_5	Q_{10}	Q_{25}	Q_{50}
07263400	38,90	6,72	135	180	0,295	24	66	134	188	267	330
07263530	83,90	3,90	130	168	0,214	14	87	121	144	172	194
07263910	6,16	6,29	127	99	0,308	26	18	24	29	34	38
07338700	41,70	9,26	135	378	0,178	21	57	101	136	189	234
07339500	471	3,52	132	256	0,148	25	434	831	1180	1730	2220
07339800	16,60	9,03	130	168	0,156	26	29	69	103	155	199
07340000	6890	0,80	132	250	0,127	40	305	2000	2500	3170	3700
07340200	27,40	2,27	124	126	0,231	22	45	75	96	122	141
07340300	232	5,66	135	381	0,265	27	432	744	959	1230	1430
07340500	932	2,94	137	271	0,125	36	799	1330	1730	2310	2790
07340530	1,66	11,50	127	120	0,165	24	5	10	13	18	21
07341000	313	4,07	140	232	0,094	35	272	501	685	952	1170
07341100	24,60	9,47	132	177	0,231	24	59	128	184	264	329
07341700	33,40	3,31	130	108	0,373	20	60	102	135	184	226
07342350	438	0,34	119	99,1	0,078	41	80	125	155	193	221
07344320	3,70	6,88	117	89,9	0,627	23	8	14	18	24	27
07346800	0,34	20,60	117	79,2	0,415	20	1	1	2	2	3
07347000	300	0,80	119	73,2	0,157	25	40	58	70	85	98
07348615	593	0,55	127	93,3	0,200	12	120	250	363	533	681
07348630	0,13	19,70	127	99,1	0,640	22	1	1	2	2	3
07348700	1570	0,66	127	88,4	0,265	36	206	411	577	817	1010
07349430	606	0,67	127	97,5	0,278	24	93	185	261	372	463
07349500	1410	0,34	124	68,9	0,112	23	124	220	290	384	457
07355800	1,68	30,10	130	347	0,156	33	6	8	11	14	16
07355900	0,41	34,90	142	366	0,190	20	1	2	3	4	4
07356000	1070	1,48	132	354	0,104	52	649	1070	1370	1780	2100
07356500	166	2,92	135	253	0,154	29	192	329	429	561	663
07356700	4,79	15,60	135	232	0,220	23	12	24	35	51	64
07357501	2860	0,98	132	268	0,082	16	1080	1750	2240	2910	3430
07357700	9,95	13,70	142	247	0,326	26	18	37	53	76	95
07359500	4110	0,84	142	247	0,058	30	1530	2510	3170	4000	4600

Suite page suivante

TAB. D.1: Données de l'Arkansas (suite)

NO	A	S	P	EL	SH	n	Q_2	Q_5	Q_{10}	Q_{25}	Q_{50}
07359520	7,77	13,70	140	162	0,115	20	8	19	31	50	68
07359700	497	3,54	140	296	0,197	35	597	1080	1440	1940	2340
07359750	6,06	16,40	147	226	0,443	22	23	47	66	94	116
07359800	808	2,33	142	265	0,127	33	734	1120	1380	1700	1940
07360150	1,04	9,87	132	104	0,443	21	2	5	8	12	17
07360800	313	3,64	135	171	0,274	41	310	531	702	945	1140
07361000	984	2,92	140	232	0,127	17	778	1440	1970	2720	3330
07361020	0,41	41,10	137	194	0,480	24	2	4	6	9	12
07361180	45,80	3	132	122	0,308	30	114	156	182	212	234
07361200	383	1,76	132	128	0,237	41	208	359	476	638	769
07361500	461	1,58	132	158	0,145	45	357	541	666	827	948
07361600	2770	1,45	137	165	0,103	11	926	1670	2220	2980	3570
07361680	3,83	9,07	130	109	0,598	26	6	13	18	26	33
07361780	8,96	4,15	132	97,5	0,253	20	13	21	26	33	39
07361800	648	1,40	132	90,2	0,121	41	480	679	814	988	1120
07362050	26,90	3,31	127	61,6	0,262	21	11	26	39	60	80
07362100	997	0,75	127	70,1	0,268	55	189	393	574	858	1110
07362330	35,20	2,12	127	57	0,248	32	24	50	72	105	133
07362450	12,90	3,79	130	96	0,554	20	19	37	50	69	84
07362500	622	1,05	130	82,3	0,237	43	139	276	380	523	634
07363000	1420	2,35	137	198	0,179	57	866	1410	1780	2260	2610
07363050	3,73	8,90	135	105	0,592	24	5	11	18	29	41
07363200	2920	0,85	137	140	0,076	56	694	1140	1470	1910	2260
07363300	528	1,30	135	114	0,120	34	201	406	563	776	941
07363330	12,60	4,20	135	88,4	0,375	22	13	27	40	58	73
07363430	1,66	12,20	130	107	0,423	21	3	7	11	17	23
07363450	0,73	12	130	74,1	0,440	23	1	3	5	7	9
07363500	5440	0,46	132	110	0,063	57	676	1190	1560	2060	2450
07364030	0,93	9,15	132	51,8	0,523	22	1	2	3	5	6
07364070	14,60	2,88	132	50,3	0,348	21	10	15	19	24	28
07364260	54,10	1,23	137	51,8	0,237	22	19	36	48	64	76

Suite page suivante

TAB. D.1: Données de l'Arkansas (suite)

NO	A	S	P	EL	SH	n	Q_2	Q_5	Q_{10}	Q_{25}	Q_{50}
07364300	702	0,63	137	43,3	0,366	24	138	294	426	618	778
07364700	365	0,68	122	61,9	0,260	22	70	158	248	412	579
07365800	466	0,96	124	76,2	0,268	38	136	322	506	819	1120
07365900	130	1,17	122	74,7	0,257	23	56	113	165	245	318
07366000	1200	0,66	122	56,4	0,213	43	168	325	464	685	885
07366200	539	0,70	122	61,3	0,239	38	108	232	342	513	664
07047820	3,57	6,30	122	98	0,170	34	15	23	28	35	40
07047880	0,21	37,90	124	113	0,690	23	1	2	3	4	6
07047975	3,19	51,30	122	607	0,327	21	6	11	15	21	26
07047990	1,74	54,90	112	463	0,283	27	5	11	17	25	32
07048000	215	5,20	114	521	0,188	38	247	469	647	906	1120
07048600	1040	2,70	114	488	0,193	30	683	1210	1620	2170	2620
07048900	2,77	19,90	109	421	0,633	34	4	8	11	16	21
07048940	58	8,10	122	585	0,365	22	85	170	240	343	430
07049000	679	1,60	114	485	0,117	35	394	686	895	1170	1370
07049500	2640	1,10	112	457	0,121	13	780	1530	2120	2940	3590
07050000	3210	0,70	112	442	0,062	37	847	1370	1760	2310	2760
07050200	7,12	26,70	114	497	0,477	21	17	37	54	79	101
07050400	1,89	24,40	109	415	0,383	20	5	9	12	16	18
07050500	1360	1,30	109	469	0,077	55	500	932	1280	1770	2180
07054400	8,83	21,20	107	317	0,224	21	33	57	76	105	129
07054450	2,20	40,90	107	323	0,354	32	8	14	20	29	37
07055550	11,30	10,60	114	390	0,302	26	17	31	43	61	78
07055650	21,60	26	119	579	0,214	21	41	94	142	214	276
07055800	15,90	44,10	114	460	0,327	22	31	64	93	136	173
07056000	2150	2	114	454	0,089	54	1060	1910	2530	3360	4000
07057000	2840	1,30	112	421	0,065	40	1090	2020	2740	3780	4630
07057300	1,97	22,40	109	259	0,229	26	8	13	17	22	25
07059000	4180	1,20	102	396	0,136	15	703	1190	1530	1950	2250
07060600	3,24	16,20	112	225	0,313	26	7	12	17	25	32
07060670	8,29	39,60	119	259	0,392	21	24	37	46	59	69

Suite page suivante

TAB. D.1: Données de l'Arkansas (suite)

NO	A	S	P	EL	SH	n	Q_2	Q_5	Q_{10}	Q_{25}	Q_{50}
07060710	150	2,90	114	198	0,234	28	122	290	436	652	830
07060830	0,70	18,50	124	320	0,360	21	2	3	4	7	9
07061100	10,10	8,80	119	140	0,290	25	23	41	55	75	92
07068000	5280	0,90	112	305	0,104	66	754	1390	1870	2530	3060
07068870	0,49	30,50	117	139	0,411	21	4	6	7	9	10
07068890	593	2	117	155	0,168	15	441	817	1100	1480	1780
07069250	1,24	29,60	112	209	0,532	25	9	15	20	27	32
07069290	5,91	14,70	112	240	0,406	21	15	28	38	52	63
07069500	3060	1,60	112	226	0,136	57	742	1390	1930	2750	3470
07071500	2050	1,90	112	305	0,125	71	265	571	815	1150	1420
07072000	2940	1,90	109	259	0,860	63	335	630	897	1330	1730
07072200	3,44	9,60	119	143	0,230	25	17	23	27	32	36
07073000	562	1,10	112	226	0,067	41	259	420	537	697	822
07073500	257	1,40	114	204	0,088	55	133	242	332	468	585
07074000	1230	1,10	114	207	0,080	58	421	760	1040	1460	1820
07074200	3,16	11,60	114	198	0,408	23	17	29	37	47	53
07074250	90,40	3,20	117	134	0,171	21	84	164	232	336	428
07074900	0,67	115	112	378	0,294	26	3	5	6	8	9
07074950	4,09	22	114	463	0,399	23	9	18	26	38	48
07075000	782	2,60	114	354	0,075	55	635	1190	1650	2350	2940
07075300	383	4	127	351	0,066	32	295	554	759	1050	1290
07075500	818	3,10	127	351	0,175	23	629	985	1220	1520	1740
07075600	3,52	19,50	127	208	0,244	30	7	13	18	26	33
07075800	0,67	37,50	127	252	0,400	30	1	3	4	6	8
07076000	2990	1,90	122	317	0,105	35	1540	2110	2440	2840	3110
07076630	1,71	16,10	127	101	0,528	33	7	11	13	15	16
07076820	12,90	6,10	127	93	0,249	21	22	32	39	46	52
07076850	430	0,50	127	101	0,131	15	177	309	403	526	618
07076870	59,60	1,30	127	79	0,283	33	61	114	156	217	269
07077100	33,10	3,90	119	129	0,604	20	84	122	145	170	188
07077200	4,09	7,70	119	137	0,193	32	10	16	19	23	26

Suite page suivante

TAB. D.1: Données de l'Arkansas (suite)

NO	A	S	P	EL	SH	n	Q_2	Q_5	Q_{10}	Q_{25}	Q_{50}
07077340	1,76	17,30	122	140	0,333	24	8	13	16	20	24
07047200	5,59	0,23	124	66	0,198	24	5	6	6	7	7
07047600	751	0,13	122	70	0,096	54	136	180	207	239	262
07047924	1,24	0,80	127	61	0,333	20	3	6	8	11	14
07047942	1390	16	124	75	0,156	23	183	332	455	636	791
07047950	2040	16	124	73	0,123	54	252	378	458	553	619
07063000	3220	1,18	112	274	0,094	26	374	788	1160	1750	2280
07064000	4530	0,68	109	223	0,053	34	345	603	795	1060	1260
07074550	16,20	0,35	122	83	0,272	21	6	14	22	34	46
07074855	14,30	0,63	124	67	0,154	20	9	14	17	22	25
07077380	1820	0,18	122	91	0,133	56	124	163	189	223	249
07077430	1,24	0,76	122	78	0,333	31	1	2	3	4	5
07077500	2690	0,16	122	79	0,054	66	181	261	318	392	449
07077680	20,50	0,25	124	70	0,147	20	8	11	13	14	15
07077700	1090	0,44	124	73	0,085	48	91	121	139	161	176
07077860	25,90	0,36	127	55	0,192	22	10	13	15	18	19
07077920	80,50	0,17	127	64	0,103	33	16	22	26	30	32
07077940	98,40	0,27	127	61	0,314	20	38	52	59	68	73
07077950	997	0,13	127	59	0,085	23	94	132	154	179	196
07078000	453	0,16	127	61	0,087	19	68	113	144	182	211
07078170	3,91	0,63	127	59	0,082	20	5	6	7	7	8
07078210	0,52	3,47	127	56	0,220	24	2	3	4	6	8
07263860	7,12	0,66	127	65	0,295	37	10	14	16	19	20
07264000	536	0,25	127	91	0,075	39	61	87	104	128	146
07264100	21,80	0,99	127	72	0,438	26	24	36	42	49	54
07364110	1,94	6,93	130	81	0,415	33	4	7	10	14	18
07364120	557	0,11	130	67	0,032	43	49	69	80	94	103
07364125	12,70	7,73	130	94	0,533	22	29	44	53	63	70
07364150	1490	0,10	130	58	0,021	54	93	127	147	170	186
07364165	48,70	2,56	132	86	0,418	21	26	46	62	89	112
07364190	3030	0,08	132	55	0,016	55	135	169	189	213	229

Suite page suivante

TAB. D.1: Données de l'Arkansas (suite)

NO	A	S	P	EL	SH	n	Q_2	Q_5	Q_{10}	Q_{25}	Q_{50}
07364200	3070	0,08	132	64	0,015	36	132	177	202	229	246
07364500	4260	0,08	132	60	0,016	53	199	256	290	331	359
07367658	2,43	0,69	130	49	0,377	26	4	6	7	8	9
07367740	4,82	0,86	137	34	0,267	23	7	8	9	11	11

Résultats de la modélisation log-linéaire

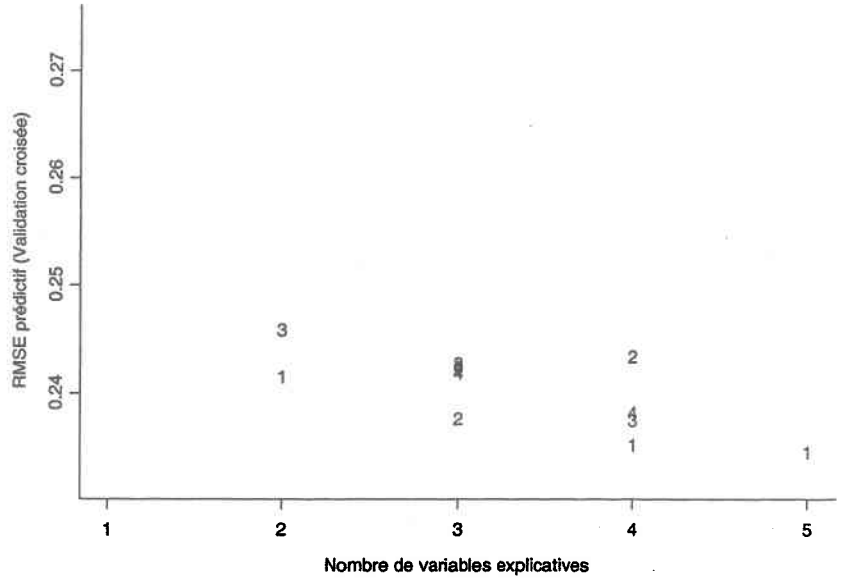

FIG. D.1: RMSE de différents modèles de régression log-linéaire pour l'estimation de Q_2 (voir la légende au tableau 5.3)

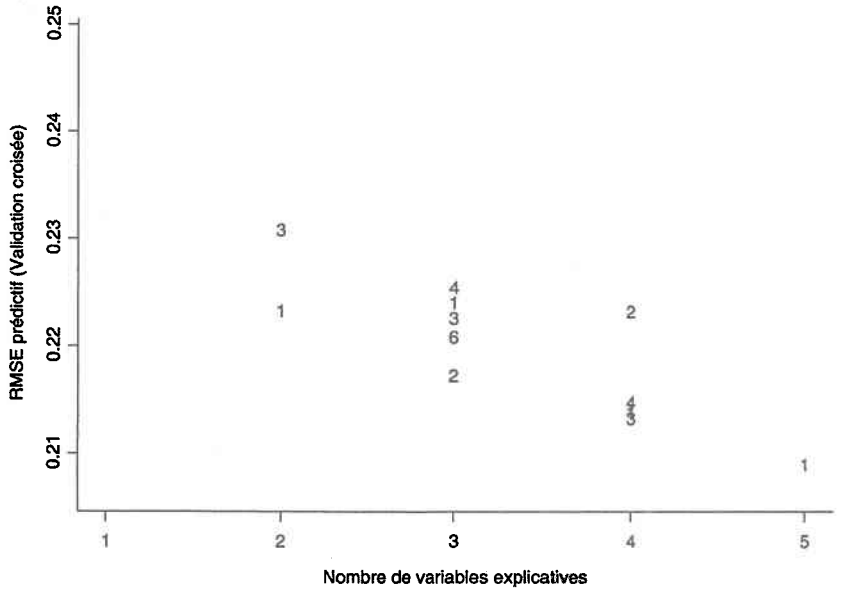

FIG. D.2: RMSE de différents modèles de régression log-linéaire pour l'estimation de Q_5 (voir la légende au tableau 5.3)

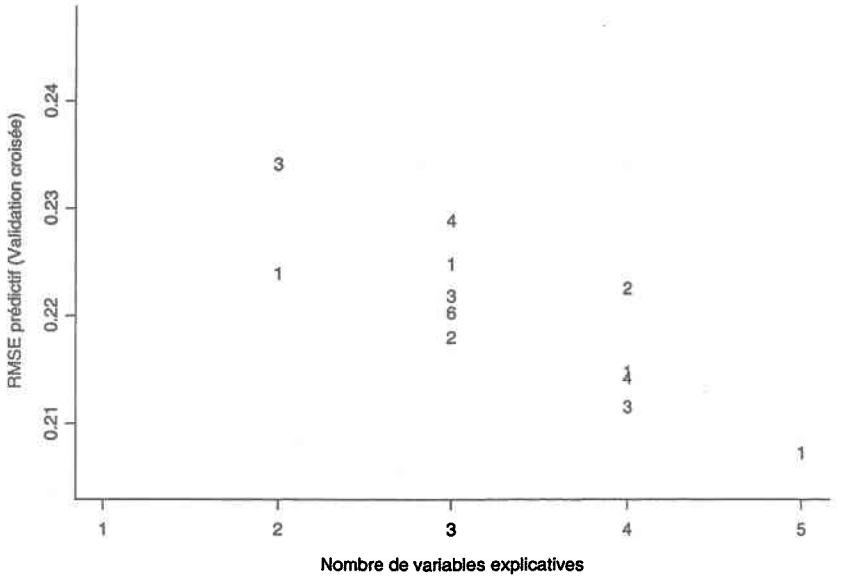

FIG. D.3: RMSE de différents modèles de régression log-linéaire pour l'estimation de Q_{10} (voir la légende au tableau 5.3)

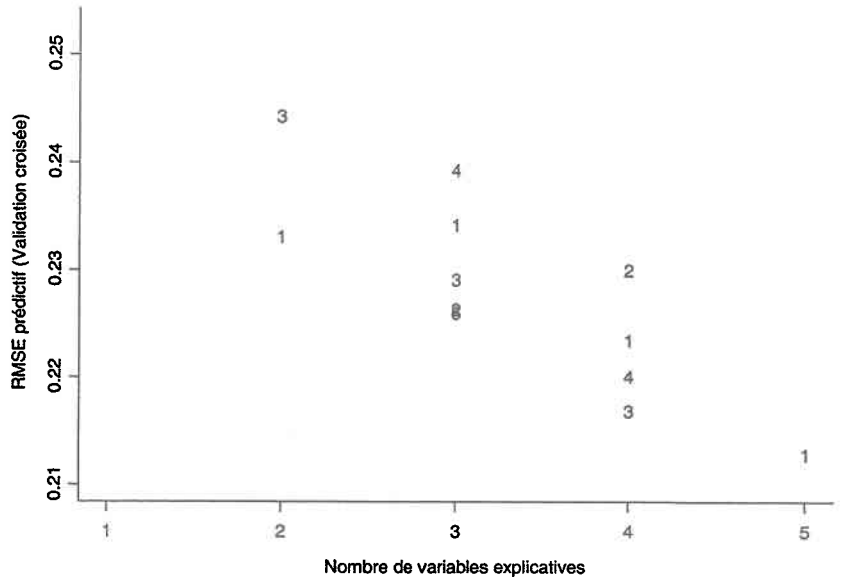

FIG. D.4: RMSE de différents modèles de régression log-linéaire pour l'estimation de Q_{25} (voir la légende au tableau 5.3)

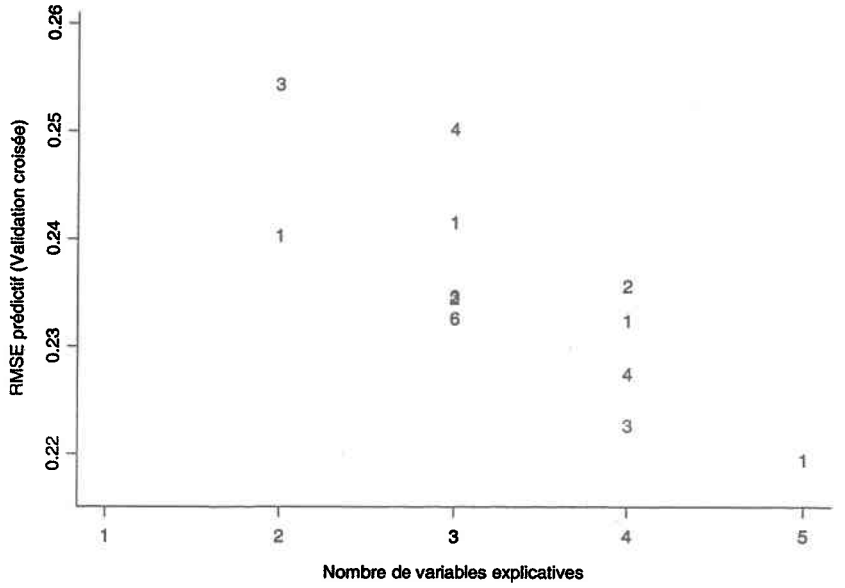

FIG. D.5: RMSE de différents modèles de régression log-linéaire pour l'estimation de Q_{50} (voir la légende au tableau 5.3)

Résultats de la modélisation par région d'influence

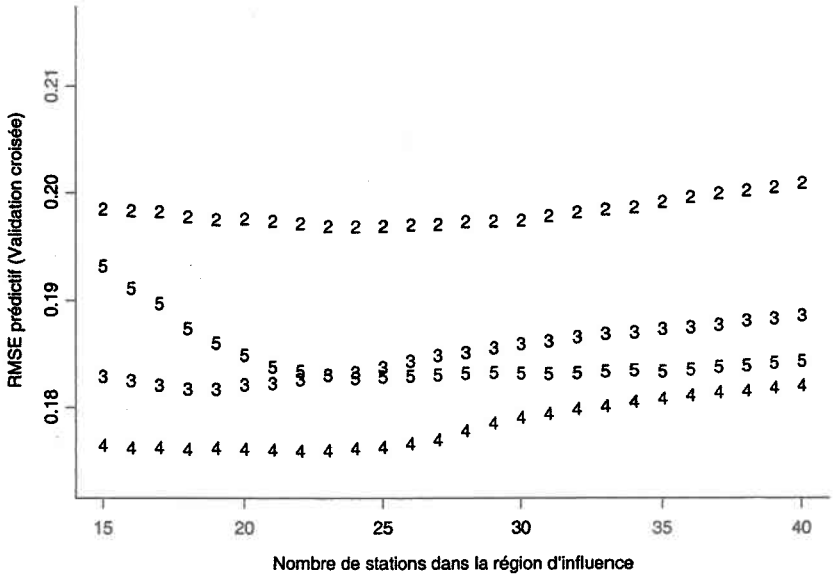

FIG. D.6: Nombre optimal de stations dans la région d'influence pour Q_2

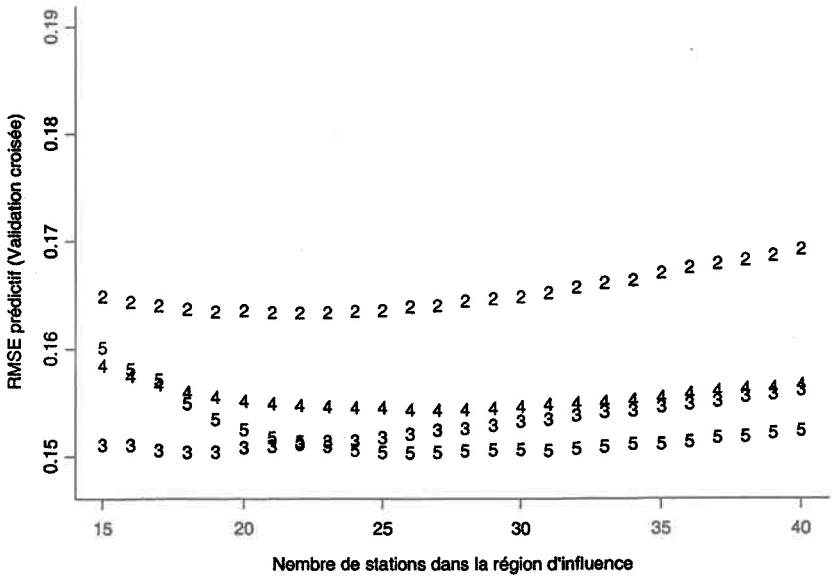

FIG. D.7: Nombre optimal de stations dans la région d'influence pour Q_5

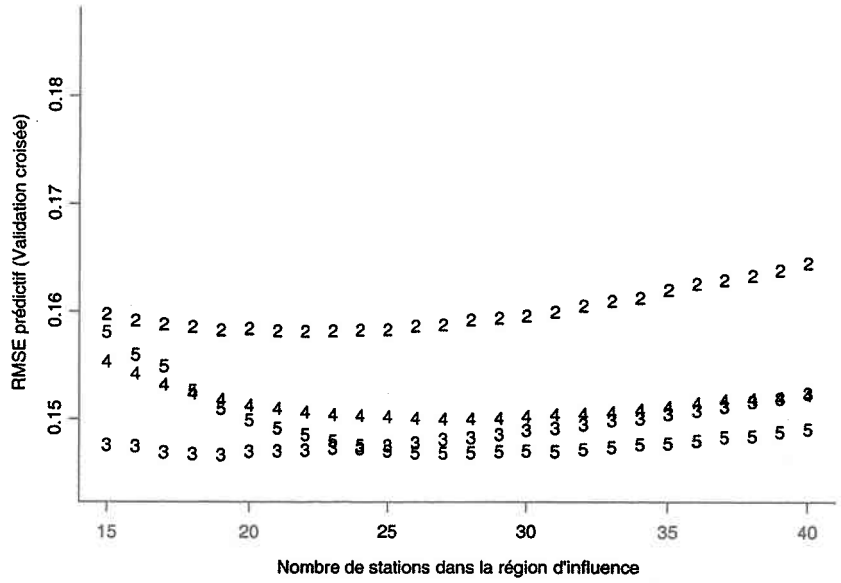

FIG. D.8: Nombre optimal de stations dans la région d'influence pour Q_{10}

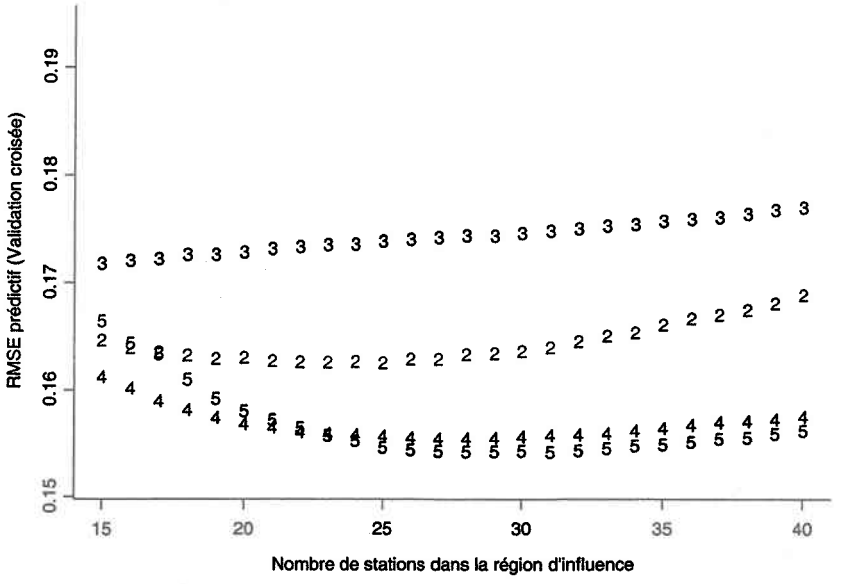

FIG. D.9: Nombre optimal de stations dans la région d'influence pour Q_{25}

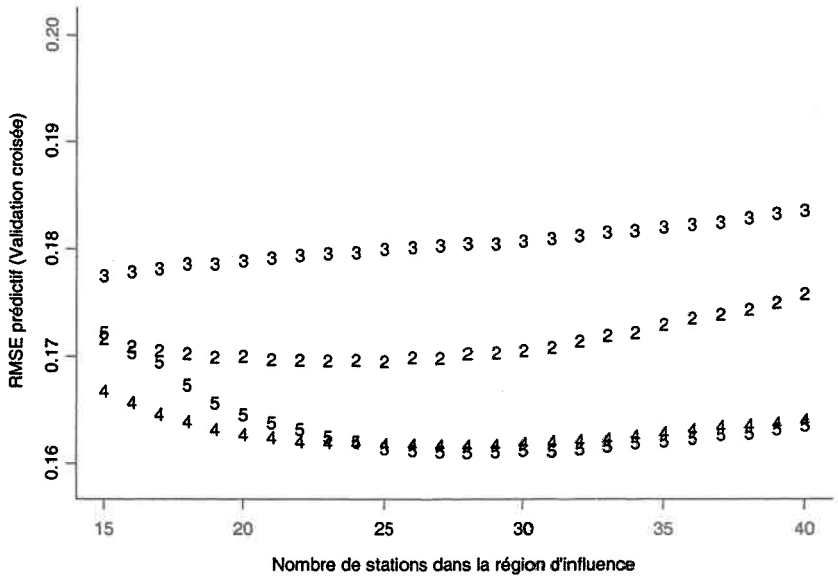

FIG. D.10: Nombre optimal de stations dans la région d'influence pour Q_{50}

Résultats de la modélisation additive

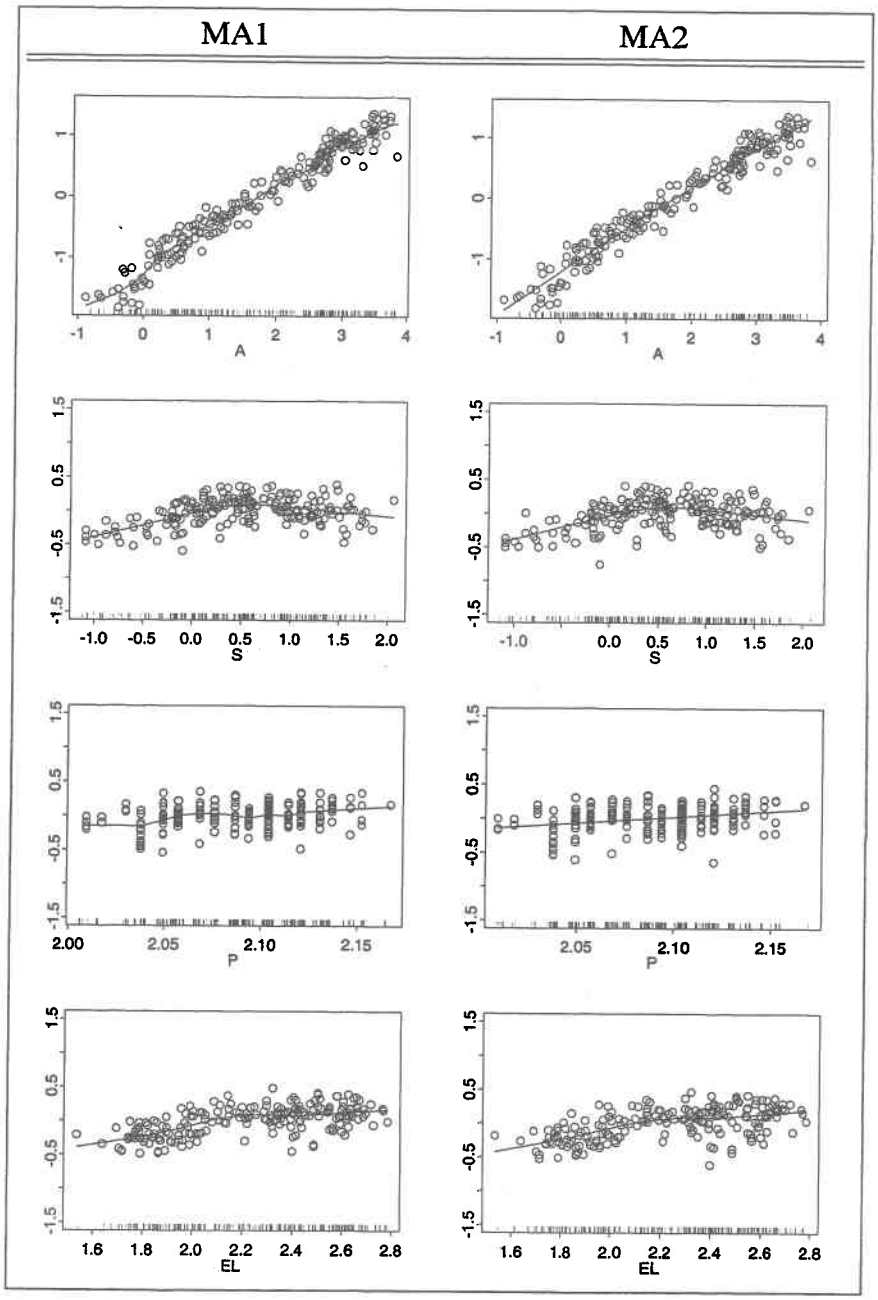

FIG. D.11: Modélisation additive de Q_2

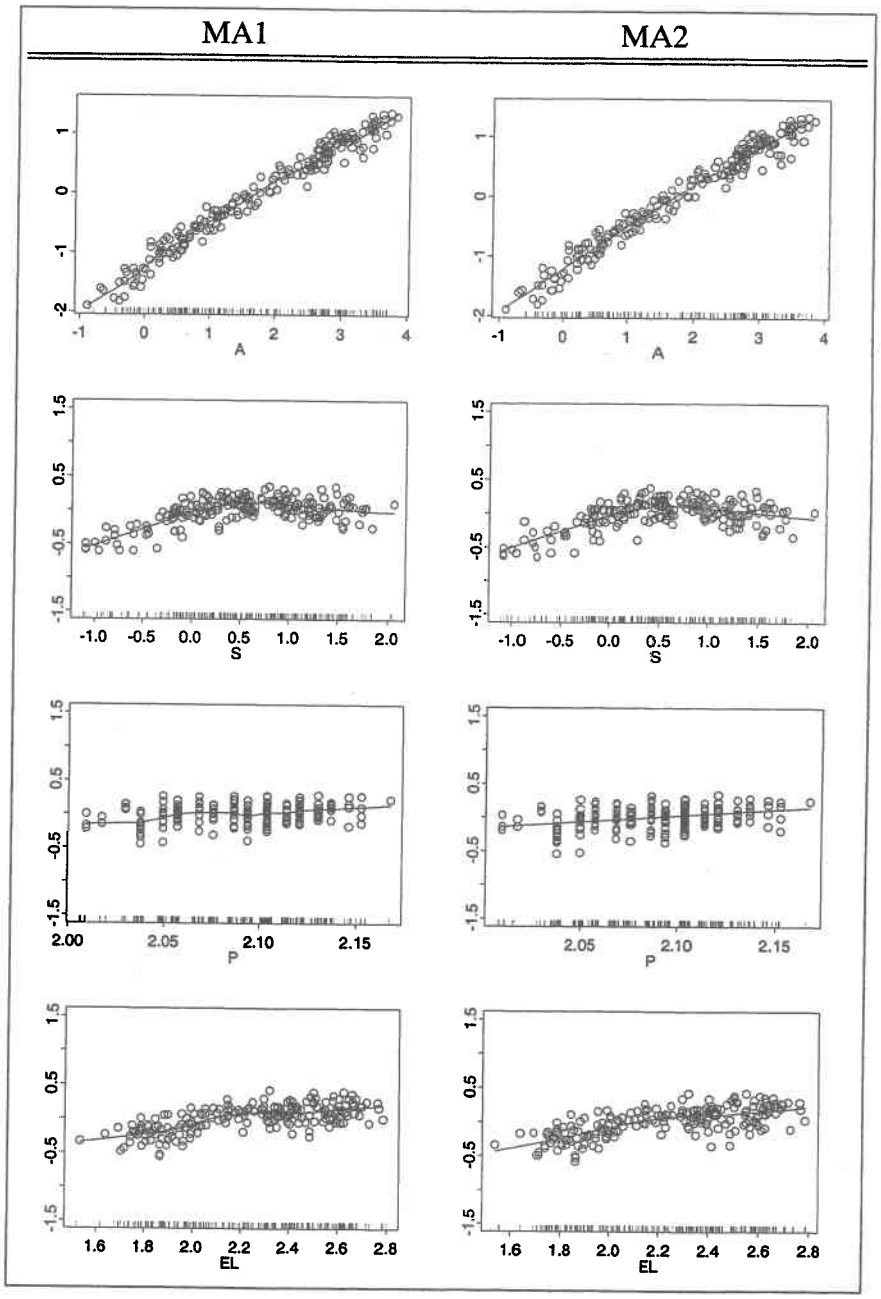

FIG. D.12: Modélisation additive de Q_5

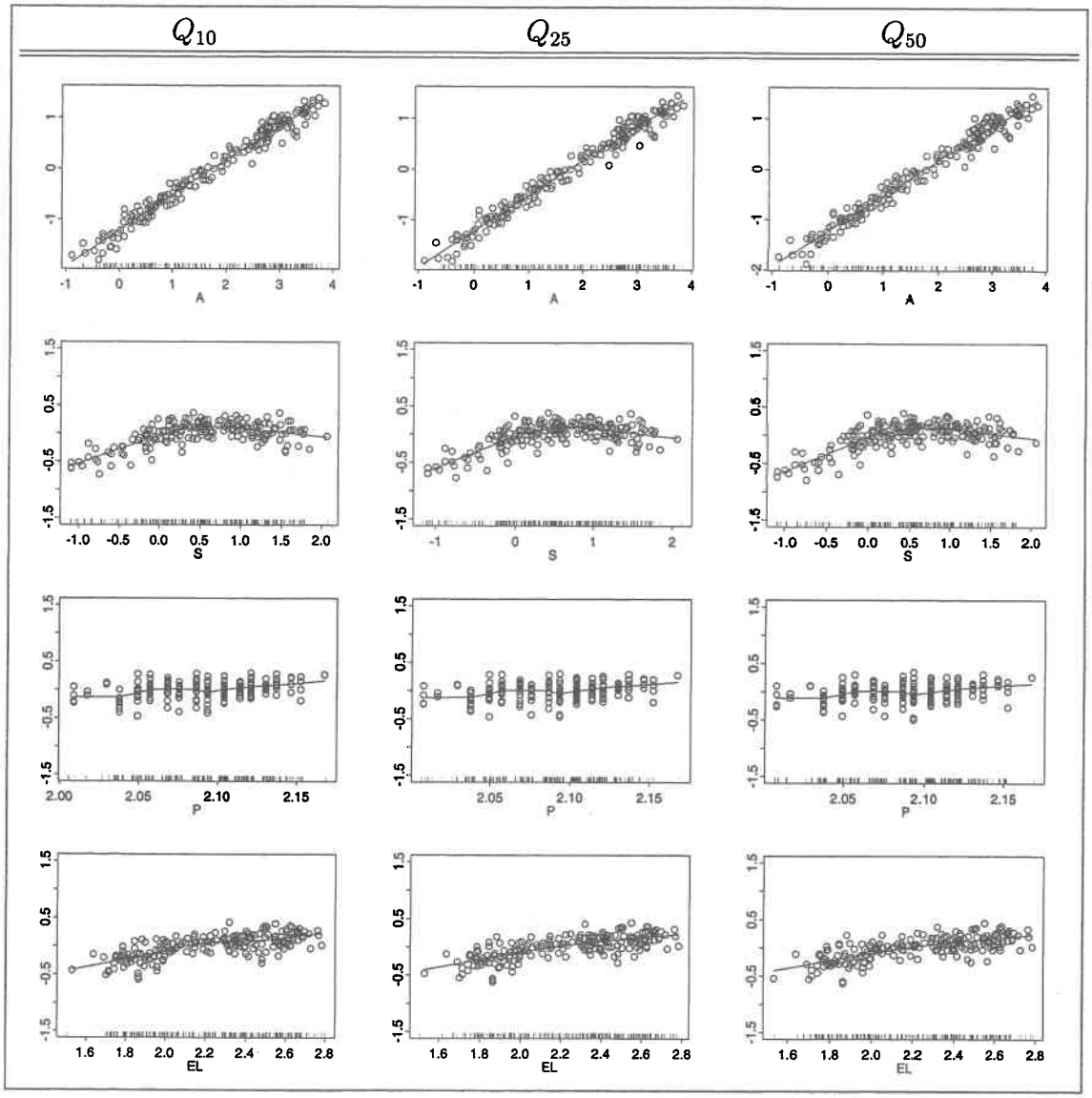

FIG. D.13: Modélisation additive ($c = 1$) de Q_{10}, Q_{25} et Q_{50}