

UNIVERSITE CATHOLIQUE DE LOUVAIN
Faculté des Sciences Economiques, Sociales et Politiques
Département des Sciences Economiques

**Une méthode simple et flexible de réduction du
biais des estimateurs de maximum de
vraisemblance des distributions GEV avec
applications aux régressions hédoniques et à
d'autres processus à valeurs extrêmes**

Vincent Scourneau

*Thèse présentée en vue de l'obtention
du grade de docteur en sciences économiques et de gestion*

Composition du jury

Promoteur :

Professeur Daniel Weiserbs (Université catholique de Louvain)

Autres Membres du jury :

Professeur Luc Bauwens (Université catholique de Louvain)

Professeur Michel Denuit (Université catholique de Louvain)

Professeur Russell Davidson (Université McGill)

Louvain-la-Neuve, mars 2011

Remerciements

Avant tout, je tiens à exprimer ma gratitude envers Daniel Weiserbs pour son soutien plus que précieux dans l'accomplissement de cette thèse. En plus de m'avoir encadré de manière merveilleuse tout au long de mon parcours, il n'a eu de cesse, de par son enthousiasme, de tenter de me transmettre sa passion pour la recherche. Je n'oublierai pas les très nombreux moments agréables passés ensemble.

Je voudrais également remercier les membres de mon jury pour leur relecture attentive et les conseils judicieux afin d'améliorer sensiblement la qualité de ma thèse. En particulier, je remercie Russell Davidson pour ses conseils éclairés à propos de la technique du bootstrap et pour ses commentaires très constructifs qui ont permis de compléter les résultats du premier chapitre. Je remercie Michel Denuit pour ses remarques très perspicaces qui ont contribué grandement à l'amélioration du cinquième chapitre. Enfin, je remercie Luc Bauwens pour sa relecture minutieuse et ses conseils prodigués dans un objectif constant de recherche de l'excellence.

Je tiens aussi à remercier les participants au séminaire de statistiques appliquées donné à l'institut de statistiques de l'UCL en décembre 2010, en particulier Johan Segers, pour leurs commentaires avisés.

Je remercie les personnes qui m'ont fourni les données utilisées dans les différentes applications de cette thèse. Il s'agit, pour les données hydrologiques, de M. P. Dewil, directeur du Service d'études hydrologiques (SETHY) du Ministère wallon de l'Équipement et des Transports, et de Mme. D. Pontegnie; de M. P. Jassogne pour les données concernant les grands vins du bordelais; des membres du service Inforville pour les loyers de Louvain-la-neuve et de M. J. Anciaux pour les données sur les coûts de sinistres automobiles.

J'aimerais également rendre hommage aux nombreuses personnes qui ont agrémenté ces six années passées à l'IRES. Il est difficile d'être exhaustif, mais je tiens à remercier tout particulièrement les membres, actuels et passés, du Service d'analyses économiques, à savoir Anne, Catherine, Cécile, Géraldine,

Guy, Jean-François, Hélène, Philippe. En outre, je remercie Vincent Bodart pour la formation de grande qualité qu'il m'a inculquée en analyse conjoncturelle, qui sera problemement très précieuse dans ma vie professionnelle future.

Je voudrais aussi remercier Anne Davister pour la bienveillance dont elle fait preuve constamment et pour le soutien logistique indispensable qu'elle m'a offert à de nombreuses reprises.

Je remercie Vanessa qui m'est venue très gentilleement en aide à plusieurs reprises lorsque je pataugeais face à des problèmes informatiques liés à la mise en page de cette thèse.

Enfin, je dédie ce travail à ma famille qui m'a toujours soutenu tout au long de ma scolarité et je remercie Audrey pour sa compréhension et son soutien, éléments essentiels à l'accomplissement de cette thèse. Audrey, tu enchantes ma vie à chaque instant et je suis très heureux de m'engager prochainement avec toi dans une belle aventure.

Table des matières

Introduction générale	1
1 Méthodologie	5
1.1 Introduction	5
1.2 Le modèle et les techniques d'estimation	8
1.2.1 La distribution GEV	8
1.2.2 Estimation par la méthode traditionnelle	9
1.2.3 Méthode d'estimation alternative (AGEV)	13
1.2.4 Interprétation des méthodes	15
1.2.5 Autres méthodes d'estimation	17
1.3 Comparaison des méthodes par simulations de Monte Carlo . .	17
1.3.1 Echantillons	17
1.3.2 Choix du paramètre α	18
1.3.3 Ajustement global	21
1.3.4 Biais et variances des estimateurs	21
1.3.5 Biais des quantiles extrêmes	22
1.3.6 Robustesse	24
1.3.7 Cas Gumbel	25
1.4 Extension 1 : régression multivariée	28
1.5 Extension 2 : données de type panel	29
1.5.1 Méthode classique d'estimation	29
1.5.2 Méthode d'estimation alternative	29
1.5.3 Résultats de la méthode alternative sur base de simulations de Monte Carlo	30
1.6 Conclusion	31
1.A Méthode des moments pondérés	32
1.B Résultats des simulations avec $\alpha = 0$	34
2 Application aux crues du bassin de l'Ourthe	37
2.1 Introduction	37
2.2 Les données	38
2.3 Estimations indépendantes (1975-2004)	39
2.4 Estimation jointe, méthode classique	41

2.5	AGEV, estimation simultanée	43
2.6	AGEV, estimation avec contraintes	44
2.7	Estimations pour la période 1975-2009	46
2.8	Quantiles extrêmes	48
2.9	Conclusion	50
2.A	Q-Q plots	51
3	Application aux prix hédoniques : les grands vins	55
3.1	Introduction	55
3.2	Les données	56
3.3	Approche standard	58
3.3.1	Le modèle général	58
3.3.2	Le modèle final	59
3.3.3	Les résultats	60
3.3.4	Qualité d'ajustement	62
3.4	Prix hédoniques et distribution GEV	63
3.4.1	Modèle GEV estimé par la méthode classique	63
3.4.2	Modèle GEV estimé par la méthode alternative	64
3.4.3	Comparaison des résultats	64
3.5	Conclusion	67
4	Application aux prix hédoniques : les loyers	69
4.1	Introduction	69
4.2	Les données	71
4.3	Approche classique	72
4.3.1	Le modèle	72
4.3.2	Les résultats	73
4.4	Distribution GEV	75
4.5	Conclusion	79
5	Application aux coûts des sinistres automobiles	81
5.1	Introduction	81
5.2	Les données	83
5.3	Fréquence des accidents	85
5.3.1	Méthode	85
5.3.2	Application	86
5.4	Coût des sinistres	89
5.4.1	Sélection de l'échantillon	89
5.4.2	Régression classique	90
5.4.3	Le modèle GEV	90
5.4.4	Distribution de Burr généralisée	94
5.5	Lien entre l'analyse des coûts et des fréquences	97
5.6	Conclusion	98
	Bibliographie	101

Introduction générale

Dans bon nombre d'applications statistiques, la variable dépendante peut prendre des valeurs extrêmes. La littérature économique empirique, et en ce qui nous concerne particulièrement celle sur les prix hédoniques, ignore généralement ce phénomène, vraisemblablement sous l'hypothèse que la distribution asymétrique de l'endogène n'empêche pas d'estimer correctement les paramètres d'intérêt. Les auteurs se contentent dès lors de tests standards avec toutefois des écarts-types corrigés pour l'hétéroscédasticité apparente des résidus¹. Or, dans bien des cas, c'est pour les quantiles les plus élevés (ou les plus faibles) que l'inférence statistique est particulièrement intéressante et c'est pour ces quantiles que l'ajustement par une régression ordinaire est le moins bon².

L'estimation de distributions avec queue lourde est courante dans d'autres sciences et peut même, comme en hydrologie et en climatologie par exemple, constituer la règle. Reposant sur le théorème des valeurs extrêmes (Fisher et Tippett, 1928), la distribution GEV (Generalized Extreme Value Distribution) introduite par Jenkinson (1955) est utilisée pour modéliser une série de maxima observés sur des intervalles donnés, tandis que la Pareto généralisée permet de modéliser les valeurs d'une série statistique excédant un certain seuil. La démarche classique consiste à estimer directement les paramètres de ces distributions par la méthode du maximum de vraisemblance et à procéder ensuite à l'évaluation de l'ajustement statistique en utilisant une définition particulière des fréquences empiriques des observations. Mais, avec des petits échantillons, les estimateurs du maximum de vraisemblance sont biaisés et il peut en résulter une sous-estimation conséquente de la probabilité d'un événement rare contre lequel on voudrait se prémunir.

1. Il peut arriver aussi que le chercheur sur base d'une analyse préalable des données écarte des observations jugées atypiques alors que, dans ce contexte, elles ne sont peut-être pas aberrantes.

2. Notons toutefois le développement ces dernières années d'études empiriques traitant des valeurs extrêmes en économie financière et en sciences actuarielles (voir par exemple Embrechts *et al.*, 1997).

Cette caractéristique de l’approche classique est bien connue et de nombreux auteurs recourent dès lors soit à une autre méthode d’estimation – la plus commune étant la méthode des moments pondérés – soit à d’autres formes fonctionnelles. Dans ce dernier cas, on renonce aux fondements théoriques de la GEV. Quant aux méthodes des moments, s’ils conduisent à des estimateurs plus efficaces pour de petits échantillons, la réduction des biais n’est pas assurée et, surtout, ils sont peu pratiques lorsque d’autres variables sont incluses dans le modèle; ils deviennent même inapplicables lorsqu’on veut tenir compte de relations de simultanéité.

Nous proposons dans cette thèse une méthode d’estimation alternative de la GEV par maximum de vraisemblance. En petit échantillon, elle améliore sensiblement la qualité de l’ajustement global et aboutit à une réduction très considérable des biais. Elle présente, en outre, une série d’avantages, à la fois sur le plan numérique et statistique. Elle reste fort simple, mis à part la nécessité de recourir à la technique du bootstrapping pour évaluer les intervalles de confiance. Dans cette recherche, nous l’analysons dans le cas des distributions de Weibull et Gumbel (deux des trois possibilités de la GEV) et pour des applications dans le champ de l’économie et de l’hydrologie. Cependant, *mutatis mutandis*, elle devrait pouvoir être utilisée dans d’autres domaines et pour d’autres distributions, pour autant qu’il y ait une expression explicite de la fonction de répartition. Un exemple particulier est la distribution de Burr généralisée qui a été testée dans les chapitres trois, quatre et cinq. Cette distribution de Burr affiche un meilleur ajustement que la GEV uniquement dans le cinquième chapitre consacré aux coûts de sinistres automobiles et les résultats des estimations de cette distribution réalisées par les deux méthodes y sont illustrés.

Le premier chapitre présente la distribution GEV ainsi que les méthodes d’estimation utilisées traditionnellement dans la littérature empirique. Ce chapitre introduit la méthode d’estimation alternative de la distribution GEV, appelée AGEV, et en compare les performances avec celles de la méthode d’estimation classique par le biais d’une analyse de Monte-carlo. Les propriétés des deux méthodes sont analysées pour des échantillons de taille faible et modérée, à la fois dans le cas univarié et dans le cas de données en panel.

Le domaine d’application le plus évident est sans conteste celui des sciences naturelles, en particulier l’hydrologie et la climatologie, disciplines pour lesquelles la caractérisation d’événements de probabilité (très) faible est d’importance cruciale. Aussi, dans le deuxième chapitre, nous appliquons la méthode d’estimation alternative de la distribution GEV à des données hydrologiques et nous comparons les résultats avec ceux obtenus par la méthode d’estimation classique.

Les données hydrologiques sont observées simultanément à différents endroits d'un même bassin hydrologique, voire d'un même cours d'eau. La modélisation de telles données requiert d'étendre la distribution GEV de deux manières. D'une part, il est important de tenir compte de toute l'information disponible plutôt que de considérer les sites de manière séparée. D'autre part, il peut s'avérer important de tenir compte de la dépendance contemporaine entre les observations des différents points de l'espace. La méthode d'estimation traditionnelle permet de tenir compte de l'hétérogénéité des observations à travers l'espace, mais ne permet pas de prendre en compte les covariances contemporaines. De ce point de vue, la méthode d'estimation alternative s'avère beaucoup plus satisfaisante puisqu'elle offre la possibilité d'estimer des distributions GEV propre à chaque site sous forme d'un système avec une matrice de variance covariance contemporaine des résidus non diagonale.

Dans le domaine de l'économie, l'estimation de prix hédoniques constitue certainement un type d'application tout désigné. En effet, pour bon nombre de biens, le prix d'une variété, même corrigé pour ses caractéristiques, peut prendre des valeurs fort élevées. Dans les troisième et quatrième chapitres, nous utilisons la distribution GEV pour modéliser, respectivement, les prix des "grands vins" de la région de Bordeaux et les loyers des biens immobiliers résidentiels dans la ville de Louvain-la-Neuve.

Aussi bien les prix des vins que les loyers affichent une importante variation, avec quelques observations qui prennent des valeurs "extrêmes". L'objectif ici est de vérifier si, conformément à la théorie, il existe des marchés implicites pour les différentes caractéristiques. Les résultats révèlent des informations économiques intéressantes. Sur le marché du vin, nous montrons entre autres l'importance des effets de réputation des producteurs, l'influence du classement officiel et les primes reçues par certaines appellations. Pour le marché immobilier, nous montrons notamment qu'il existe des sous-marchés homogènes au sein de l'espace urbain, même dans le cas d'une petite ville comme Louvain-la-Neuve et que, de manière générale, les loyers ont augmenté plus rapidement que l'indice des prix sur lequel ils sont indexés.

Il existe bien sûr d'autres champs d'application en économie, notamment dans le domaine de la finance. On sait en effet qu'une des nombreuses causes de la récente crise financière réside dans l'incapacité qu'ont affichée les gestionnaires à modéliser les risques des différents instruments financiers complexes de manière correcte³. Dans le cinquième chapitre, nous proposons une application de la méthode dans le domaine des assurances où l'appréciation correcte du risque de survenance d'un événement rare au coût élevé est également de grande importance. Nous modélisons le coût de sinistres automobiles supportés par une compagnie d'assurance en tenant compte des différentes caractéristiques des assurés. Même si le coût pour la compagnie d'assurance dépend principalement des caractéristiques de la tierce partie impliquée dans l'accident, nous montrons que certains facteurs propres à l'assuré, tels que le type de couverture choisi ou le degré d'urbanisation de son lieu de résidence peuvent également influencer le coût supporté par la compagnie d'assurance en cas d'accident. Notre analyse porte sur un échantillon original dont l'intérêt principal est la taille (plus de 10 000 observations) et la richesse des informations qu'il contient sur les assurés et leur véhicule. Tout comme dans le chapitre précédent, où l'échantillon est également d'une taille importante (730 observations), les estimations sont fort proches pour les deux méthodes. Mais, par rapport à des régressions ordinaires, le gain dans l'ajustement statistique réalisé grâce à l'utilisation de la fonction GEV est spectaculaire. On montre aussi qu'une distribution de Burr généralisée, qui possède un paramètre de courbure additionnel par rapport à la GEV, offre un gain supplémentaire en précision dans cette application.

3. La plupart des gestionnaires de risque ont basé leur analyse sur des techniques statistiques classiques (par exemple, le modèle de Black-Scholes, utilisé traditionnellement pour modéliser le prix des options, repose sur l'hypothèse d'une distribution log-normale pour le rendement du produit sous-jacent) alors qu'il aurait probablement été plus judicieux d'utiliser des techniques statistiques issues de la théorie des valeurs extrêmes.

Chapitre 1

Méthodologie

1.1 Introduction

Sauf à remettre en cause la validité du théorème des valeurs extrêmes (Fisher et Tippett, 1928), seule la distribution GEV (Generalized Extreme Value Distribution) introduite par Jenkinson (1955) devrait être utilisée pour modéliser une série de maxima observés sur des intervalles donnés. Par le même argument, la distribution de Pareto généralisée devrait prévaloir pour modéliser les valeurs d'une série statistique excédant un certain seuil.

La démarche classique consiste à estimer directement les paramètres de ces distributions par la méthode du maximum de vraisemblance et à procéder ensuite à l'évaluation de l'ajustement statistique en utilisant une définition particulière des fréquences empiriques des observations.

Dans la présente recherche, nous proposons une méthode d'estimation alternative de la distribution GEV, notée AGEV, qui considère le modèle statistique formé conjointement par les deux spécifications que sont la distribution GEV, exprimée ici sous forme inverse, et la distribution empirique. Par rapport à la démarche classique, cette méthode possède une série d'avantages qui seront détaillés à travers les différentes sections de ce chapitre. En voici les principaux.

1° Les biais des estimateurs des paramètres de la GEV sont considérablement réduits et, en conséquence, ceux des quantiles élevés le sont aussi¹. En outre, lorsque la taille de l'échantillon croît, les estimateurs des paramètres

1. Pour apprécier l'importance de cet effet, on peut se mettre dans la position assez fréquente en hydrologie de devoir estimer la valeur d'une crue de période de retour de 25 ans (probabilité de 4% de survenir à chaque période) à partir d'un échantillon d'une trentaine d'observations. L'approche classique sous-estimera la valeur de 10% en moyenne tandis que l'AGEV la surestimera en moyenne de 0.5%.

convergent bien plus rapidement vers leurs vraies valeurs.

2° La méthode d'estimation traditionnelle donne une somme des résidus parfois largement différente de zéro, ce qui rend délicat l'utilisation de tests statistiques de mauvaise spécification (en particulier les tests traditionnels d'autocorrélation et de choix de la forme fonctionnelle). En contraste avec cette méthode, dans le cas de l'AGEV la somme des résidus est par définition quasiment nulle. De plus, l'ajustement global du modèle, tel que mesuré par la somme des carrés des résidus, est sensiblement meilleur.

3° La méthode présente des propriétés numériques avantageuses. En contraste avec l'approche traditionnelle, d'une part, l'estimation par maximum de vraisemblance converge toujours, rapidement et en respectant les conditions qui définissent l'espace admissible des paramètres et, d'autre part, elle n'est pas sensible au choix des valeurs initiales.

4° L'AGEV se prête aisément à diverses extensions du modèle de base. En particulier, elle permet de tenir compte de la corrélation contemporaine entre les observations, ce qui se rencontre très souvent dans les données de type panel.

Notons que certains chercheurs utilisent des méthodes *ad hoc* basées sur le principe des moments pour l'estimation des paramètres de la GEV. Ces méthodes n'ont pas les propriétés théoriques du maximum de vraisemblance, mais elles s'avèrent plus précises que la méthode classique pour des petits échantillons (voir par exemple Hosking *et al.*, 1985). Cependant les méthodes des moments ne permettent pas de réduire le biais des estimateurs, et à cet égard s'avèrent moins bonnes et moins robustes que la méthode alternative. En outre, elles deviennent impraticables en présence de covariables. Dès lors, nous nous bornerons à les traiter brièvement en annexe.

Par ailleurs, dans le souci de baser l'analyse sur une spécification plus flexible que la GEV, d'autres chercheurs ont proposé récemment des distributions ayant davantage de paramètres libres, telles que la distribution de Burr à quatre paramètres (Shao *et al.*, 2004). Toutefois, le degré de non-linéarité de cette approche la rend peu praticable dans le cas d'échantillons de petite taille. Dès lors, nous ne traitons pas ces distributions dans ce chapitre et nous renvoyons au dernier chapitre consacré aux coûts de sinistres automobiles pour une application de la Burr sur un échantillon de très grande taille.

Il est assez fréquent que les praticiens, en particulier en hydrologie, discriminent des distributions statistiques, sur base d'un rapport de vraisemblance si elles sont imbriquées (par exemple Gumbel vs. GEV) ou sur base des critères de sélection AIC et BIC. Or, en petit échantillon le test du rapport de vraisemblance n'a pas une distribution standard. Ainsi, via des simulations pour différentes valeurs des paramètres et différentes tailles d'échantillon, la GEV a été testée contre la Gumbel; il apparaît que les valeurs critiques du test sont plus faibles que celle de la distribution χ^2 . On trouve par exemple que la valeur critique à 5% obtenue par bootstrap est de 3.20 au lieu de 3.84 pour la distribution χ^2 standard². Par ailleurs, quand on compare la distribution GEV avec une distribution Gamma sur base des critères AIC et BIC on est amené à rejeter la GEV dans plus de la moitié des cas alors qu'elle a servi à générer les données.

L'objectif de ce chapitre est de présenter la méthode d'estimation alternative de la GEV (AGEV) et d'établir ses propriétés pour des échantillons de taille faible et modérée en comparant ses performances à celles de la méthode d'estimation traditionnelle.

Dans la deuxième section, nous présenterons la distribution GEV et les méthodes d'estimation.

Dans la troisième section, nous comparons les propriétés des méthodes dans le cas uni dimensionnel pour différentes tailles d'échantillon.

Les quatrième et cinquième sections présentent des extensions du modèle, d'une part pour tenir compte de la présence de variables explicatives, d'autre part pour permettre l'analyse de données en panel.

2. Pour plus de détails sur les tests par bootstrap, voir Davidson et MacKinnon (2004). On y apprend notamment que pour calculer la valeur critique d'un test par bootstrap au niveau α , il faut choisir le nombre d'échantillons simulés B tel que $\alpha(B+1)$ soit un nombre entier. Ainsi, il est plus correct de calculer le test sur base de 4999 échantillons plutôt que sur 5000.

1.2 Le modèle et les techniques d'estimation

1.2.1 La distribution GEV

D'après le théorème de Fisher et Tippett (1928), toute suite de maxima d'une série statistique quelconque suit asymptotiquement une loi appartenant à l'une des familles de distribution à valeurs extrêmes, soit les distributions de Weibull, de Fréchet et de Gumbel. La distribution GEV, proposée par Jenkinson (1955), englobe ces trois types de distribution. Sa fonction de répartition a la forme suivante :

$$G(x_t) = \exp \left\{ - \left[1 + \xi \left(\frac{x_t - \mu}{\sigma} \right) \right]^{-1/\xi} \right\}, \quad (1.1)$$

où x_t représente une suite de maxima obtenus à chaque période t . Lorsque le paramètre ξ prend une valeur négative on obtient, moyennant une reparamétrisation de la fonction, une distribution de Weibull à trois paramètres. Cette distribution se distingue de la Weibull traditionnelle car elle a un paramètre supplémentaire et possède une borne supérieure au lieu d'une borne inférieure à 0. Lorsque ξ est positif il s'agit d'une distribution de Fréchet et d'une distribution de Gumbel lorsque ξ est égal à 0.

Dans le cas où le modèle prend en compte une série de k variables explicatives Z_k , le paramètre μ dans (1.1) est remplacé par $\sum_k \mu_k Z_k$.

Les conditions suivantes doivent être satisfaites :

$$\begin{aligned} -\infty < \mu < \infty, \quad \sigma > 0 \quad \text{et} \quad -\infty < \xi < \infty, \\ 1 + \xi \left(\frac{x_t - \mu}{\sigma} \right) > 0. \end{aligned} \quad (1.2)$$

La relation (1.2) détermine le support de x_t :

$$\begin{aligned} [\mu - \sigma/\xi, +\infty) & \quad \text{si} \quad \xi > 0 \\ (-\infty, +\infty) & \quad \text{si} \quad \xi = 0 \\ (-\infty, \mu - \sigma/\xi] & \quad \text{si} \quad \xi < 0 \end{aligned}$$

La moyenne de la distribution (1.1) est égale à :

$$\begin{aligned} \mu + (\sigma/\xi) * (\Gamma(1 - \xi) - 1) & \quad \text{si} \quad \xi < 1 \quad \text{et} \quad \xi \neq 0, \\ \mu + \sigma\gamma & \quad \text{si} \quad \xi = 0, \\ \text{et n'existe pas} & \quad \text{si} \quad \xi \geq 1, \end{aligned}$$

où γ est la constante d'Euler. La variance de la distribution (1.1) est donnée par :

$$\begin{aligned} & \sigma^2 (g_2 - g_1^2) / \xi^2 \quad \text{si } \xi < 1/2 \quad \text{et } \xi \neq 0, \\ & \sigma^2 \pi^2 / 6 \quad \text{si } \xi = 0, \\ & \text{et n'existe pas} \quad \text{si } \xi \geq 1/2, \end{aligned}$$

avec $g_k = \Gamma(1 - k\xi)$.

Notons que dans bon nombre d'applications, l'utilisation d'une distribution de Fréchet pour la modélisation des maxima d'une série statistique n'a pas beaucoup de sens étant donné qu'elle ne possède pas de limite maximum finie. Cela peut conduire à des résultats absurdes dans la prédiction d'événements de probabilité (très) faible. Les distributions de type Weibull, qui possèdent un plafond, sont, pour les cas qui nous intéressent, beaucoup plus réalistes. Il en est de même pour la distribution de Gumbel qui peut être considérée comme le cas limite. Dans la littérature appliquée, la plupart des chercheurs n'attachent pas beaucoup d'importance à ce problème, ce qui marque une préférence à la vraisemblance statistique par rapport à la vraisemblance physique ou économique. Dans ce chapitre nous ne considérons que les distributions de type Weibull et de Gumbel qui en est la limite. Ces distributions sont les plus réalistes dans les domaines de l'économie et de l'hydrologie qui sont étudiés dans les chapitres empiriques. La transposition au cas Fréchet des techniques d'estimation présentée dans ce chapitre ne pose aucun problème.

1.2.2 Estimation par la méthode traditionnelle

Méthodologie

Traditionnellement, l'estimation des paramètres de la distribution (1.1) s'effectue par la méthode de maximisation de la vraisemblance.

La fonction de densité pour une observation est donnée par :

$$g(x_t) = \frac{1}{\sigma} \exp \left\{ - \left[1 + \xi \left(\frac{x_t - \mu}{\sigma} \right) \right]^{(-1/\xi)} \right\} \left[1 + \xi \left(\frac{x_t - \mu}{\sigma} \right) \right]^{-(1+1/\xi)}.$$

La méthode classique consiste à maximiser le logarithme de la vraisemblance du modèle (1.1), défini par :

$$\begin{aligned} L(x \mid \mu, \sigma, \xi) = & -n \log \sigma - (1 + 1/\xi) \sum_t \log \left[1 + \xi \left(\frac{x_t - \mu}{\sigma} \right) \right] \\ & - \sum_t \left[1 + \xi \left(\frac{x_t - \mu}{\sigma} \right) \right]^{-1/\xi}, \end{aligned} \quad (1.3)$$

par rapport au vecteur des paramètres (μ, σ, ξ) .

Comme il n'existe pas de solution analytique à ce problème, il doit être résolu par une procédure numérique. Même si les estimateurs sont biaisés en petits échantillons, cette méthode possède les propriétés asymptotiques de consistance et d'efficacité sous les conditions de régularité usuelles.

Ajustement global du modèle

L'évaluation de l'ajustement statistique du modèle (1.1) est basé sur une analyse des résidus donnés par la différence entre les observations $x_{(i)}$ classées par ordre croissant et les quantiles estimés correspondant définis par :

$$\hat{x}_{(i)} = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}} \left[1 - (-\log(F_{(i)}))^{-\hat{\xi}} \right], \quad (1.4)$$

où $F_{(i)}$ est une définition particulière de la fréquence empirique des observations. La fréquence empirique des observations est généralement définie par :

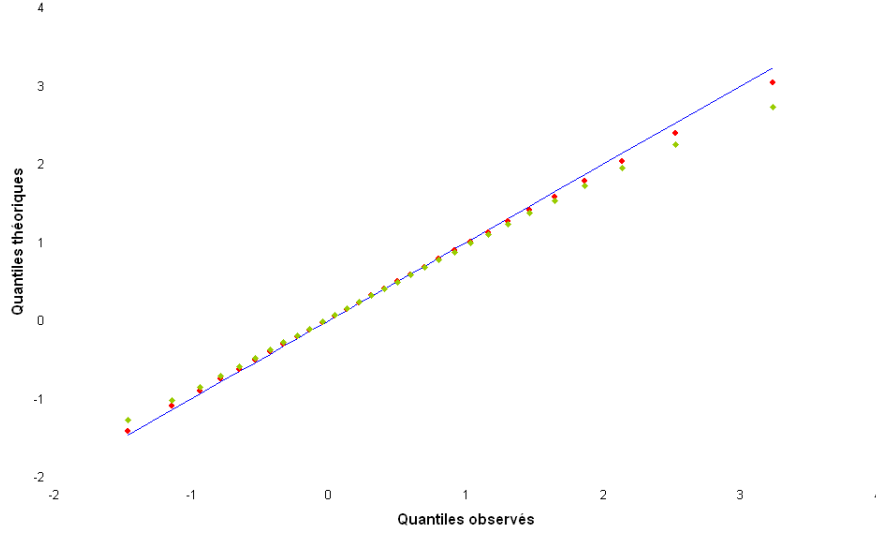
$$F_{(i)}(\alpha) = \frac{i - \alpha}{n + 1 - 2\alpha}, \quad (1.5)$$

pour $\alpha < 1$.

Dans la littérature empirique utilisant les distributions à valeurs extrêmes, la valeur de 0 pour le paramètre α a été celle utilisée initialement (Gumbel, 1958 ; Weibull, 1939) et reste la plus commune en pratique, bien que Cunnane (1978) ait montré qu'une valeur de 0.4 donnait le meilleur ajustement. En exposant sa méthode des moments pondérés, Hosking *et al.* (1985) privilégient la valeur de 0.35. Nous reviendrons sur le choix de ce paramètre ultérieurement (cf. section 1.3.2).

Lorsque les deux séries, $x_{(i)}$ et $\hat{x}_{(i)}$, sont mises en relation et présentées graphiquement, on obtient le traditionnel Q-Q plot à partir duquel la qualité de l'ajustement s'évalue en comparant l'alignement des points $(x_{(i)}, \hat{x}_{(i)})$ dans le plan avec une droite à 45 degrés.

Afin de visualiser l'importance du paramètre α , nous avons créé un échantillon de 30 observations issues d'une distribution GEV à partir d'une définition des fréquences empiriques (1.5) dans laquelle la valeur du α a été fixée à 0.4. Nous avons ensuite estimé les paramètres de cette GEV par la méthode classique et avons calculé les résidus du modèle en utilisant dans la fonction (1.5) alternativement les valeurs de 0 et de 0.4 pour le α . Les Q-Q plot qui en résultent sont présentés à la figure 1.1, avec en rouge les résultats pour $\alpha = 0.4$ et en vert ceux pour $\alpha = 0$. La figure 1.1 tend à indiquer que la qualité d'ajustement du modèle dépend sensiblement du choix du paramètre α , en particulier dans les extrémités de la distribution.

FIGURE 1.1 – Influence du α sur la qualité d'ajustement du modèle

Inférence sur les paramètres

L'inférence traditionnelle repose sur une approximation quadratique de la matrice de variance-covariance des estimateurs. En effet, sous certaines conditions de régularité (Coles, 2001 ; Smith, 1985)³, le vecteur des estimateurs des paramètres de la distribution $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ est distribué asymptotiquement comme une normale multivariée dont la moyenne est la valeur des paramètres dans la population (μ, σ, ξ) et dont la matrice de variance-covariance est donnée par l'inverse de la matrice d'information évaluée en pratique à la valeur estimée des paramètres :

$$I(\mu, \sigma, \xi)^{-1} = E \begin{bmatrix} \frac{-\partial^2(\mu, \sigma, \xi)}{\partial \mu^2} & \frac{-\partial^2(\mu, \sigma, \xi)}{\partial \mu \partial \sigma} & \frac{-\partial^2(\mu, \sigma, \xi)}{\partial \mu \partial \xi} \\ \frac{-\partial^2(\mu, \sigma, \xi)}{\partial \sigma \partial \mu} & \frac{-\partial^2(\mu, \sigma, \xi)}{\partial \sigma^2} & \frac{-\partial^2(\mu, \sigma, \xi)}{\partial \sigma \partial \xi} \\ \frac{-\partial^2(\mu, \sigma, \xi)}{\partial \xi \partial \mu} & \frac{-\partial^2(\mu, \sigma, \xi)}{\partial \xi \partial \sigma} & \frac{-\partial^2(\mu, \sigma, \xi)}{\partial \xi^2} \end{bmatrix}^{-1}. \quad (1.6)$$

3. Smith a démontré que les conditions habituelles de régularité ne sont plus respectées lorsque $\xi < -0.5$.

Si on dénomme les éléments diagonaux de l'inverse de la matrice d'information par $\psi_{i,i}$, la distribution asymptotique des estimateurs du maximum de vraisemblance est donnée par :

$$\begin{aligned}\hat{\mu} &\sim N(\mu, \psi_{1,1}) \\ \hat{\sigma} &\sim N(\sigma, \psi_{2,2}) \\ \hat{\xi} &\sim N(\xi, \psi_{3,3})\end{aligned}$$

Périodes de retour

En substituant la valeur estimée des paramètres dans (1.4), on obtient les estimateurs des quantiles de la distribution donnés par :

$$\hat{Q}_{(T)} = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}} \left[1 - \left(-\log \left(1 - \frac{1}{T} \right) \right)^{-\hat{\xi}} \right], \quad (1.7)$$

où T est la période de retour pour laquelle on désire une estimation du quantile⁴. En utilisant la méthode du delta, la théorie asymptotique présentée ci-dessus peut être transposée pour faire de l'inférence sur les estimateurs des quantiles. En effet, en grand échantillon, on peut montrer que :

$$\hat{Q}_{(T)} \sim N(Q_{(T)}, s_{Q_{(T)}}), \quad (1.8)$$

où :

$$s_{Q_{(T)}} \approx \Delta Q'_{(T)} I(\mu, \sigma, \xi)^{-1} \Delta Q_{(T)}, \quad (1.9)$$

avec $I(\mu, \sigma, \xi)^{-1}$ donnée par (1.6) et $\Delta Q'_{(T)} = \left[\frac{\partial Q_{(T)}}{\partial \mu}, \frac{\partial Q_{(T)}}{\partial \sigma}, \frac{\partial Q_{(T)}}{\partial \xi} \right]$.

Limites de l'approche classique

La méthode classique d'estimation par maximum de vraisemblance rencontre plusieurs inconvénients. D'une part, celle-ci est connue pour présenter un biais en petit échantillon (et dans la pratique, il semble que tous les échantillons

4. Le concept de période de retour prête souvent à confusion et il n'est pas rare de le voir interprété de manière erronée y compris dans des travaux d'experts. Soit une variable x_t dont les réalisations en deux points quelconques du temps sont des événements indépendants (ce qui exclut toute régularité temporelle!) et soit p la probabilité que sa valeur en tout t soit égale ou supérieure à un seuil x_0 . La période de retour, T , est définie comme l'inverse de cette probabilité : $T = 1/p$. Une période de retour de 25 ans par exemple signifie que l'on a chaque année une probabilité de 4% d'observer $x \geq x_0$. Cela ne veut pas dire qu'on observera en moyenne pareil événement au moins une fois en 25 ans et encore moins une seule fois. En effet, à une période de retour de 25 ans, correspond sur cet intervalle une probabilité de 64% d'occurrence de $x \geq x_0$ au moins une fois (37,5% que ce soit une seule fois). Ce qui est vrai par contre, mais peu utile en pratique, est que, sur une très longue période, 25 ans sera l'écart le plus fréquent entre deux survenances de $x \geq x_0$ et c'est sans doute cela qui a donné naissance au terme.

soient petits). Dans la section 1.3, nous illustrerons ceci à l'aide de simulations de Monte Carlo. D'autre part, l'argument du logarithme apparaissant dans la fonction (1.3) doit être strictement positif, ce qui n'est pas toujours vérifié numériquement, auquel cas l'estimation n'est pas convergente. Par ailleurs, la méthode traditionnelle n'offre pas un traitement satisfaisant des données observées simultanément en différents points de l'espace.

1.2.3 Méthode d'estimation alternative (AGEV)

Méthodologie

Comme explicité dans la section précédente, l'approche standard dans la littérature empirique consiste à maximiser la vraisemblance de la fonction GEV (1.1) par rapport aux paramètres conditionnellement aux observations. Ensuite, l'évaluation de la qualité de l'ajustement est réalisée en utilisant une définition particulière de la fréquence empirique des observations donnée souvent par une version de la fonction (1.5) où le paramètre α est choisi de manière *ad hoc*. Aucune hypothèse sur les aléas n'est formalisée à priori. Nous proposons ici une méthode d'estimation alternative des paramètres de la distribution GEV. Si on prend le logarithme des deux côtés de l'expression (1.1) et qu'on solutionne pour x , on obtient :

$$x = \mu - \frac{\sigma}{\xi} \left[1 - (-\log(G(x)))^{-\xi} \right]. \quad (1.10)$$

Ensuite, on classe les observations, on remplace $G(x)$ par les fréquences empiriques (1.5) et on ajoute un aléa d'espérance nulle et de variance constante. On obtient alors le modèle de régression suivant :

$$x_{(i)} = \mu - \frac{\sigma}{\xi} \left[1 - (-\log(F_{(i)}(\alpha)))^{-\xi} \right] + \epsilon_{(i)}, \quad (1.11)$$

qui peut être estimé par la méthode du quasi maximum de vraisemblance.

La méthode d'estimation alternative consiste en fait à trouver la valeur des paramètres qui minimise la distance entre les quantiles théoriques et observés.

Nous verrons dans les chapitres suivants consacrés aux prix hédoniques et à l'étude de crues que les différents tests de mauvaise spécification sont plus rarement significatifs lorsque la méthode alternative est utilisée.

Remarquons que, en estimant le modèle (1.11), on peut soit estimer simultanément le α (par *grid search*), soit estimer les coefficients de la distribution conditionnellement à une valeur du α fixée arbitrairement. On verra que les coefficients d'intérêt changent très peu, mais que la variance est plus faible avec la première approche.

La spécification (1.11) s'accommode très bien d'un trend linéaire ainsi que de différents types d'effets dynamiques. Le test traditionnel d'autocorrélation est un bon signal pour ce genre de processus. Une telle hypothèse n'a de sens que dans le deuxième chapitre consacré aux crues et aucun effet de ce type n'y a été détecté.

Avantages de la méthode AGEV

1. Tout d'abord, notons la différence conceptuelle substantielle qui réside dans l'estimation du modèle AGEV (1.11) par rapport à l'estimation traditionnelle. La méthode traditionnelle cherche la valeur des paramètres qui maximise la vraisemblance du modèle conditionnellement aux observations et fait ensuite de l'inférence sur la qualité de l'ajustement en fonction d'une spécification de la fréquence empirique. L'évaluation de l'ajustement du modèle est donc tributaire du choix, opéré de manière *ad hoc*, de la valeur du paramètre α dans la fonction (1.5). En contraste avec la procédure classique, la méthode proposée cherche, quant à elle, la valeur des paramètres qui minimise la somme des carrés résiduelle du modèle statistique (1.11) conditionnellement aux observations et à une spécification de la fréquence empirique (1.5).
2. La contrainte (1.2) est automatiquement satisfaite lors de l'estimation des paramètres ce qui n'est pas le cas lorsque la méthode classique est utilisée. L'estimation est toujours convergente, ce qui n'est pas le cas de la méthode classique, en particulier lorsque ξ est proche ou inférieur à -0.5 et que l'échantillon est de taille réduite.
3. La méthode alternative impliquant une estimation par moindres carrés, toute l'inférence habituelle associée à ce processus d'estimation peut être utilisée. Ainsi, nous montrons dans la section 1.3 que la somme des résidus est pratiquement nulle lorsque cette méthode est utilisée même pour des échantillons de très petite taille, contrastant fortement avec les résultats découlant des estimations réalisées par la méthode d'estimation classique.
4. En réécrivant le modèle, il est aisé d'utiliser la méthode alternative dans des applications impliquant des données de panel avec une matrice de variance-covariance contemporaine des résidus non diagonale, ce qui n'est pas permis par la méthode d'estimation classique.

Inférence sur les paramètres

La variable explicative $F_{(i)}(\alpha)$, construite sur base du classement des observations est une variable non observée et, par conséquent, l'estimation du modèle (1.11) est sujette à la critique concernant les problèmes liés aux régresseurs générés (cf. Pagan, 1984)⁵. Ainsi, bien que les estimateurs des paramètres du

5. Nous remercions Russell Davidson pour ses remarques et conseils très judicieux à ce sujet.

modèle (1.11) conservent leurs propriétés désirables, les estimateurs des écart-types des paramètres sont biaisés et inconsistants, invalidant l'inférence traditionnelle. Cependant, une estimation des écart-types des paramètres du modèle (1.11) peut être obtenue par bootstrapping. En effet, les estimateurs des paramètres étant consistants, on peut considérer que le modèle suivant est une bonne approximation du processus ayant généré les données (*Data Generating Process*) :

$$G(x_t) = \exp \left\{ - \left[1 + \hat{\xi} \left(\frac{x_t - \hat{\mu}}{\hat{\sigma}} \right) \right]^{-1/\hat{\xi}} \right\}.$$

A partir de cette expression, on effectue 5000 tirages aléatoires d'échantillons de même taille que l'échantillon original et on estime (1.11) pour chacun de ces tirages. Les écart-types des paramètres calculés sur l'ensemble des échantillons simulés constituent une bonne approximation des écarts-type des paramètres de l'estimation initiale. L'approche suivie est décrite en détail dans Davidson et MacKinnon (2004).

1.2.4 Interprétation des méthodes

Pour comprendre les propriétés des deux méthodes, on peut examiner le comportement de la fonction de vraisemblance de chacune d'elles autour du maximum. Nous avons réalisé l'exercice pour un échantillon de 50 observations issues d'une loi GEV de paramètres $\mu = 0$, $\sigma = 1$ et pour différentes valeurs du paramètre ξ . Les figures 1.2 et 1.3 sont obtenues sur base des résultats pour une valeur du ξ de -0.1, les conclusions étant essentiellement identiques pour une valeur du ξ plus éloignée de 0. La figure 1.2 représente la valeur maximale de la fonction de vraisemblance lorsque l'on fait varier la valeur du paramètre ξ autour de sa valeur optimale en maintenant la valeur des autres paramètres constante, tandis que la figure 1.3 représente la dérivée partielle de la fonction de vraisemblance⁶. On remarque que le maximum de la fonction de vraisemblance dans la méthode classique est atteint lorsque le paramètre ξ prend une valeur sensiblement différente de sa vraie valeur (-0.14 au lieu de -0.1). On voit aussi que la fonction de vraisemblance est très plate dans les alentours du maximum. En outre, il est à noter que la valeur optimale du paramètre ξ obtenue par la méthode classique ne converge que très lentement vers la valeur vraie lorsque la taille de l'échantillon augmente. La fonction de vraisemblance qui découle de la méthode d'estimation alternative atteint son maximum exactement à la valeur vraie du paramètre de courbure, tandis que la fonction autour du maximum est beaucoup moins plate et beaucoup plus symétrique que celle de la méthode classique.

6. Dans les figures, les résultats de la méthode alternative sont en rouge et ceux de la méthode classique en bleu.

FIGURE 1.2 – Forme des fonctions de vraisemblance

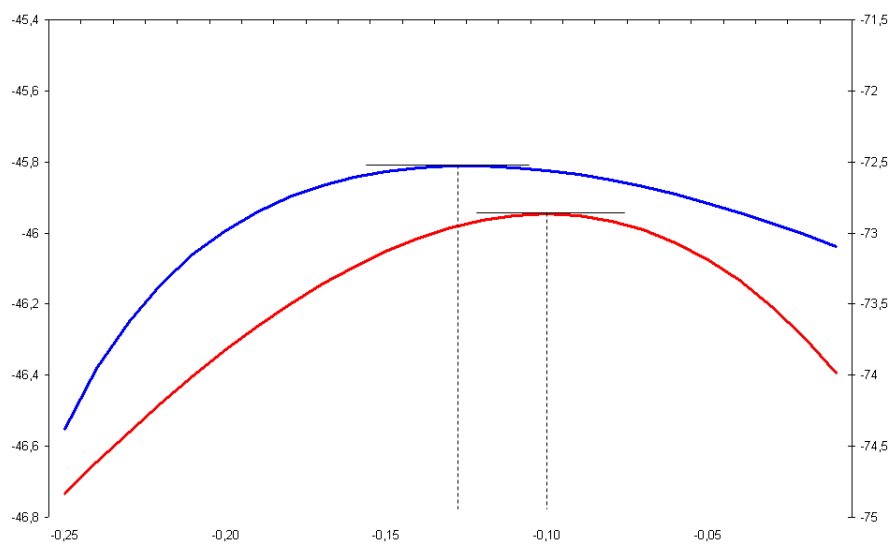
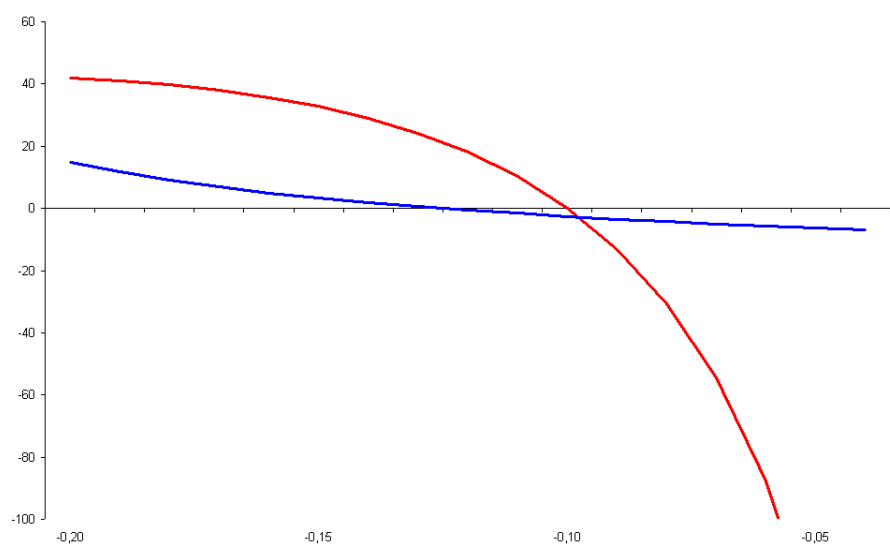


FIGURE 1.3 – Dérivées des fonctions de vraisemblance



1.2.5 Autres méthodes d'estimation

D'autres méthodes d'estimation apparaissent dans la littérature, comme par exemple la méthode des moments pondérés (Hosking *et al.*, 1985). Il faut souligner que cette méthode n'est pas plus robuste face à l'ajout de nouvelles observations, ne réduit pas le biais mais offre un gain d'efficacité en petit échantillon (Hosking *et al.*, 1985). Ensuite et surtout, elle ne permet pas aisément l'inclusion de variables explicatives dans l'analyse ce qui la rend moins utile dans les applications économiques⁷. C'est pourquoi nous ne la détaillerons pas dans cette recherche.

1.3 Comparaison des méthodes par simulations de Monte Carlo

1.3.1 Échantillons

Dans cette section, nous comparons les performances de la méthode alternative avec celles de la méthode d'estimation classique pour des échantillons de petite et de moyenne taille à l'aide de simulations de Monte Carlo. En l'occurrence, 5000 échantillons de taille $n = 30, 50, 100$ et 200 ont été tirés aléatoirement d'une distribution GEV pour différentes valeurs du paramètre de courbure comprises entre -0.1 et -0.5 (distribution Weibull)⁸.

Les méthodes d'estimations utilisées dans ce chapitre étant invariables à une transformation linéaire près, nous pouvons normaliser la valeur des autres paramètres, ($\mu = 0$ et $\sigma = 1$) sans perdre en généralité (Hosking *et al.*, 1985).

Pour chaque échantillon et pour chacune des deux méthodes, les paramètres sont estimés, la somme des résidus et l'écart-type de la régression sont calculés ainsi que les quantiles de la distribution correspondant aux périodes de retour de 25, 50, 100 et 200 ans. Les valeurs des paramètres ainsi que des quantiles estimés par les deux méthodes sont comparées avec les vraies valeurs. Les résultats sont présentés à la table 1.1 pour les paramètres et les résidus et à la table 1.2 pour les quantiles.

7. Notre travail de recherche comportant l'estimation de distributions GEV sur des données hydrologiques, nous présentons brièvement la méthode des moments pondérés en annexe étant donné que celle-ci est largement répandue dans la littérature appliquée en hydrologie.

8. Pour éviter des problèmes numériques survenant lorsque la valeur tirée de la loi uniforme est trop proche d'une des bornes de l'intervalle $[0, 1]$ (ce qui donnerait d'ailleurs une valeur de x complètement irréaliste), nous avons très légèrement rétréci cet intervalle, d'une part en prenant 10^{-3} comme borne inférieure et, d'autre part, de manière à exclure des valeurs de période de retour excédant le quadruple de la taille de l'échantillon. Cela explique peut-être que le nombre de cas de non convergence pour la méthode classique soit réduit de moitié par rapport à celui que donne Hosking *et al.* (1985).

Avec la méthode d'estimation classique, dans un certain nombre de cas (pouvant atteindre presque 2% pour $n = 30$), le processus de maximisation de la fonction de vraisemblance ne converge pas ou converge vers des valeurs des paramètres qui violent la condition (1.2)⁹. L'expérience met aussi en évidence la sensibilité de la procédure au choix des valeurs initiales des paramètres. Avec la méthode d'estimation alternative ces deux problèmes sont inexistantes. Par conséquent, les résultats de la méthode alternative fournissent des valeurs initiales fiables pour l'estimation classique.

1.3.2 Choix du paramètre α

L'analyse a été réalisée avec différentes valeurs pour le paramètre α apparaissant dans la spécification de la fréquence empirique (1.5).

Dans le cas de la méthode classique, le choix du paramètre α n'intervient pas au moment de l'estimation du modèle. Sa valeur influence uniquement l'évaluation de l'ajustement global. Néanmoins, après chaque estimation des paramètres de la GEV, nous avons procédé à l'estimation du paramètre α qui minimisait la somme des carrés des résidus de l'échantillon. On obtient une valeur moyenne assez proche du 0.4 de Cunnane (1978).

Dans le cas de l'estimation du modèle par la méthode AGEV, la valeur du paramètre α a été obtenue par une procédure de grid search conjointement avec les valeurs des autres paramètres. Plus précisément, la valeur estimée de α est celle qui minimise la somme des carrés des résidus du modèle (1.11) sous la contrainte $\alpha < 1$. Trois résultats émergent. D'abord, la valeur moyenne des α est quasi nulle avec une distribution extrêmement plate. Ensuite, la variation du maximum estimé de la fonction de vraisemblance lorsque α est modifié est d'autant plus faible (en valeur absolue) que la taille de l'échantillon est élevée ou que la valeur du paramètre ξ se rapproche de -0.25. Enfin, et surtout, les valeurs des paramètres de la GEV sont peu affectées, mais l'estimation du α conduit à un gain en efficience. En d'autres termes, les différences entre les deux méthodes proviennent avant tout de la maximisation de la fonction de vraisemblance statistique de (1.11) plutôt que de (1.3).

Les résultats présentés dans les tables 1.1 et 1.2 ont été obtenus sur base de la valeur optimale de ce paramètre. En agissant ainsi, au lieu de prendre la valeur $\alpha = 0$ on réduit sensiblement (parfois de plus de 50% dans certains échantillons) la somme et la variance estimée des résidus. Néanmoins, comme il est nettement plus facile d'estimer le modèle sous la contrainte $\alpha = 0$, qu'en pratique c'est la valeur la plus communément utilisée et que les effets de cette restriction sur les biais des paramètres et des quantiles sont faibles, nous présentons également les résultats obtenus avec cette contrainte en annexe.

9. Dans ces cas, l'échantillon tiré a été éliminé pour les deux méthodes.

TABLE 1.1 – Paramètres et résidus

$\xi =$	-0.1		-0.2		-0.3		-0.4		-0.5	
	Cl.	Al.	Cl.	Al.	Cl.	Al.	Cl.	Al.	Cl.	Al.
<i>N=30</i>										
μ	.029	-.010	.032	-.012	.035	-.014	.037	-.016	.039	.019
	<i>.211</i>	<i>.205</i>	<i>.208</i>	<i>.204</i>	<i>.206</i>	<i>.203</i>	<i>.203</i>	<i>.204</i>	<i>.201</i>	<i>.205</i>
σ	.950	1.034	.954	1.034	.959	1.035	.964	1.035	.970	1.036
	<i>.144</i>	<i>.177</i>	<i>.142</i>	<i>.169</i>	<i>.144</i>	<i>.167</i>	<i>.147</i>	<i>.169</i>	<i>.152</i>	<i>.174</i>
ξ	-.156	-.123	-.261	-.228	-.365	-.332	-.469	-.436	-.574	-.541
	<i>.160</i>	<i>.151</i>	<i>.157</i>	<i>.148</i>	<i>.155</i>	<i>.151</i>	<i>.157</i>	<i>.160</i>	<i>.159</i>	<i>.174</i>
$\sum \epsilon$	.315	.000	.089	.000	-.112	.000	-.332	.000	-.576	.000
<i>s.e.</i>	.163	.143	.150	.132	.142	.125	.139	.121	.139	.119
$ \sum \epsilon $	.403	.000	.291	.000	.286	.000	.400	.000	.605	.000
<i>N=50</i>										
μ	.020	-.005	.022	-.007	.024	-.007	.026	-.008	.028	-.008
	<i>.160</i>	<i>.158</i>	<i>.158</i>	<i>.157</i>	<i>.157</i>	<i>.157</i>	<i>.155</i>	<i>.157</i>	<i>.154</i>	<i>.158</i>
σ	.966	1.021	.969	1.020	.972	1.021	.975	1.022	.979	1.022
	<i>.109</i>	<i>.132</i>	<i>.107</i>	<i>.125</i>	<i>.108</i>	<i>.123</i>	<i>.110</i>	<i>.124</i>	<i>.114</i>	<i>.127</i>
ξ	-.136	-.113	-.238	-.217	-.341	-.320	-.443	-.424	-.547	-.528
	<i>.107</i>	<i>.105</i>	<i>.104</i>	<i>.103</i>	<i>.102</i>	<i>.105</i>	<i>.103</i>	<i>.112</i>	<i>.106</i>	<i>.124</i>
$\sum \epsilon$	.446	.000	.178	.000	-.062	.000	-.318	.000	-.599	.000
<i>s.e.</i>	.136	.122	.123	.111	.117	.105	.114	.101	.114	.100
$ \sum \epsilon $	.504	.000	.308	.000	.251	.000	.363	.000	.609	.000
<i>N=100</i>										
μ	.013	-.002	.014	-.002	.016	-.002	.017	-.003	.018	-.003
	<i>.112</i>	<i>.112</i>	<i>.111</i>	<i>.111</i>	<i>.110</i>	<i>.111</i>	<i>.109</i>	<i>.111</i>	<i>.108</i>	<i>.112</i>
σ	.979	1.010	.981	1.010	.982	1.010	.984	1.011	.986	1.011
	<i>.076</i>	<i>.090</i>	<i>.075</i>	<i>.085</i>	<i>.075</i>	<i>.084</i>	<i>.077</i>	<i>.084</i>	<i>.079</i>	<i>.087</i>
ξ	-.120	-.105	-.221	-.208	-.323	-.310	-.424	-.412	-.525	-.515
	<i>.068</i>	<i>.070</i>	<i>.065</i>	<i>.067</i>	<i>.063</i>	<i>.068</i>	<i>.063</i>	<i>.073</i>	<i>.064</i>	<i>.080</i>
$\sum \epsilon$	.658	.000	.302	.000	-.010	.000	-.347	.000	-.706	.000
<i>s.e.</i>	.104	.095	.092	.085	.086	.079	.084	.076	.085	.076
$ \sum \epsilon $	.737	.000	.416	.000	.245	.000	.374	.000	.708	.000
<i>N=200</i>										
μ	.009	.000	.010	.000	.011	.000	.012	.000	.012	.000
	<i>.079</i>	<i>.079</i>	<i>.078</i>	<i>.079</i>	<i>.078</i>	<i>.079</i>	<i>.077</i>	<i>.079</i>	<i>.076</i>	<i>.079</i>
σ	.986	1.003	.986	1.004	.987	1.004	.988	1.004	.989	1.003
	<i>.053</i>	<i>.061</i>	<i>.052</i>	<i>.058</i>	<i>.052</i>	<i>.057</i>	<i>.053</i>	<i>.058</i>	<i>.055</i>	<i>.060</i>
ξ	-.109	-.100	-.210	-.202	-.312	-.303	-.412	-.405	-.513	-.506
	<i>.045</i>	<i>.048</i>	<i>.043</i>	<i>.045</i>	<i>.041</i>	<i>.046</i>	<i>.040</i>	<i>.049</i>	<i>.041</i>	<i>.055</i>
$\sum \epsilon$	.949	.000	.461	.000	.040	.000	-.425	.000	-.899	.000
<i>s.e.</i>	.078	.072	.068	.063	.063	.058	.062	.056	.062	.056
$ \sum \epsilon $	1.099	.000	.633	.000	.296	.000	.443	.000	.906	.000

TABLE 1.2 – Biais des quantiles extrêmes

$\xi =$	-0.1		-0.2		-0.3		-0.4		-0.5	
	Cl.	Al.	Cl.	Al.	Cl.	Al.	Cl.	Al.	Cl.	Al.
<i>N=30</i>										
Q_{10}	-.175	.020	-.150	.007	-.127	-.002	-.105	-.008	-.086	-.013
Q_{25}	-.275	.038	-.232	.012	-.193	-.002	-.158	-.010	-.127	-.014
Q_{50}	-.341	.064	-.284	.023	-.233	.003	-.187	-.007	-.148	-.011
Q_{100}	-.394	.106	-.326	.042	-.264	.013	-.209	-.001	-.163	-.007
Q_{200}	-.431	.166	-.357	.069	-.286	.027	-.224	.008	-.173	.000
<i>N=50</i>										
Q_{10}	-.118	.013	-.101	.003	-.085	-.003	-.070	-.007	-.056	-.010
Q_{25}	-.190	.021	-.160	.002	-.132	-.007	-.107	-.012	-.085	-.014
Q_{50}	-.243	.033	-.201	.005	-.163	-.007	-.130	-.013	-.102	-.016
Q_{100}	-.290	.051	-.236	.011	-.188	-.006	-.147	-.013	-.114	-.016
Q_{200}	-.331	.078	-.267	.020	-.209	-.002	-.161	-.011	-.122	-.014
<i>N=100</i>										
Q_{10}	-.069	.011	-.059	.004	-.049	.000	-.041	-.003	-.032	-.005
Q_{25}	-.113	.017	-.096	.004	-.079	-.003	-.064	-.007	-.050	-.009
Q_{50}	-.147	.025	-.122	.005	-.098	-.005	-.078	-.009	-.060	-.010
Q_{100}	-.179	.036	-.146	.008	-.116	-.005	-.090	-.010	-.068	-.011
Q_{200}	-.209	.050	-.167	.012	-.130	-.004	-.099	-.010	-.074	-.011
<i>N=200</i>										
Q_{10}	-.039	.008	-.034	.004	-.029	.001	-.023	-.001	-.018	-.002
Q_{25}	-.064	.014	-.055	.006	-.046	.000	-.037	-.002	-.029	-.003
Q_{50}	-.082	.020	-.069	.008	-.057	.000	-.045	-.003	-.035	-.004
Q_{100}	-.100	.029	-.083	.010	-.067	.001	-.052	-.003	-.040	-.004
Q_{200}	-.116	.039	-.095	.014	-.076	.002	-.058	-.003	-.043	-.004
<i>Q vrais</i>										
Q_{10}	2.015		1.812		1.636		1.484		1.351	
Q_{25}	2.737		2.363		2.056		1.804		1.596	
Q_{50}	3.231		2.709		2.299		1.975		1.716	
Q_{100}	3.687		3.007		2.495		2.103		1.799	
Q_{200}	4.111		3.266		2.653		2.199		1.858	

1.3.3 Ajustement global

Notons tout d'abord que si la somme des résidus ($\sum \epsilon$) est quasi nulle dans l'estimation par la méthode AGEV, ce n'est pas du tout le cas pour les estimations réalisées par la méthode classique, ce qui rend délicat l'utilisation de certains tests statistiques tels que les tests traditionnels d'autocorrélation des résidus. On remarque qu'avec la méthode classique, la somme des résidus diminue de manière continue avec la valeur du paramètre ξ , franchissant le seuil de 0 pour une valeur de ce paramètre comprise entre (-0.2) et (-0.3)¹⁰. C'est dans cette zone que la distribution GEV prend une forme presque symétrique.

Mais surtout, l'ajustement global par la méthode AGEV est meilleur, comme l'indique un écart-type des résidus (*s.e.*) systématiquement plus faible.

1.3.4 Biais et variances des estimateurs

En ce qui concerne les estimations réalisées par la méthode classique, nous retrouvons les résultats de l'étude de Hosking *et al.* (1985). Ainsi, la méthode classique donne un estimateur des paramètres σ et ξ présentant un biais systématique vers le bas et un estimateur du paramètre μ biaisé vers le haut, et ce, quelle que soient la valeur de ξ et la taille de l'échantillon.

Les biais des estimateurs des paramètres μ et σ par la méthode alternative sont de signe opposé aux biais de la méthode classique, mais surtout, ils leur sont presque toujours inférieurs. Les vraies valeurs de ces paramètres semblent être comprises dans l'intervalle constitué par les résultats des estimations réalisées par chacune des deux méthodes.

Le biais de l'estimateur du paramètre ξ par la méthode alternative est de même signe que celui de la méthode classique tout en lui étant systématiquement inférieur en valeur absolue.

Par rapport aux résultats des estimations réalisées par la méthode classique, les écarts-types des paramètres obtenus par la méthode alternative sont, soit identiques, soit légèrement supérieurs¹¹. Sans surprise, quelle que soit la méthode, lorsque la taille de l'échantillon augmente, le biais et la variance des estimateurs des différents paramètres se réduisent. Ainsi, il est à noter que pour des échantillons créés à partir d'une distribution possédant un ξ de -0.2

10. La table 1.1 donne également la moyenne de la somme des résidus en valeur absolue qui découle des estimations ($|\sum \epsilon|$).

11. Les écarts-types présentés dans les tables sont obtenus à partir de la variance des paramètres et des quantiles estimés à travers les 5000 tirages. Surtout dans la méthode alternative, ils sont quasi toujours plus élevés que les écarts-types moyens qui ressortent des estimations individuelles.

ce paramètre ξ n'est pas estimé une seule fois supérieur à zéro lorsque la taille des échantillons simulés dépasse 100 observations, alors que c'est bien souvent le cas pour des échantillons de taille plus restreinte.

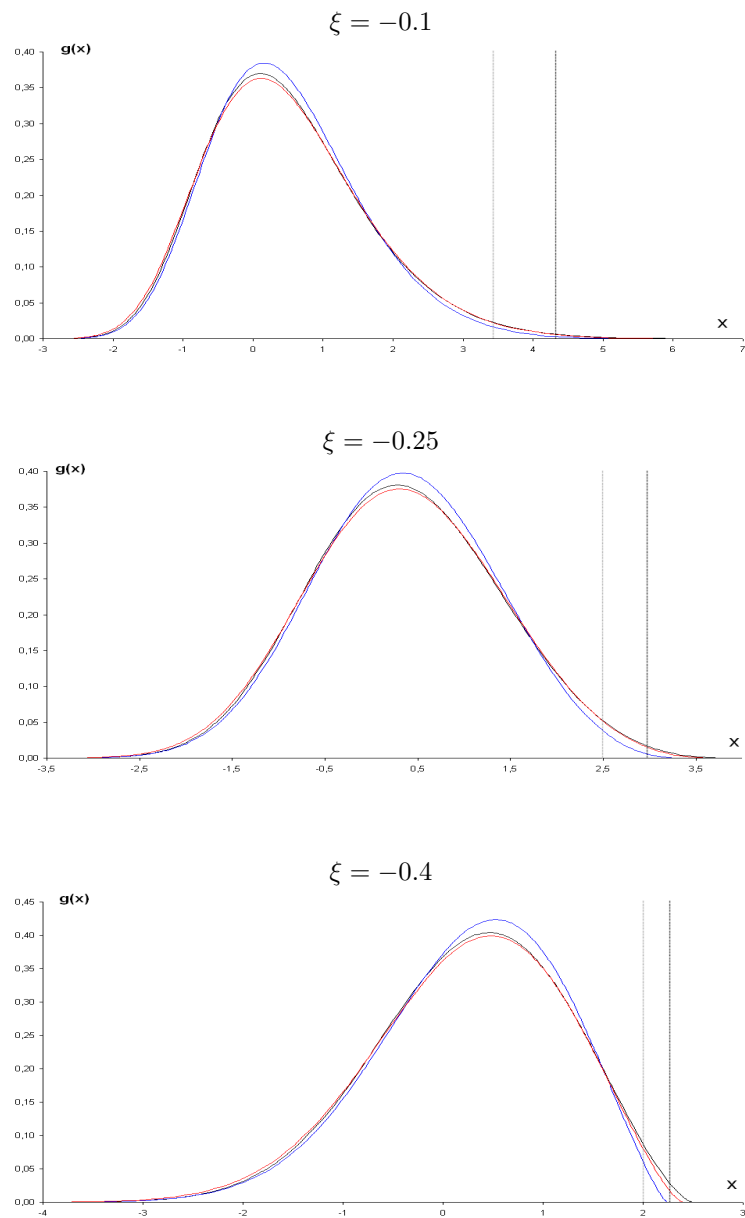
Afin de donner une image synthétique de l'effet du biais des différents paramètres sur la forme de la distribution, il est plus facile de procéder à une analyse graphique. Ainsi, la figure 1.4 représente les fonctions de densité de la GEV telles que données par les valeurs moyennes des paramètres estimés par chacune des deux méthodes et la fonction de densité d'une GEV possédant les vraies valeurs des paramètres. Les résultats pour la méthode classique sont en bleu, pour la méthode alternative en rouge et la GEV vraie est tracée en noir. Les lignes verticales indiquent respectivement les quantiles correspondant à une période de retour de 25 et 100 ans pour la vraie distribution. Les fonctions de fréquences présentées ont été construites à partir des résultats des estimations réalisées sur base d'échantillons de 50 observations et nous avons présentés les résultats pour les valeurs du paramètre ξ de -0.1, -0.25 et -0.4. La méthode AGEV domine nettement dans les parties modales et extrêmes.

1.3.5 Biais des quantiles extrêmes

Les estimateurs des quantiles extrêmes de la distribution obtenus par la méthode classique présentent un biais systématique vers le bas.

Les estimateurs de ces quantiles obtenus par la méthode alternative présentent au contraire un biais vers le haut pour $\xi > -0.25$, mais surtout, les biais sont toujours plus petits en valeur absolue que ceux des estimateurs classiques. La supériorité de la méthode alternative est bien souvent spectaculaire, en particulier dans des échantillons de petite taille.

FIGURE 1.4 – Fonctions de densité



1.3.6 Robustesse

Afin d'apprécier la robustesse de la méthode alternative en comparaison avec la méthode classique, nous avons procédé à un exercice dans lequel chaque échantillon simulé a été perturbé par l'ajout d'une observation supplémentaire, correspondant alternativement : 1° au quantile de la distribution à 2.5% ; 2° à la médiane de la distribution ; 3° au quantile de la distribution à 97.5%. L'exercice a été effectué pour des échantillons de 30, 50 et 100 observations et pour des valeurs du paramètre ξ de -0.1, -0.25 et -0.4. La table (1.3) reprend l'écart type moyen de la variation des paramètres σ et ξ ainsi que de la variance estimée des résidus.

L'estimateur du paramètre ξ par la méthode alternative est généralement beaucoup moins affecté par l'ajout d'une observation supplémentaire que l'estimateur de la méthode classique. La robustesse de la méthode alternative au niveau de l'estimation de ce paramètre est particulièrement frappante face à l'ajout d'une observation dont la valeur correspond à la valeur médiane de la distribution. Il n'y a que dans le cas d'échantillons de 30 observations et lorsque la perturbation consiste en l'ajout d'une observation prenant une valeur relativement élevée que la méthode alternative est légèrement moins robuste que la méthode classique.

L'estimateur du paramètre σ obtenu par la méthode alternative est également moins affecté que l'estimateur classique lorsque l'on ajoute une observation dans le centre de la distribution. Par contre, lorsque l'on ajoute une observation dans l'une des deux extrémités de la distribution, l'estimateur du paramètre σ par la méthode alternative est légèrement moins robuste que l'estimateur obtenu par la méthode classique.

Dans tous les cas, la qualité de l'ajustement du modèle statistique telle que mesurée par la variance des résidus est beaucoup moins affectée par l'ajout d'une observation dans le cas de la méthode alternative.

TABLE 1.3 – Ecart type moyen de la variation des estimateurs suite à l'ajout d'une observation

N		$\xi = -0.1$			$\xi = -0.25$			$\xi = -0.4$		
		<i>Bas</i>	<i>Ctre.</i>	<i>Ht.</i>	<i>Bas</i>	<i>Ctre.</i>	<i>Ht.</i>	<i>Bas</i>	<i>Ctre.</i>	<i>Ht.</i>
ξ										
30	Cl.	.149	.125	.096	.144	.116	.096	.142	.115	.100
	Al.	.034	.009	.133	.045	.010	.097	.061	.011	.070
50	Cl.	.096	.084	.060	.090	.076	.060	.085	.069	.060
	Al.	.016	.004	.060	.021	.004	.044	.028	.004	.032
100	Cl.	.056	.052	.041	.053	.047	.041	.049	.042	.040
	Al.	.006	.002	.019	.008	.001	.015	.010	.001	.011
σ										
30	Cl.	.046	.057	.022	.055	.054	.022	.066	.052	.026
	Al.	.053	.022	.043	.065	.019	.033	.076	.018	.028
50	Cl.	.022	.034	.010	.029	.033	.011	.036	.030	.012
	Al.	.030	.013	.028	.037	.012	.021	.043	.011	.017
100	Cl.	.009	.018	.006	.011	.015	.006	.016	.014	.007
	Al.	.014	.007	.016	.017	.006	.012	.020	.005	.009
$Var(\epsilon)$										
30	Cl.	.022	.020	.020	.011	.009	.010	.012	.010	.009
	Al.	.004	.003	.012	.004	.003	.007	.006	.002	.004
50	Cl.	.009	.010	.010	.006	.005	.005	.005	.004	.004
	Al.	.002	.001	.005	.002	.001	.002	.003	.001	.002
100	Cl.	.005	.005	.005	.002	.002	.003	.002	.002	.002
	Alt.	.001	.000	.002	.001	.000	.001	.001	.000	.000

1.3.7 Cas Gumbel

La distribution Gumbel est un cas particulier de la GEV, obtenu lorsque $\xi = 0$ dans (1.1). La fonction de répartition de la Gumbel est donnée par :

$$G(x_t) = \exp \left\{ - \exp \left[- \left(\frac{x_t - \mu}{\sigma} \right) \right] \right\}. \quad (1.12)$$

La fonction de log-vraisemblance (1.3) se simplifie alors en :

$$L(x \mid \mu, \sigma) = -n \log \sigma - \sum_t \exp \left[- \left(\frac{x_t - \mu}{\sigma} \right) \right] - \sum_t \left(\frac{x_t - \mu}{\sigma} \right), \quad (1.13)$$

et la définition des quantiles estimés (1.4) devient :

$$\hat{x}_{(i)} = \hat{\mu} - \hat{\sigma} \log \left(- \log(F_{(i)}(\alpha)) \right). \quad (1.14)$$

Dans cette section, nous comparons les performances des deux méthodes d'estimation pour des échantillons de mêmes tailles que celles utilisées dans les sections précédentes (30, 50, 100 et 200 observations). A nouveau, pour chaque taille, 5000 échantillons sont tirés aléatoirement d'une distribution Gumbel (1.12) avec des valeurs des paramètres fixées à $\mu = 0$ et $\sigma = 1$. Les résultats des estimations des paramètres, des quantiles extrêmes, de la somme et de l'écart-type des résidus sont présentés dans la table 1.4.

Il ressort des résultats que l'ajustement global du modèle AGEV est meilleur, comme en témoigne un écart-type estimé des résidus (*s.e.*) systématiquement plus faible que dans l'estimation classique, ce qui est conforme aux résultats obtenus dans le cas des distributions de Weibull.

Avec la méthode classique, l'estimateur du paramètre μ présente un biais vers le haut tandis que celui de σ est biaisé vers le bas. On a exactement le résultat inverse avec la méthode alternative. Pour les deux méthodes, les biais des estimateurs diminuent bien lorsque la taille de l'échantillon augmente, mais c'est moins rapide dans le cas de la méthode classique si bien que dans des échantillons de 200 observations le biais de l'estimateur du paramètre σ par la méthode classique reste près de trois fois plus grand que celui de la méthode alternative.

Conformément aux résultats obtenus dans le cas des distributions de Weibull, les estimateurs des quantiles obtenus par la méthode d'estimation classique sont systématiquement biaisés vers le bas. Par contre, ceux de la méthode alternative sont cette fois systématiquement biaisés vers le haut. En valeur absolue, les biais des estimateurs des quantiles de la méthode alternative sont nettement plus faibles, en particulier dans des échantillons de taille plus réduite. Il en résulte qu'une pondération des deux méthodes donne une estimation (quasi) sans biais des périodes de retour élevées. Pour illustrer cette proposition, on a calculé pour chacun des échantillons (20 000) les périodes de retour correspondant à des probabilités allant de 0.10 (période de retour de 10 ans) à 0.005 (période de retour de 200 ans). Le biais moyen avec la méthode classique est de 0.118 (écart-type de 0.140) et de 0.051 (écart-type de 0.061) avec l'AGEV. On peut alors chercher quelle pondération minimise le biais moyen. On trouve que la taille de l'échantillon n'a qu'un effet négligeable et une règle fort simple se dégage : donner un poids de 30% à la méthode classique et de 70% à la méthode alternative (le R^2 entre quantiles vrais et quantiles ainsi pondérés est de 0.9999!). Le biais moyen des quantiles pondérés est de -0.0003 (écart-type de 0.001). Nous n'avons pas encore de réponse satisfaisante au pourquoi du signe opposé des biais entre les deux méthodes.

En ce qui concerne le paramètre α , on retrouve à quelques nuances près les mêmes résultats que dans le cas Weibull. Avec la méthode alternative, α est en moyenne très proche de zéro avec une distribution quasi uniforme. Avec la méthode classique la valeur moyenne de α est légèrement réduite par rapport aux résultats pour la distribution Weibull, celle-ci se rapprochant de 0.2. Comme la valeur de zéro est la plus couramment utilisée dans la pratique, nous présentons aussi dans la table 1.5 les résultats auxquels elle conduit. Dans ce cas, l'écart-type estimé des résidus est sensiblement plus élevé, davantage pour des échantillons de petite taille et pour la méthode classique. Par contre, les biais des estimateurs des paramètres et des quantiles obtenus par la méthode AGEV ne sont que légèrement affectés et les différences deviennent négligeables lorsque la taille de l'échantillon augmente.

Les valeurs vraies des quantiles de la distribution correspondant aux périodes de retour de 25, 50, 100 et 200 ans, dont les estimations par les deux méthodes figurent dans les tables 1.4 et 1.5, sont respectivement de 3.199; 3.902; 4.600 et 5.296.

TABLE 1.4 – Résultats des simulations pour la Gumbel, α libre.

$N =$	30		50		100		200	
	Clas.	Alter.	Clas.	Alter.	Clas.	Alter.	Clas.	Alter.
μ	.003	-.009	.003	-.004	.003	-.001	.003	.001
	<i>.186</i>	<i>.188</i>	<i>.146</i>	<i>.147</i>	<i>.105</i>	<i>.106</i>	<i>.074</i>	<i>.075</i>
σ	.941	1.028	.960	1.018	.976	1.010	.984	1.006
	<i>.129</i>	<i>.163</i>	<i>.102</i>	<i>.127</i>	<i>.075</i>	<i>.091</i>	<i>.053</i>	<i>.065</i>
$\sum \epsilon$	.675	.000	.867	.000	1.202	.000	1.643	.000
<i>s.e.</i>	.207	.186	.175	.159	.138	.127	.109	.100
Q_{25}	3.013	3.279	3.075	3.252	3.125	3.231	3.152	3.218
Q_{50}	3.674	4.002	3.750	3.968	3.811	3.942	3.845	3.925
Q_{100}	4.331	4.719	4.421	4.678	4.493	4.648	4.532	4.627
Q_{200}	4.986	5.434	5.089	5.386	5.172	5.321	5.217	5.327

TABLE 1.5 – Résultats des simulations pour la Gumbel, $\alpha = 0$.

$N =$	30		50		100		200	
	Clas.	Alter.	Clas.	Alter.	Clas.	Alter.	Clas.	Alter.
μ	.003	-.007	.003	-.003	.003	-.001	.003	.001
	<i>.186</i>	<i>.191</i>	<i>.146</i>	<i>.149</i>	<i>.105</i>	<i>.107</i>	<i>.074</i>	<i>.076</i>
σ	.941	1.022	.960	1.017	.976	1.011	.984	1.007
	<i>.129</i>	<i>.148</i>	<i>.102</i>	<i>.120</i>	<i>.075</i>	<i>.090</i>	<i>.053</i>	<i>.065</i>
$\sum \epsilon$	.033	.000	.025	.000	.016	.000	.010	.000
$s.e.$	.240	.203	.200	.174	.157	.139	.122	.110
Q_{25}	3.013	3.262	3.075	3.249	3.125	3.234	3.152	3.222
Q_{50}	3.674	3.981	3.750	3.965	3.811	3.945	3.845	3.930
Q_{100}	4.331	4.695	4.421	4.675	4.493	4.651	4.532	4.633
Q_{200}	4.986	5.405	5.089	5.382	5.172	5.355	5.217	5.334

1.4 Extension 1 : régression multivariée

Dans beaucoup de domaines, l'intérêt principal du chercheur porte sur la mesure des effets de diverses variables explicatives. C'est notamment le cas en économie pour ce qu'on appelle les régressions hédoniques. Une manière simple d'incorporer des variables exogènes consiste à rendre le paramètre de position (μ) du modèle (1.1) fonction de ces variables. En notant par Z_k le vecteur des k variables explicatives d'intérêt et par β_k le vecteur des k paramètres associés, le modèle (1.1) devient :

$$G(x) = \exp \left\{ - \left[1 + \xi \left(\frac{x - \mu_0 - Z'_k \beta_k}{\sigma} \right) \right]^{-1/\xi} \right\}. \quad (1.15)$$

L'estimation des paramètres de (1.15) peut à nouveau s'effectuer par la méthode du maximum de vraisemblance suivant le schéma présenté dans la section 1.2. Dans la méthode AGEV, on procède en deux étapes. Dans la première étape, on estime la relation entre x et Z_k par moindres carrés ordinaires :

$$x = \mu_0 + Z'_k \beta_k + \epsilon. \quad (1.16)$$

Ensuite, après classement, on postule que les aléas ϵ suivent une distribution GEV à un bruit blanc près et on estime alors par la méthode du maximum de vraisemblance le modèle :

$$x_{(i)} = \mu_0 + Z'_{(i)} \beta - \frac{\sigma}{\xi} \left[1 - (-\log(F_{(i)}(\alpha)))^{-\xi} \right] + \nu_{(i)}, \quad (1.17)$$

où, comme précédemment, $F_{(i)}(\alpha) = \frac{i-\alpha}{n+1-2\alpha}$.

Dans ce contexte, on peut trouver une interprétation particulière de la méthode d'estimation alternative. En effet, la relation (1.16) peut être considérée comme une approximation linéaire (généralement log-linéaire pour des variables économiques) dans le voisinage d'un point Z_0 , tandis que (1.17) formule l'hypothèse que la distance entre tout $\hat{x}_{(i)}$ et $\hat{x}_{(0)}$ est adéquatement approximée par une distribution GEV. En quelque sorte, l'AGEV présente une forte analogie avec les méthodes d'approximation des résidus qui furent proposées en économétrie appliquée dans la foulée du développement des formes fonctionnelles flexibles au sens de Diewert¹².

1.5 Extension 2 : données de type panel

Le cadre d'analyse de la section 1.2 peut également être étendu pour tenir compte de la dimension spatiale présente dans les données. Ainsi, on peut traiter assez facilement des données composées de k séries de n observations de maxima (annuels ou autres) correspondant à k sites différents (ou k variétés du même bien ou service).

1.5.1 Méthode classique d'estimation

Par exemple en hydrologie, à partir d'observations de crues annuelles maximales pour k sites différents d'un même bassin, il est tout à fait possible d'estimer une GEV unique en laissant la possibilité aux paramètres de prendre une valeur distincte par site, c'est à dire d'estimer le modèle :

$$G(x_{jt}) = \exp \left\{ - \left[1 + \sum_j^k D_j \xi_j \left(\frac{x_{jt} - \mu_j}{\sigma_j} \right) \right]^{-1/\xi_j} \right\}, \quad (1.18)$$

la variable binaire D_j prenant la valeur 1 si l'observation provient du site j et 0 sinon. Le modèle (1.18) généralise l'approche de Northrop (2004) qui impose un paramètre ξ identique à travers les sites.

Si la spécification (1.18) constitue un net progrès par rapport à la méthode consistant à estimer les distributions site par site, elle a l'inconvénient majeur de considérer comme indépendantes les données observées simultanément dans les différents sites.

1.5.2 Méthode d'estimation alternative

Un avantage substantiel de la méthode d'estimation alternative réside dans la possibilité de procéder assez facilement à une estimation simultanée des

12. Voir par exemple Simmons et Weiserbs (1979), Galant (1981) et Barnett *et al.* (1985), ces auteurs utilisant, respectivement, des séries de Taylor, de Fourier et de Laurent.

différentes équations permettant de passer outre l'inconvénient rencontré par la méthode d'estimation traditionnelle. Il suffit en effet de réécrire le modèle de la sorte :

$$x_{jt} = \mu_j - \frac{\sigma_j}{\xi_j} \left[1 - (-\log(F_{jt}(\alpha)))^{-\xi_j} \right] + \epsilon_{jt} \quad j = 1, k, \quad (1.19)$$

avec un vecteur ϵ d'espérance nulle et de matrice de variance covariance (Σ) de forme non spécifiée a priori. Les $F_j(\alpha)$ sont les fréquences empiriques (1.5) correspondant au classement propre à chaque site j .

Le modèle (1.19) permet de tenir compte d'une corrélation contemporaine entre les résidus des différents sites j , ce qui peut constituer une source d'efficacité dans l'estimation des paramètres. Il est possible de tester, par exemple à l'aide d'un test du multiplicateur de Lagrange (Breusch et Pagan, 1980), si la matrice de variance-covariance des résidus est diagonale ou non. De toute façon, un gain d'efficacité est obtenu si les corrélations ne sont pas nulles. Il est également possible de tester, à l'aide d'une procédure standard, si le paramètre de courbure (ξ) n'est pas statistiquement différent à travers les sites. Par ailleurs, il est également possible d'imposer des contraintes de plausibilité théorique permettant de s'assurer, par exemple, que pour toute période de retour raisonnable les débits estimés d'un cours d'eau augmentent bien d'amont en aval.

1.5.3 Résultats de la méthode alternative sur base de simulations de Monte Carlo

Afin de tester les mérites de la méthode alternative dans le cas de données en panel, nous avons à nouveau procédé à des estimations sur des échantillons simulés. Des séries de trois échantillons de 30 données (cas assez fréquent en hydrologie) ont été tirées aléatoirement de trois distributions GEV possédant un paramètre ξ identique et des valeurs distinctes des paramètres μ et σ . Un choc commun a été rajouté aux observations des trois sites afin d'avoir une structure de dépendance dans les données.

Lorsqu'on estime simultanément ces distributions GEV par la méthode alternative (1.19), on obtient les résultats suivants conformes aux attentes (compte tenu des résultats précédents et des propriétés des estimateurs *SUR*).

Primo, que l'on prenne en compte la dépendance contemporaine des résidus ou non, la méthode alternative présente des biais plus petits que ceux de la méthode classique. Ainsi, par exemple, pour un ξ de -0.3 le biais de l'estimateur du quantile correspondant à une période de retour de 100 ans est de l'ordre de 10% vers le bas pour la méthode traditionnelle, alors qu'il est négligeable pour la méthode alternative (cf. table 1.2).

Secundo, les écarts-types des estimateurs des paramètres et des quantiles élevés obtenus sur base des observations des trois sites sont beaucoup plus faibles que ceux obtenus dans les estimations réalisées pour chaque site séparément.

Tertio, la prise en compte de la corrélation contemporaine des aléas (méthode Zellner) offre un léger gain supplémentaire en précision.

Les propriétés de la méthode alternative seront explorées plus en détail dans le chapitre 2 qui sera consacré à la modélisation des crues annuelles maximales observées en différents endroits d'un même bassin hydrologique.

1.6 Conclusion

Dans ce premier chapitre, nous avons exposé une méthode alternative, dénotée AGEV, d'estimation de la fonction GEV par maximum de vraisemblance.

Pour un chercheur qui travaille sur un échantillon de petite taille, au sein duquel il y a des covariables et les observations sont réalisées à différents points de l'espace, cette méthode AGEV s'avère très intéressante. Nous avons en effet montré qu'elle est très simple à estimer, qu'elle possède de meilleures propriétés de convergence numérique, et est plus pratique au stade de l'inférence statistique que la méthode d'estimation classique. En outre, elle est aussi plus flexible dès qu'il s'agit d'intégrer l'effet de covariables et de liens de simultanéité. Nous avons établi au travers d'une analyse du type Monte Carlo que, par rapport à la démarche classique, elle fournissait, pour des échantillons de petite taille, une réduction très substantielle du biais des estimateurs des paramètres et des quantiles extrêmes.

Notons que les avantages de la méthode AGEV se réduisent au fur et à mesure que la taille de l'échantillon augmente et asymptotiquement les deux méthodes donnent des résultats identiques.

Les chapitres suivants seront consacrés à l'application de la méthode AGEV dans divers domaines où la distribution GEV constitue une spécification pertinente.

1.A Méthode des moments pondérés

Les moments pondérés d'une distribution¹³ sont la généralisation des moments traditionnels d'une fonction de distribution.

Pour une variable aléatoire X possédant une fonction de distribution $F(x) = P(X \leq x)$, les moments pondérés sont définis par :

$$M_{p,r,s} = E[X^p \{F(X)\}^r \{1 - F(X)^s\}], \quad (1.20)$$

où p , r et s sont des nombres réels. Les quantités $M_{p,1,0}$ ($p = 1, 2$) sont les moments traditionnels non centrés de la variable X . Cependant, les moments de poids unitaire ($p = 1$) sont parfois préférables pour estimer les paramètres de la distribution étant donné que dans ce cas la présence uniquement de la première puissance de la variable X dans l'expression (1.20) implique que la relation entre les paramètres et les moments prend souvent une forme beaucoup plus simple que lorsque les moments traditionnels sont utilisés pour estimer les paramètres. Dans ce cas, lorsque la valeur de s est fixée à 0 et que l'on ne prend que des nombres entiers pour r , l'expression (1.20) se simplifie en :

$$\beta_r = M_{1,r,0} = E[X \{F(X)\}^r]. \quad (1.21)$$

Si l'on classe les observations par ordre croissant ($x_1 \leq x_2 \leq \dots \leq x_n$), un estimateur consistant de β_r est donné par :

$$\hat{\beta}_r = \frac{1}{n} \sum_{j=1}^n p_{j,n}^r x_j, \quad (1.22)$$

où $p_{j,n}^r$ est une définition particulière de la fréquence empirique des observations, par exemple $p_{j,n}^r = (j - \alpha)/(n + 1 - 2\alpha)$ avec $\alpha < 1$.

Pour la distribution GEV (1.1) les moments pondérés sont définis par :

$$\beta_r = (r + 1)^{-1} (\mu - \sigma \{1 - (r + 1)^\xi \Gamma(1 - \xi)\} / \xi). \quad (1.23)$$

A partir de l'expression (1.23), il est possible d'obtenir les trois expressions suivantes :

$$\beta_0 = (\mu - \sigma \{1 - \Gamma(1 - \xi)\} / \xi), \quad (1.24)$$

$$2\beta_1 - \beta_0 = \sigma \Gamma(1 - \xi) (2^\xi - 1) \xi, \quad (1.25)$$

$$(3\beta_2 - \beta_0) / (2\beta_1 - \beta_0) = (1 - 3^\xi) / (1 - 2^\xi). \quad (1.26)$$

13. Cette section se base sur l'exposé théorique de Hosking *et al.* (1985).

Les estimateurs des paramètres (μ, σ, ξ) par la méthode des moments pondérés sont trouvés en résolvant numériquement le système (1.24)-(1.26) lorsque les β_r sont remplacés par leurs estimateurs $\hat{\beta}_r$ dont la définition est donnée par (1.22).

Nous avons évalué les performances empiriques de la méthode des moments pondérés sur base de simulations de Monte Carlo. Notre analyse a permis de mettre en évidence les résultats suivants.

Primo, conformément aux travaux publiés par Hosking *et al.* (1985), nos résultats montrent que cette méthode affiche une efficacité accrue des estimateurs en petit échantillon par rapport à la méthode d'estimation traditionnelle.

Secundo, nous avons constaté une grande sensibilité des résultats au choix de la fonction pour la fréquence empirique, tout particulièrement pour les estimations en petits échantillons. En effet, lorsque la fréquence empirique (1.5) est utilisée avec le paramètre α fixé à une valeur de 0.5 ou de 0, le biais des estimateurs des quantiles augmente sensiblement par rapport au 0.35 de Hosking *et al.* (1985).

Tertio, l'ajustement global du modèle GEV lorsqu'il est estimé par cette méthode est nettement moins bon que l'ajustement de la méthode AGEV. En effet, dans les simulations réalisées, la variance des résidus est dans tous les cas sensiblement plus grande lorsque les estimations sont réalisées par la méthode des moments en comparaison de la méthode AGEV.

1.B Résultats des simulations avec $\alpha = 0$

TABLE 1.6 – Paramètres et résidus

$\xi =$	-0.1		-0.2		-0.3		-0.4		-0.5	
	Cl.	Al.	Cl.	Al.	Cl.	Al.	Cl.	Al.	Cl.	Al.
$N=30$										
μ	.029	-.007	.032	-.010	.035	-.012	.037	-.014	.039	-.015
	<i>.211</i>	<i>.203</i>	<i>.208</i>	<i>.202</i>	<i>.206</i>	<i>.201</i>	<i>.203</i>	<i>.202</i>	<i>.201</i>	<i>.203</i>
σ	.950	1.029	.954	1.031	.959	1.032	.964	1.034	.970	1.035
	<i>.144</i>	<i>.158</i>	<i>.142</i>	<i>.150</i>	<i>.144</i>	<i>.149</i>	<i>.147</i>	<i>.151</i>	<i>.152</i>	<i>.156</i>
ξ	-.156	-.125	-.261	-.230	-.365	-.336	-.469	-.442	-.574	-.550
	<i>.160</i>	<i>.148</i>	<i>.157</i>	<i>.146</i>	<i>.155</i>	<i>.152</i>	<i>.157</i>	<i>.165</i>	<i>.159</i>	<i>.174</i>
$\sum \epsilon$	.553	.000	.143	.000	-.214	.000	-.542	.000	-.848	.000
<i>s.e.</i>	.191	.152	.172	.140	.161	.133	.157	.129	.157	.128
$ \sum \epsilon $	.671	.000	.462	.000	.455	.000	.623	.000	.883	.000
$N=50$										
μ	.020	-.004	.022	-.006	.024	-.007	.026	-.007	.028	-.007
	<i>.160</i>	<i>.157</i>	<i>.158</i>	<i>.156</i>	<i>.157</i>	<i>.156</i>	<i>.155</i>	<i>.156</i>	<i>.154</i>	<i>.157</i>
σ	.966	1.019	.969	1.020	.972	1.022	.975	1.023	.979	1.024
	<i>.109</i>	<i>.120</i>	<i>.107</i>	<i>.113</i>	<i>.108</i>	<i>.111</i>	<i>.110</i>	<i>.112</i>	<i>.114</i>	<i>.116</i>
ξ	-.136	-.114	-.238	-.218	-.341	-.322	-.443	-.427	-.547	-.534
	<i>.107</i>	<i>.103</i>	<i>.104</i>	<i>.101</i>	<i>.102</i>	<i>.105</i>	<i>.103</i>	<i>.115</i>	<i>.106</i>	<i>.130</i>
$\sum \epsilon$	.760	.000	.265	.000	-.170	.000	-.554	.000	-.895	.000
<i>s.e.</i>	.154	.129	.137	.118	.128	.111	.125	.107	.126	.107
$ \sum \epsilon $	.796	.000	.446	.000	.389	.000	.596	.000	.904	.000
$N=100$										
μ	.013	-.001	.014	-.002	.016	-.002	.017	-.002	.018	-.002
	<i>.112</i>	<i>.112</i>	<i>.111</i>	<i>.111</i>	<i>.110</i>	<i>.110</i>	<i>.109</i>	<i>.111</i>	<i>.108</i>	<i>.112</i>
σ	.979	1.009	.981	1.010	.982	1.011	.984	1.012	.986	1.012
	<i>.076</i>	<i>.084</i>	<i>.075</i>	<i>.079</i>	<i>.075</i>	<i>.077</i>	<i>.077</i>	<i>.078</i>	<i>.079</i>	<i>.081</i>
ξ	-.120	-.106	-.221	-.208	-.323	-.311	-.424	-.414	-.525	-.517
	<i>.068</i>	<i>.069</i>	<i>.065</i>	<i>.066</i>	<i>.063</i>	<i>.068</i>	<i>.063</i>	<i>.074</i>	<i>.064</i>	<i>.084</i>
$\sum \epsilon$	1.052	.000	.427	.000	-.118	.000	-.591	.000	-.999	.000
<i>s.e.</i>	.114	.100	.099	.089	.092	.083	.090	.080	.091	.081
$ \sum \epsilon $	1.066	.000	.526	.000	.346	.000	.609	.000	.999	.000
$N=200$										
μ	.009	.000	.010	.000	.011	.000	.012	.000	.012	.000
	<i>.079</i>	<i>.079</i>	<i>.078</i>	<i>.078</i>	<i>.078</i>	<i>.078</i>	<i>.077</i>	<i>.078</i>	<i>.076</i>	<i>.079</i>
σ	.986	1.003	.986	1.003	.987	1.004	.988	1.004	.989	1.004
	<i>.053</i>	<i>.059</i>	<i>.052</i>	<i>.055</i>	<i>.052</i>	<i>.053</i>	<i>.053</i>	<i>.054</i>	<i>.055</i>	<i>.056</i>
ξ	-.109	-.101	-.210	-.202	-.312	-.304	-.412	-.405	-.513	-.507
	<i>.045</i>	<i>.046</i>	<i>.043</i>	<i>.044</i>	<i>.041</i>	<i>.046</i>	<i>.040</i>	<i>.050</i>	<i>.041</i>	<i>.057</i>
$\sum \epsilon$	1.360	.000	.602	.000	-.060	.000	-.637	.000	-1.131	.000
<i>s.e.</i>	.084	.076	.072	.066	.066	.061	.065	.059	.066	.060
$ \sum \epsilon $	1.390	.000	.715	.000	.361	.000	.649	.000	1.131	.000

TABLE 1.7 – Biais des quantiles extrêmes

$\xi =$	-0.1		-0.2		-0.3		-0.4		-0.5	
	Cl.	Al.	Cl.	Al.	Cl.	Al.	Cl.	Al.	Cl.	Al.
<i>N=30</i>										
Q_{10}	-.175	.004	-.150	-.005	-.127	-.012	-.105	-.017	-.086	-.020
Q_{25}	-.275	.007	-.232	-.009	-.193	-.019	-.158	-.024	-.127	-.026
Q_{50}	-.341	.020	-.284	-.007	-.233	-.019	-.187	-.024	-.148	-.026
Q_{100}	-.394	.045	-.326	.003	-.264	-.015	-.209	-.022	-.163	-.024
Q_{200}	-.431	.085	-.357	.020	-.286	-.077	-.224	-.016	-.173	-.019
<i>N=50</i>										
Q_{10}	-.118	.007	-.101	.000	-.085	-.005	-.070	-.009	-.056	-.012
Q_{25}	-.190	.010	-.160	-.004	-.132	-.012	-.107	-.017	-.085	-.019
Q_{50}	-.243	.016	-.201	-.005	-.163	-.015	-.130	-.020	-.102	-.022
Q_{100}	-.290	.028	-.236	-.002	-.188	-.016	-.147	-.021	-.114	-.023
Q_{200}	-.331	.048	-.267	.003	-.209	-.015	-.161	-.021	-.122	-.022
<i>N=100</i>										
Q_{10}	-.069	.007	-.059	.002	-.049	-.001	-.041	-.004	-.032	-.006
Q_{25}	-.113	.010	-.096	.001	-.079	-.005	-.064	-.009	-.050	-.011
Q_{50}	-.147	.015	-.122	.000	-.098	-.007	-.078	-.011	-.060	-.013
Q_{100}	-.179	.022	-.146	.001	-.116	-.008	-.090	-.013	-.068	-.014
Q_{200}	-.209	.033	-.167	.004	-.130	-.008	-.099	-.013	-.074	-.014
<i>N=200</i>										
Q_{10}	-.039	.006	-.034	.003	-.029	.000	-.023	-.001	-.018	-.003
Q_{25}	-.064	.010	-.055	.003	-.046	-.001	-.037	-.003	-.029	-.004
Q_{50}	-.082	.015	-.069	.004	-.057	-.001	-.045	-.004	-.035	-.005
Q_{100}	-.100	.021	-.083	.006	-.067	-.001	-.052	-.004	-.040	-.005
Q_{200}	-.116	.030	-.095	.009	-.076	.000	-.058	-.004	-.043	-.005

Chapitre 2

Application aux crues du bassin de l'Ourthe

2.1 Introduction

L'objectif de ce chapitre est d'appliquer la méthode AGEV sur des données hydrologiques et de comparer les résultats avec ceux de la méthode d'estimation classique de la GEV.

Les données hydrologiques possèdent des caractéristiques particulières. Dans la plupart des cas, les observations, d'une part, couvrent une période de temps relativement courte et, d'autre part, sont disponibles pour plusieurs points du même bassin.

La démarche habituelle consiste à traiter les différents sites de manière indépendante. Northrop (2004) fait exception en les estimant de manière simultanée, mais en ignorant toute covariance contemporaine des aléas. La GEV, au contraire de l'AGEV ne permet pas d'agir d'autrement.

En outre, pour être crédible, un modèle statistique appliqué à des crues doit être en mesure de délivrer des résultats qui garantissent que la prévision d'un quantile pour un site soit, pour toute période de retour raisonnable, inférieure à celle pour un site se trouvant plus en aval sur la rivière (ou que le débit d'un affluent n'excède pas celui de son confluent). Cette propriété physique de monotonie n'est pas garantie lorsque les paramètres de la distribution GEV sont estimés par la méthode classique, mais elle peut aisément être imposée dans le cadre de l'AGEV¹.

1. Le principe de traiter globalement les crues d'un cours d'eau ou d'un bassin n'a évidemment rien d'original [voir par exemple Chow *et al.* (1988), Reiss et Thomas (1997)]. Il en va de même pour l'imposition des contraintes physiques. A cet égard, on peut aussi

Notre application porte sur les débits maxima de trois sites de l’Ourthe pour la période utilisée par le Ministère de la Région wallonne lors de son élaboration des zones inondables. Les données sont présentées à la section 2.2. Dans la section 2.3, on estime une distribution GEV par la méthode classique et la méthode alternative pour chacun des sites pris de manière indépendante. L’estimation conjointe des sites suivant l’approche proposée par Northrop est réalisée à la section 2.4. Ensuite, nous procédons à l’estimation simultanée en appliquant la méthode AGEV (section 2.5). Différentes hypothèses restrictives sont testées à la section suivante de manière à sélectionner le “meilleur” modèle. Enfin, à la section 2.7 l’analyse est reconduite en ajoutant les dernières observations disponibles.

2.2 Les données

L’analyse porte sur les crues de l’Ourthe. Nous disposons du débit le plus élevé de chaque année hydrologique depuis 1975 pour les sites de Tabreux, Sauhaid et Angleur qui se succèdent d’amont en aval dans cet ordre . Les sites de Nisramont, Hotton et Durbuy, où les limnigraphes ont été installés en 1978 (1993 pour Durbuy), n’ont pas été pris en compte pour ne pas raccourcir la période d’estimation.

Les débits sont mesurés en mètre cube par seconde. Les statistiques descriptives des données² sont présentées à la table 2.1. Les maxima annuels sont fortement corrélés, avec un coefficient de corrélation minimum de 0.87 entre les sites de Tabreux et Angleur.

TABLE 2.1 – Statistiques descriptives

	Tabreux	Sauhaid	Angleur
Moy.	216.98	397.34	490.60
Ec.-type	94.68	165.85	184.64

Dans un premier temps, l’analyse est réalisée pour la période 1975-2004 de manière à comparer nos résultats avec ceux du Ministère de la Région wallonne lors de leur étude préalable à l’établissement des zones inondables de la région. Ensuite, l’estimation est faite pour toute la période disponible (1975-2009).

recourir à une approche bayésienne telle qu’initiée par Kuczera (1982) dont un résumé est donné dans Reiss et Thomas. Toutefois, le choix de la distribution (lognormale ou gamma inverse) manque de fondement théorique.

2. Nous remercions vivement Paul Dewil, Directeur au Service d’études hydrologiques du Ministère wallon de l’Équipement et des Transports, d’avoir eu l’amabilité de nous fournir les données.

2.3 Estimations indépendantes (1975-2004)

Dans une première étape, l'estimation de la distribution GEV (1.1) a été réalisée pour chaque site séparément par la méthode traditionnelle, qui consiste à maximiser la fonction de vraisemblance (1.3), ainsi que par la méthode alternative, impliquant l'estimation du modèle (1.11). Les résultats sont présentés dans la table 2.2. Pour chaque site et pour chacune des deux méthodes, les paramètres de la distribution GEV ont été estimés et les quantiles correspondant à une période de retour de 25, 50, 100 et 200 ans ainsi que la somme des résidus et la somme des carrés des résidus ont été calculés.

Même si les échantillons sont de taille réduite dans cette application, il apparaît que les résultats des estimations ne sont que très peu affectés par la valeur du paramètre α dans (1.5) si bien que nous l'avons systématiquement fixé à 0 dans chacune des estimations par souci de simplicité.

La valeur estimée du paramètre de courbure (ξ) pour chacun des sites prend une valeur négative, tendant à indiquer que les crues maximales suivent une distribution appartenant à la famille des Weibull. Toutefois, pour les sites de Tabreux et Sauhaid, la valeur de ce paramètre est très faible, et n'apparaît pas du tout significative.

La valeur estimée du paramètre ξ ainsi que celle du paramètre de position (μ) n'est pas sensiblement différente en fonction de la méthode utilisée, alors que la valeur estimée du paramètre de dispersion (σ) est systématiquement plus élevée lorsque l'estimation est réalisée par la méthode AGEV.

De manière cohérente avec les résultats des simulations présentés dans le chapitre 1, les prévisions des quantiles élevés réalisées sur base des résultats de la méthode AGEV sont systématiquement supérieures aux prévisions obtenues par la méthode classique.

Sur base de la somme des carrés des résidus, l'AGEV est clairement la méthode qui permet d'obtenir le meilleur ajustement du modèle aux données. En outre, contrairement à l'AGEV pour laquelle la somme des résidus est pratiquement nulle, la somme des résidus est largement différente de zéro dans le cas de la méthode classique.

Par ailleurs, la méthode AGEV passe généralement mieux les tests statistiques standards de mauvaise spécification. La table 2.3 reprend, pour chaque site, les résultats du test de normalité de Jarque-Bera ($JB, \chi^2_{(2)}$), du test d'hétéroscédasticité du multiplicateur de Lagrange ($LMhet, \chi^2_{(1)}$) et du test de mauvaise spécification de Ramsey d'ordre 4 ($Rsey4, \chi^2_{(3)}$). Il ressort de la table 2.3, que l'hypothèse de normalité des résidus n'est jamais rejetée

lorsque les estimations sont réalisées par la méthode AGEV tandis qu'elle est rejetée pour les sites de Sauhaid et Tabreux avec la méthode classique. L'homoscédasticité des résidus n'est rejetée que pour le site de Sauhaid avec la méthode AGEV et est rejetée pour les sites de Sauhaid et d'Angleur dans le cas de la méthode classique. De même, le test de Ramsey d'ordre 4 indique un problème de mauvaise spécification de la forme fonctionnelle pour le seul site de Sauhaid lorsque la GEV est estimée par la méthode alternative tandis que ce test est significatif pour tous les sites avec la méthode classique.

Les Q-Q plot des estimations par site de la GEV sont présentés en annexe. Il apparaît clairement que l'ajustement de la distribution lorsqu'elle est estimée par la méthode AGEV est toujours meilleur que celui de la méthode classique de manière globale, et en particulier dans les queues de la distribution.

TABLE 2.2 – Résultats des estimations, période 1975-2004

$N = 30$	Tabreux		Sauhaid		Angleur	
	Clas.	Alter.	Clas.	Alter.	Clas.	Alter.
μ	175.04 <i>15.30</i>	174.49 <i>17.87</i>	324.40 <i>26.99</i>	322.70 <i>31.32</i>	421.75 <i>34.07</i>	413.78 <i>37.56</i>
σ	77.52 <i>14.12</i>	88.48 <i>15.87</i>	136.20 <i>25.32</i>	154.97 <i>27.86</i>	171.58 <i>33.91</i>	187.40 <i>30.68</i>
ξ	-.05 <i>.29</i>	-.08 <i>.18</i>	-.06 <i>.32</i>	-.08 <i>.18</i>	-.22 <i>.28</i>	-.19 <i>.16</i>
$\sum \epsilon$	96.96	.00	164.26	.00	67.87	.00
Q_{25}	404.18 <i>96.01</i>	425.28 <i>65.26</i>	723.50 <i>179.72</i>	763.30 <i>115.18</i>	814.18 <i>122.59</i>	865.51 <i>90.41</i>
Q_{50}	449.84 <i>138.88</i>	472.61 <i>99.20</i>	802.15 <i>262.59</i>	846.80 <i>175.52</i>	868.93 <i>168.87</i>	933.65 <i>124.16</i>
Q_{100}	493.61 <i>188.91</i>	517.12 <i>147.12</i>	877.23 <i>359.05</i>	925.42 <i>260.98</i>	915.42 <i>217.73</i>	993.02 <i>166.14</i>
Q_{200}	535.72 <i>245.52</i>	559.14 <i>213.30</i>	949.19 <i>467.90</i>	999.76 <i>379.47</i>	955.07 <i>267.39</i>	1044.98 <i>216.01</i>
SSR	8107	4930	22958	12919	20001	8606

TABLE 2.3 – Tests de spécification

	Tabreux		Sauhaid		Angleur	
	Clas.	Alter.	Clas.	Alter.	Clas.	Alter.
JB	19.89	5.62	11.51	.35	3.39	1.38
$LMhet$	2.88	2.64	8.84	13.97	7.95	3.52
$Rsey4$	8.21	2.00	21.94	9.68	10.10	1.54

2.4 Estimation jointe, méthode classique

Avec la méthode classique, il est possible d'estimer les paramètres des différents sites conjointement. Il suffit en effet d'estimer une distribution GEV unique pour l'ensemble des sites avec des paramètres distincts par site. Le modèle prend alors la forme :

$$G(x_{jt}) = \exp \left\{ - \left[1 + \sum_{j=1}^3 \xi_j S_{jt} \left(\frac{x_{jt} - \mu_j}{\sigma_j} \right) \right]^{-1/\xi_j} \right\}, \quad (2.1)$$

les variables binaires S_{jt} prenant la valeur 1 si l'observation t provient du site j et 0 sinon. Dans la littérature appliquée en hydrologie, un modèle de ce type a notamment été utilisé par Northrop (2004), celui-ci imposant a priori un paramètre ξ identique pour toutes les observations.

Etant donné que la dépendance entre les observations des différents sites n'est pas prise en compte par le modèle, les coefficients estimés sont identiques à ceux obtenus par les estimations réalisées site par site dont les résultats se trouvent dans la table 2.2.

La contrainte d'un paramètre ξ constant à travers les trois sites ($\chi^2_{(2)}$, *valeur* : .64) n'est pas rejetée par le test traditionnel du rapport de vraisemblance. Il en est de même pour la contrainte d'un paramètre ξ égal à zéro ($\chi^2_{(3)}$, *valeur* : 1.744) pour chacun des sites³. Les résultats de l'estimation du modèle (2.1) avec cette contrainte sont présentés dans la table 2.4.

Lorsque cette contrainte est imposée, les écarts-types des estimateurs des quantiles extrêmes diminuent.

Par ailleurs, pour le site d'Angleur, dont les observations semblaient suivre une distribution de Weibull dans les estimations indépendantes, les prédictions de ces quantiles extrêmes sont sensiblement supérieures à celles présentées dans la table 2.2.

L'ajustement global du modèle est médiocre, avec une somme des carrés des résidus totale qui est sensiblement plus grande que lorsque les estimations de la GEV sont réalisées site par site. La forme du Q-Q plot (figure 2.1) renforce le sentiment que cette spécification n'est pas très appropriée pour

3. Comme expliqué dans le premier chapitre, la distribution de ces tests en petit échantillon n'est pas standard. Il apparaît en effet que les valeurs critiques obtenues par bootstrap sont légèrement plus faibles que les valeurs critiques de la distribution χ^2 standard. Dans le cas présent, les statistiques de test sont de loin inférieures aux valeurs critiques obtenues par bootstrap.

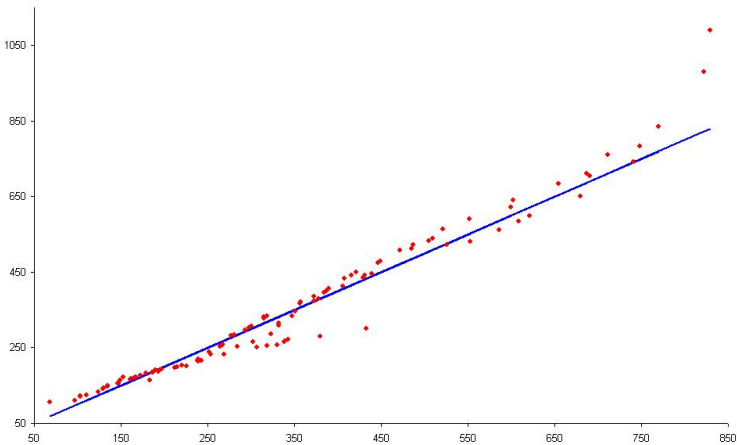
modéliser les données en présence.

Ces mauvaises performances sont probablement dues au fait que la spécification traditionnelle (2.1) rencontre une limite importante. En effet, celle-ci suppose que la corrélation contemporaine entre les résidus des différents sites est nulle. Cette hypothèse nous paraît très restrictive dans le cas de crues maximales provenant de différents endroits d’un même bassin hydrologique, en particulier lorsque les sites sont très proches.

TABLE 2.4 – Estimation jointe, méthode classique

$N = 90$	Tabreux	Sauhaid	Angleur
μ	172.99 <i>14.64</i>	320.40 <i>25.69</i>	402.11 <i>31.11</i>
σ	76.06 <i>11.01</i>	133.46 <i>19.25</i>	161.05 <i>22.54</i>
Q_{25}	416.26 <i>37.52</i>	747.27 <i>65.87</i>	917.24 <i>83.78</i>
Q_{50}	469.76 <i>47.29</i>	841.15 <i>83.10</i>	1030.53 <i>107.34</i>
Q_{100}	522.86 <i>58.90</i>	934.33 <i>103.57</i>	1142.98 <i>134.39</i>
Q_{200}	575.77 <i>72.06</i>	1027.17 <i>126.73</i>	1255.02 <i>164.24</i>
SSR	7655	21125	24271

FIGURE 2.1 – Q-Q plot estimation jointe, méthode classique



2.5 AGEV, estimation simultanée

Un avantage substantiel de la méthode d'estimation alternative (1.11) est qu'elle permet de procéder assez facilement à une estimation simultanée des distributions propres à chaque site en tenant compte de la dépendance contemporaine entre les termes d'erreur à travers les sites. Cette propriété de la méthode alternative est une source d'efficacité dans l'estimation des paramètres. Lorsque l'estimation est réalisée conjointement, le modèle est :

$$x_{j,t} = \mu_j - \frac{\sigma_j}{\xi_j} \left[1 - (-\log(F_{j,t}(\alpha_j)))^{-\xi_j} \right] + \epsilon_{jt}, \quad j = 1, 2, 3 \quad (2.2)$$

avec un vecteur ϵ d'espérance nulle et dont la matrice de variance-covariance (Σ) n'est pas contrainte a priori. Les observations t sont contemporaines dans les trois sites et les $F_{jt}(\alpha_j)$ sont les fréquences empiriques (1.5) correspondant au classement propre à chaque site j .

L'estimation des paramètres du modèle (2.2) s'effectue par la méthode du quasi maximum de vraisemblance. On peut vérifier si la matrice de variance-covariance des résidus (Σ) est diagonale ou non, notamment à l'aide d'un test de Lagrange (Breusch et Pagan, 1980). Lorsque les erreurs contemporaines ne sont pas corrélées à travers les sites, les résultats des estimations sont identiques aux résultats des estimations réalisées site par site. Cependant l'estimation jointe des différentes distributions peut malgré tout être intéressante puisqu'elle permet d'imposer certaines contraintes sur les paramètres à travers les différents sites. Par exemple, on peut tester à l'aide d'une procédure standard si le paramètre de courbure (ξ) n'est pas statistiquement différent à travers les sites. Si la contrainte d'un paramètre unique n'est pas rejetée, l'estimateur du paramètre en question obtenu sur base de l'ensemble des observations est plus efficace que les estimateurs obtenus dans les régressions séparées. Les résultats de l'estimation du système d'équation (2.2) sont présentés dans la table 2.5.

Sur base du test de Breusch et Pagan ($\chi^2_{(3)}$, valeur de 35.99), on rejette clairement l'hypothèse d'une matrice de variance-covariance des résidus diagonale, impliquant qu'il est préférable de réaliser l'estimation des paramètres des différents sites conjointement.

Par ailleurs, il apparaît que la restriction d'un paramètre de courbure constant à travers les différents sites est rejetée sur base du test traditionnel du rapport de la vraisemblance ($\chi^2_{(2)}$, valeur de 7.12).

Par rapport aux résultats des estimations réalisées pour chaque site séparément présentés dans la table 2.2, les écarts-types des estimateurs sont très légèrement plus faibles lorsque les estimations sont réalisées conjointement, tandis que la somme des carrés des résidus est légèrement plus élevée. En outre,

la valeur des coefficients estimés change légèrement, induisant des prédictions pour les quantiles extrêmes quelque peu différentes. Par exemple, pour le site de Sauhaid la prédiction du quantile correspondant à une période de retour de 100 ans est approximativement 5% plus faible par rapport aux résultats obtenus sur base de l'estimation réalisée à partir de ce site uniquement.

TABLE 2.5 – AGEV, estimations simultanées

$N = 90$	Tabreux	Sauhaid	Angleur
μ	176.02 <i>17.4</i>	328.39 <i>30.76</i>	415.23 <i>37.54</i>
σ	86.18 <i>15.23</i>	152.98 <i>26.14</i>	187.56 <i>30.52</i>
ξ	-.08 <i>0.18</i>	-.12 <i>0.17</i>	-.20 <i>0.16</i>
Q_{25}	417.61 <i>61.48</i>	734.02 <i>94.06</i>	858.72 <i>85.82</i>
Q_{50}	462.54 <i>92.30</i>	804.01 <i>135.96</i>	923.75 <i>116.59</i>
Q_{100}	504.58 <i>135.09</i>	867.83 <i>191.31</i>	979.88 <i>154.42</i>
Q_{200}	544.08 <i>193.09</i>	926.25 <i>261.99</i>	1028.55 <i>198.73</i>
SSR	5189	15399	8736

2.6 AGEV, estimation avec contraintes

Partant du modèle général (2.2), plusieurs tests de restriction sur les paramètres ont été effectués afin de trouver la spécification la plus parcimonieuse. Le modèle final retenu est un modèle dans lequel le paramètre ξ est égal à 0 dans les deux premiers sites (Tabreux et Sauhaid) et est laissé libre pour le dernier site (Angleur). Cet ensemble de restriction est le seul à ne pas être rejeté par le test traditionnel du rapport de vraisemblance.

Ce modèle final suggère qu'en amont de la rivière (Tabreux et Sauhaid) les observations suivent une distribution de Gumbel, impliquant qu'il n'existe pas de limite maximale finie pour les crues de ces sites, tandis qu'en aval (Angleur) les observations semblent plutôt suivre une loi appartenant à la famille des distributions de Weibull. Par conséquent, il se peut que pour une période de retour (très) élevée, la prédiction du débit maximal des sites se trouvant en amont soit supérieure à la prédiction pour le site le plus en aval, ce qui va à l'encontre de toute logique physique. Dans le cas présent, il s'avère que la

prédiction du quantile correspondant à une période de retour de 500 ans pour le site de Sauhaid (1191.05) dépasse la prédiction du quantile de même période de retour pour le site d'Angleur (1129.78). Afin d'éviter cette incohérence, il est possible d'imposer une contrainte d'inégalité sur les paramètres entre les équations des sites concernés. Lorsque les observations dans au moins un des deux sites concernés suivent une distribution de Gumbel, comme c'est le cas dans notre exemple, il est trivial d'imposer la contrainte d'inégalité. En effet, pour le site d'Angleur (distribution de Weibull), la définition des quantiles correspondant à une période de retour T est donnée par :

$$\hat{Q}_{(T)}^a = \hat{\mu}^a - \frac{\hat{\sigma}^a}{\hat{\xi}^a} \left(1 - \left(-\log\left(1 - \frac{1}{T}\right) \right)^{-\hat{\xi}^a} \right), \quad (2.3)$$

tandis que pour le site de Sauhaid (distribution de Gumbel), le quantile de même période de retour est donné par :

$$\hat{Q}_{(T)}^s = \hat{\mu}^s - \hat{\sigma}^s \log\left(-\log\left(1 - \frac{1}{T}\right)\right), \quad (2.4)$$

Dès lors, la contrainte que $\hat{Q}_{(T)}^s \leq \hat{Q}_{(T)}^a$ pour tout T implique :

$$\hat{\mu}^s - \hat{\sigma}^s \log\left(-\log\left(1 - \frac{1}{T}\right)\right) + \frac{\hat{\sigma}^a}{\hat{\xi}^a} \left(1 - \left(-\log\left(1 - \frac{1}{T}\right) \right)^{-\hat{\xi}^a} \right) \leq \hat{\mu}^a. \quad (2.5)$$

Cette contrainte peut être imposée pour une période de retour arbitrairement grande, ce qui implique qu'elle est également vérifiée pour toutes les périodes de retour inférieures. Les résultats de l'estimation du modèle final avec la contrainte que la prévision du quantile correspondant à une période de retour de 1000 ans à Angleur doit être supérieure à celle pour le site de Sauhaid sont présentés dans la table 2.6.

Lorsque la contrainte (2.5) est imposée, la valeur estimée du paramètre de courbure pour le site d'Angleur devient plus faible en valeur absolue, impliquant des prévisions plus élevées pour les quantiles extrêmes de la distribution. Notons que l'imposition de cette contrainte ne détériore pas sensiblement la qualité d'ajustement du modèle.

TABLE 2.6 – AGEV, modèle final

$N = 90$	Tabreux	Sauhaid	Angleur
μ	173.96 <i>15.21</i>	323.68 <i>26.04</i>	409.04 <i>36.71</i>
σ	79.66 <i>15.85</i>	136.33 <i>27.09</i>	179.92 <i>33.62</i>
ξ	*	*	-0.12 <i>0.20</i>
Q_{25}	428.76 <i>55.34</i>	759.72 <i>94.61</i>	890.24 <i>118.28</i>
Q_{50}	484.79 <i>66.14</i>	855.61 <i>113.06</i>	974.84 <i>181.14</i>
Q_{100}	540.41 <i>76.95</i>	950.79 <i>131.55</i>	1051.14 <i>275.49</i>
Q_{200}	595.83 <i>87.80</i>	1045.63 <i>150.08</i>	1121.83 <i>421.32</i>
SSR	6037	18075	10841

2.7 Estimations pour la période 1975-2009

Les résultats des estimations indépendantes réalisées pour l'ensemble de la période avec chacune des deux méthodes se trouvent dans la table 2.7. Les cinq observations supplémentaires ont des valeurs relativement modérées par rapport aux maxima observés sur la période qui précède. Si la méthode AGEV est assez robuste face à l'ajout d'une observation supplémentaire (cf. section 1.3.6), l'ajout de cinq observations dans la partie centrale de la répartition d'un échantillon de 30 données affecte substantiellement les résultats. On note d'abord que pour chacun des trois sites et pour les deux méthodes d'estimation utilisées, les valeurs des paramètres de position (μ) et d'échelle (σ) ont diminué sensiblement. Parallèlement la valeur du paramètre de courbure (ξ) a sensiblement diminué en valeur absolue pour l'ensemble des sites, devenant pratiquement égale à zéro pour les sites de Tabreux et Sauhaid. Les estimations ont dès lors été refaites en imposant une distribution de Gumbel pour ces deux sites (table 2.7)⁴.

Le comportement des deux méthodes dans le cas de la distribution Gumbel a fait l'objet d'une analyse par simulations de Monte Carlo au chapitre 1. Il ressortait entre autre de cette analyse que les estimateurs des paramètres et des quantiles extrêmes obtenus par la méthode AGEV sont moins biaisés que ceux de la méthode classique.

4. L'hypothèse d'une distribution Gumbel pour ces deux sites n'est pas rejetée, que ce soit par le t-test ou par le test du rapport de vraisemblance.

TABLE 2.7 – Résultats des estimations, période 1975-2009.

$N = 35$	Tabreux		Sauhaid		Angleur	
	Clas.	Alter.	Clas.	Alter.	Clas.	Alter.
μ	164.17 <i>12.89</i>	162.87 <i>13.80</i>	304.70 <i>21.94</i>	302.34 <i>23.83</i>	392.76 <i>27.81</i>	389.84 <i>30.70</i>
σ	69.22 <i>10.41</i>	79.34 <i>14.71</i>	120.16 <i>17.55</i>	136.97 <i>25.37</i>	150.09 <i>24.31</i>	166.95 <i>26.54</i>
ξ	*	*	*	*	-.09 <i>.27</i>	-.10 <i>.16</i>
$\sum \epsilon$	145.88	.00	235.18	.00	157.02	.00
Q_{25}	385.57 <i>109.51</i>	416.64 <i>51.13</i>	689.04 <i>181.76</i>	740.44 <i>88.21</i>	804.38 <i>147.41</i>	844.80 <i>100.16</i>
Q_{50}	434.26 <i>166.99</i>	472.45 <i>61.17</i>	773.56 <i>278.46</i>	836.78 <i>105.52</i>	878.74 <i>211.12</i>	926.29 <i>145.15</i>
Q_{100}	482.59 <i>239.99</i>	527.85 <i>71.22</i>	857.46 <i>399.79</i>	932.42 <i>122.85</i>	947.61 <i>283.33</i>	1001.55 <i>204.28</i>
Q_{200}	530.74 <i>330.22</i>	583.04 <i>81.30</i>	941.05 <i>547.57</i>	1027.70 <i>140.21</i>	1011.67 <i>362.54</i>	1071.35 <i>279.27</i>
SSR	9891	6830	29308	21378	23460	13416

En outre, les vraies valeurs des quantiles se trouvent systématiquement dans l'intervalle formé par les prévisions de chacune des deux méthodes. Plus précisément, si l'on fait varier la probabilité d'excès en une période $P(x \geq x_0)$ de 0.10 à 0.005 pour différentes tailles d'échantillon, on trouve qu'une pondération $0.7 \times \text{AGEV} + 0.3 \times \text{GEV}$ élimine le biais. Appliquant cette règle pour les prévisions des quantiles élevés des sites de Tabreux et Sauhaid donne :

	Tabreux	Sauhaid
Q_{25}	407	725
Q_{50}	461	818
Q_{10}	514	910

Les résultats des estimations du modèle final (cf. section 2.6, supra) réalisées sur l'ensemble de la période sont présentés dans la table 2.8. La contrainte (2.5) est également liante lorsque l'on ajoute les nouvelles données.

On remarque que l'estimation simultanée conduit à des valeurs légèrement plus faibles des débits correspondant aux différentes périodes de retour (table 2.8 vs. table 2.7). On se retrouve ainsi fort proche de celles obtenues en prenant une moyenne pondérée des deux méthodes, ce qui (en principe) supprime les biais. Cela pourrait suggérer que l'estimation simultanée a également

pour effet de réduire ces biais, mais il serait sûrement prétentieux de généraliser ce résultat à partir d'une seule expérience.

TABLE 2.8 – AGEV, modèle final période 1975-2009

$N = 90$	Tabreux	Sauhaid	Angleur
μ	164.58 <i>13.23</i>	306.58 <i>22.36</i>	392.27 <i>30.03</i>
σ	75.99 <i>14.10</i>	128.48 <i>23.78</i>	161.06 <i>28.03</i>
ξ	*	*	-0.10 <i>0.18</i>
Q_{25}	407.57 <i>48.97</i>	717.36 <i>82.71</i>	832.59 <i>99.62</i>
Q_{50}	461.01 <i>58.59</i>	807.68 <i>98.93</i>	911.83 <i>150.33</i>
Q_{100}	514.05 <i>68.21</i>	897.34 <i>115.18</i>	985.13 <i>224.44</i>
Q_{200}	566.90 <i>77.86</i>	986.67 <i>131.46</i>	1053.23 <i>335.55</i>
SSR	7345	24669	14594

2.8 Quantiles extrêmes

En guise de synthèse, nous comparons à la table 2.9 nos estimations des quantiles extrêmes avec celles réalisées par le SETHY. Ces dernières figurent à la ligne I de chaque site. La ligne II donne les prévisions qui découlent de l'AGEV pour la même période, la ligne III donnant celles obtenues quand la période inclut les observations 2005-9. L'excès durant une année hydrologique du montant figurant dans la colonne Q_{100} (probabilité de 1%) est donné à titre purement indicatif; il est communément admis qu'estimer une période de retour de plus du double de la taille de l'échantillon est hasardeux sinon présomptueux. Les nombres entre parenthèses sont la borne supérieure de l'intervalle de confiance à 95%.

Il convient en premier lieu de remarquer que, non sans raison, les statisticiens du SETHY ont écarté la fonction GEV (classique) au profit de la Gumbel pour Tabreux et Sauhaid et ont choisi la loi Gamma pour Angleur. En général, les services d'études hydrologiques estiment diverses fonctions (lognormale, exponentielle, gamma et GEV; aux USA, la Pearson III a été longtemps la spécification recommandée) qu'ils discriminent sur base du critère d'information d'Akaike (AIC) ou de Schwarz (BIC). Les prévisions issues de

la GEV classique donnent des valeurs encore plus faibles que celles figurant dans la table 2.9, particulièrement pour Angleur ($Q_{25} = 814$). Rappelons que d'après les simulations du chapitre précédent, l'estimation de la GEV classique avec $\xi = -.2$ et $N = 30$ comporte un biais négatif d'environ 10%. L'estimation classique des distributions Gumbel et Gamma donnent un biais plus petit, mais également négatif. Rappelons aussi, en ce qui concerne Angleur que la Gamma est une fonction à deux paramètres, la GEV est donc plus flexible.

L'ajout de cinq observations modifie à la baisse les prévisions de ces quantiles. Selon notre analyse, les estimations du SETHY pour déterminer les zones inondables en 2005 comportaient une sous-estimation des débits extrêmes. Les observations supplémentaires semblent donc avoir l'heureux effet de gommer les écarts. Mais cela souligne aussi, quand on travaille avec de petits échantillons, la nécessité de revoir régulièrement pareilles prévisions. Certes, l'établissement d'une carte des zones inondables constitue un travail considérable, mais il semble bien qu'une révision quinquennale soit un minimum.

TABLE 2.9 – Prévisions des quantiles extrêmes

	Q_{25}		Q_{50}		Q_{100}	
Tabreux						
I	407	(489)	451	(548)	494	(606)
II	429	(541)	485	(619)	540	(697)
III	408	(508)	461	(581)	514	(653)
Sauhaid						
I	726	(866)	803	(968)	876	(1067)
II	760	(952)	856	(1086)	951	(1219)
III	717	(886)	808	(1010)	897	(1131)
Angleur						
I	867	(1025)	953	(1139)	1035	(1249)
II	890	(1130)	975	(1343)	1051	(1612)
III	833	(1037)	912	(1218)	985	(1442)

2.9 Conclusion

Dans ce chapitre, nous avons modélisé les crues annuelles maximales observées à trois endroits différents du bassin de l'Ourthe. Notre analyse s'est basée sur la distribution Généralisée à Valeurs Extrêmes qui possède des fondements théoriques solides pour ce genre d'application. Les performances empiriques de la méthode d'estimation alternative (AGEV) de cette distribution ont été comparées avec celles de la méthode d'estimation classique.

Quelle que soit la méthode, les résultats donnent des prédictions des quantiles extrêmes qui sont contre-intuitives. En effet, à partir d'une période de retour, certes élevée (500 ans avec l'AGEV), les valeurs prédites pour le site de Sauhaid dépassent celles prédites pour le site d'Angleur plus en aval de la rivière. Nous montrons qu'il est très facile avec la méthode d'estimation alternative d'imposer une contrainte sur les valeurs des paramètres des distributions concernées qui garantit que les prévisions vérifient bien la logique physique.

Dans le chapitre 1, nous avons montré que, dans le cas d'une distribution Gumbel, utiliser une moyenne pondérée des deux méthodes dans l'estimation des quantiles élevés aboutit à une quasi suppression des biais. Dans cette application, la distribution Gumbel est celle qui caractérise le mieux les débits maxima de deux des sites et nous avons illustré l'intérêt de cette procédure.

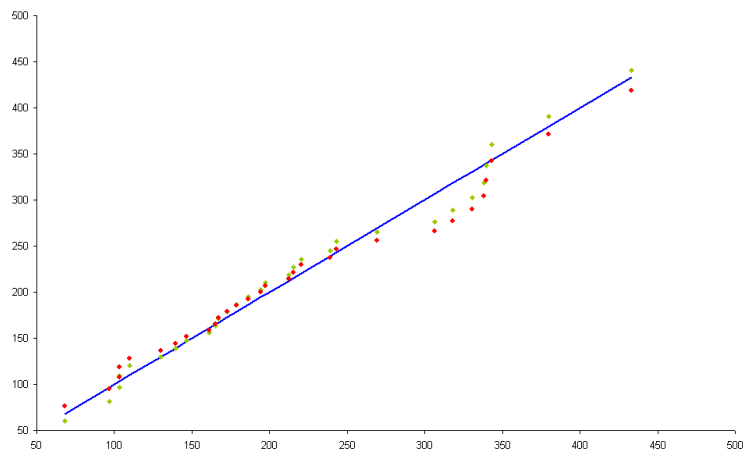
Bien sûr, l'existence de biais en petit échantillon dans l'estimation de la fonction Gumbel par maximum de vraisemblance est connue depuis longtemps et diverses procédures ont été proposées pour pallier ce problème. À notre connaissance, la plus populaire est la correction de Fiorentino et Gabriele (1984). Toutefois, outre le caractère alambiqué de leur formule qui consiste à modifier les valeurs estimées de μ et σ , la correction apportée est largement insuffisante en très petit échantillon (de l'ordre d'une trentaine d'observations).

À titre de remarque finale, on observe un nombre croissant d'études ajoutant une tendance linéaire dans le modèle. L'hypothèse que le paramètre μ augmente d'un même montant à chaque période n'est probablement pas la meilleure manière d'appréhender un éventuel effet de l'évolution climatique. À cet égard, comme mentionné au premier chapitre, l'AGEV permet très facilement d'estimer l'influence du temps sous différentes formes, en particulier des processus d'ajustement dynamique. Cependant aucune des spécifications que nous avons estimées ne révèle un phénomène de ce type dans le cas des crues de l'Ourthe.

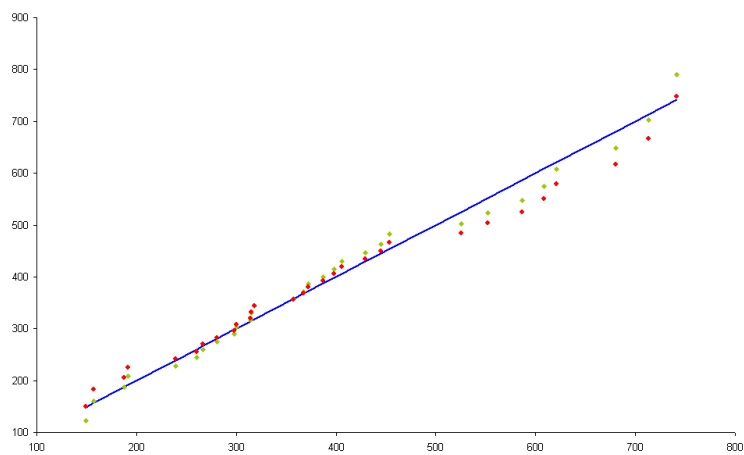
2.A Q-Q plots

Période 1975-2004

Tabreux⁵

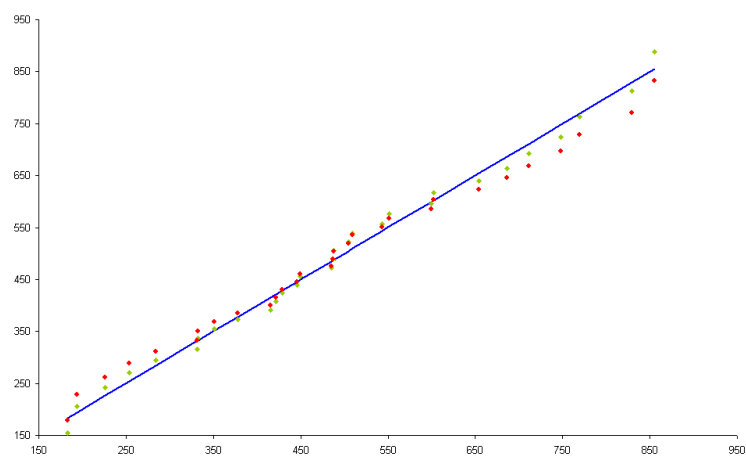


Sauhaid



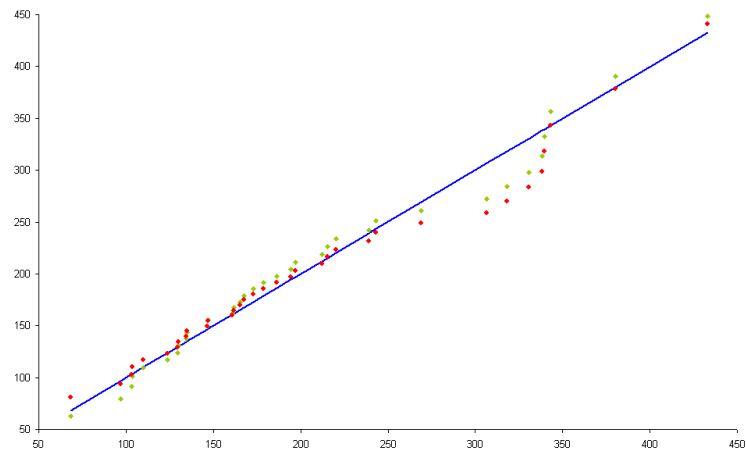
5. losanges verts : méthode alternative, points rouges : méthode classique.

Angleur

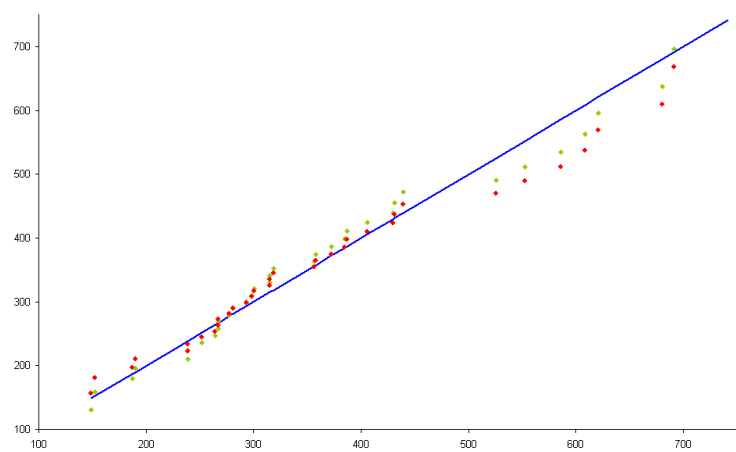


Période 1975-2009

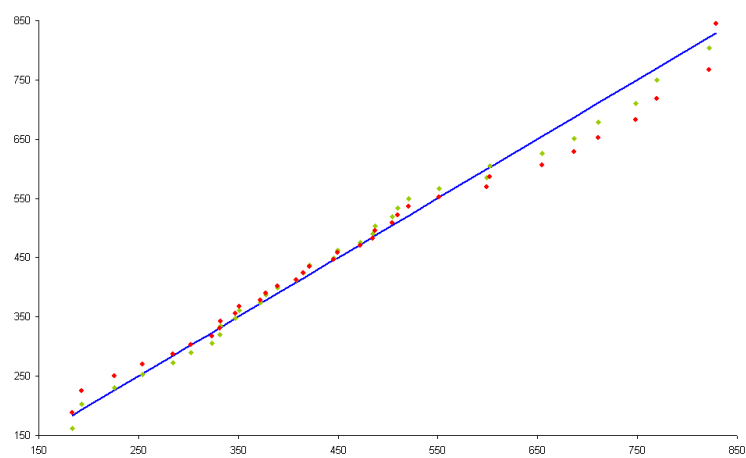
Tabreux



Sauhaïd



Angleur



Chapitre 3

Application aux prix hédoniques : les grands vins

3.1 Introduction

Ce chapitre ainsi que le suivant sont centrés sur la problématique des prix hédoniques. Se basant sur la théorie de la demande développée par Lancaster (1971) selon laquelle les consommateurs demandent un produit avant tout pour les caractéristiques qu'il contient, Rosen (1974) proposa de régresser le prix des produits sur les différentes caractéristiques pertinentes afin d'en obtenir les prix implicites.

Les régressions de prix hédoniques sont régulièrement appliquées au prix des biens durables. On trouve en effet dans la littérature empirique des études sur le marché des véhicules automobiles (Atkinson et Halvorsen, 1984 ; Griliches, 1961), des ordinateurs personnels (Baker, 1997 ; Triplett, 1989) et sur le marché immobilier (Beron *et al.*, 1997 ; Can et Megbolugbe, 1997 ; Mason et Quigley, 1996). On trouve également dans la littérature économique des applications à des produits de consommation courante tels que les céréales (Stanley et Tschirhart, 1991), les oeufs (Gorman, 1980) et, plus récemment, le vin (Combris *et al.*, 1997 ; Di Vittorio et Ginsburgh, 1996 ; Golan et Shalit, 1993 ; Landon et Smith, 1997 ; Nerlove, 1995 ; Oczkowski, 1994 ; 2001 ; Schamel, 2000 ; Schamel et Anderson, 2003).

Toutes ces études empiriques proposent l'estimation d'un modèle économétrique de régression classique par la méthode des moindres carrés ordinaires. Dans certaines de ces études, des problèmes de spécification du modèle sont mis en évidence par les tests usuels de mauvaise spécification (voir par exemple Landon et Smith, 1997 et Oczkowski, 1994). Cependant, les réponses apportées par les auteurs sont tantôt inexistantes, tantôt peu

satisfaisantes, ceux-ci se contentant par exemple d'utiliser des écart-types robustes à la présence d'hétéroscédasticité dans les résidus.

Dans ce chapitre, nous estimons une régression hédonique des prix pour les grands vins de la région de Bordeaux en utilisant une distribution GEV qui, étant donné la nature de la variable dépendante, semble être un outil approprié dans ce genre d'applications. En effet, La modélisation du prix des grands vins par une distribution GEV améliore sensiblement l'ajustement global de la régression hédonique, en particulier dans les queues de la distribution. Cette propriété est appréciable du point de vue des consommateurs étant donné que pour bon nombre d'entre eux, c'est pour les quantiles les plus élevés (ou les plus faibles) que l'inférence est particulièrement intéressante. En outre, nous testons le mérite empirique de la méthode d'estimation alternative des paramètres de la GEV (AGEV) possédant plusieurs propriétés intéressantes qui ont été explicitées dans le premier chapitre. Cette méthode offre un léger gain supplémentaire en précision par rapport à la méthode classique.

La suite du papier est structurée de la manière suivante. Les données sont présentées dans la section 3.2. Dans la section 3.3, nous présentons les résultats de l'estimation d'un modèle de régression hédonique classique. Les résultats des estimations de la distribution GEV par la méthode traditionnelle ainsi que par la méthode alternative sont présentés dans la section 3.4.

3.2 Les données

Les vins constituant l'échantillon sont tous issus de la région de Bordeaux. L'échantillon contient 215 observations. Les données ont été récoltées durant le printemps 2004. La sélection s'est faite en fonction de la disponibilité des informations jugées pertinentes pour l'analyse. Les variables retenues sont les suivantes.

- *Le Prix (P)* : prix en euros observé dans l'un des deux plus grands supermarchés belges¹. Le prix le plus faible dans l'échantillon se situe à 10 euros, le prix médian est de 30 euros, le prix moyen de 50 euros et le vin le plus cher s'établit à un prix de 300 euros. La distribution des prix affiche un caractère hautement asymétrique (figure 3.1).
- *L'origine géographique* : l'échantillon contient des vins originaires des 8 entités suivantes, appelées “communes” : Margaux (Ma), Pauillac (Pa), Saint-Julien (Ju), Saint-Estèphe (Es), Haut-Médoc (Hm), Saint-Emilion

1. Les vins dans ces deux supermarchés sont généralement différents. Lorsque le même vin est proposé simultanément par les deux supermarchés, les différences sont faibles et le prix considéré est le maximum des deux. Toute forme de promotion est ignorée.

(*Em*), Pomerol (*Po*) et Graves (*Gr*).

- *Le classement officiel* : il y a deux grands classements officiels dans la région, le classement des *Grands Crus du Haut-Médoc* et celui des *Grands Crus* de Saint-Emilion. Dans la sous-région du Haut-Médoc² composée de cinq communes (Margaux, Pauillac, Saint-Julien, Saint-Estèphe et Haut-Médoc), le classement officiel répartit les vins en cinq catégories. La composition de l'échantillon en vins issus de ces cinq communes est illustrée dans la table 3.1. Dans la commune de Saint-Emilion, les vins se répartissent en trois catégories : Em_0 dénote les vins non-classés (9 observations), Em_1 contient les *Premiers Grands Crus Classés* (31 observations) et Em_2 les *Grands Crus Classés* (15 observations). Il n'y a pas de classement officiel pour les vins issus des communes de Pomerol (42 observations) et de Graves (19 observations).
- *Le millésime (M)* : les vins présents dans l'échantillon sont de millésimes 1997 (6 observations), 1998 (38 observations), 1999 (73 observations), 2000 (41 observations) et 2001 (57 observations).
- Une variable appelée *Grand Millésime (GM)* : cette variable mesure l'effet croisé des millésimes réputés de 1998 et 2000 dans les communes les plus prestigieuses (Margaux, Pauillac, Saint-Emilion et Pomerol).
- *La qualité (Q)* : cote attribuée par R. Parker (2003) sur une échelle de 100 mesurant les qualités sensorielles et gustatives du vin.
- *La réputation du producteur (R)* : classement établi par R. Parker sur base des performances des producteurs durant les 40 dernières années. Les observations sont classées en six catégories. La répartition des observations par catégories allant de la meilleure, appelée R_1 , à la moins bonne, appelée R_6 , est la suivante : 28, 41, 25, 19, 42 et 60.
- *Le nombre de bouteilles produites (B)* : nombre total de bouteilles produites pour le millésime concerné (en milliers).

2. Le terme Haut-Médoc peut prêter à confusion. En effet, il est utilisé alternativement pour désigner la commune et la sous-région bordelaise incluant cinq communes.

FIGURE 3.1 – Fréquence des prix

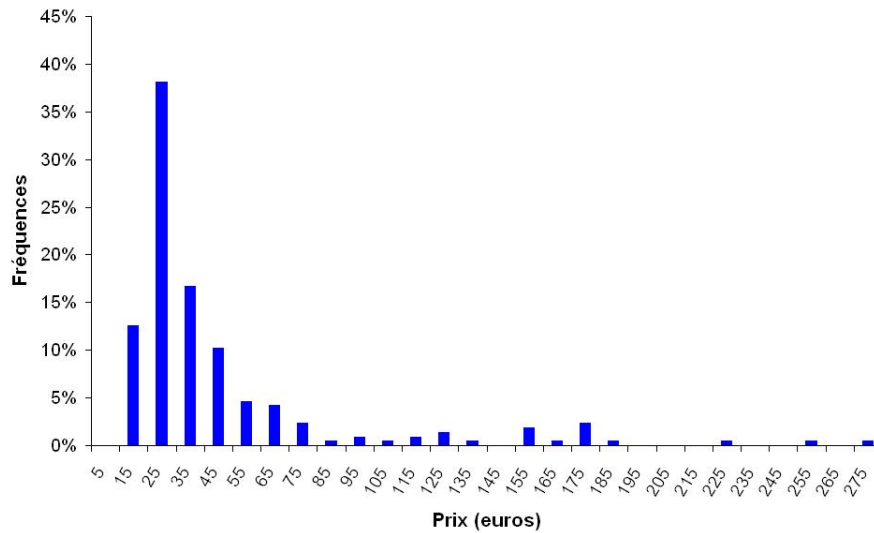


TABLE 3.1 – Grands Crus Classés du Haut-Médoc

	1 ^{er}	2 ^e	3 ^e	4 ^e	5 ^e	n.c.	Total
Margaux (<i>Ma</i>)	3	4	2	5	2	0	16
Pauillac (<i>Pa</i>)	7	5	0	3	18	0	33
St-Estèphe (<i>Es</i>)	0	6	1	4	0	6	17
St-Julien (<i>Ju</i>)	0	15	3	7	0	0	25
Ht-Médoc (<i>Hm</i>)	0	0	1	1	4	2	8

3.3 Approche standard

3.3.1 Le modèle général

Nous partons d'un modèle général qui englobe les deux grandes hypothèses proposées par Landon et Smith (1997).

Une des hypothèses, qui sous-tend leur modèle dit à *information parfaite*, postule que les consommateurs établissent leur choix principalement sur base de la qualité intrinsèque des produits disponibles sur le marché.

L'autre hypothèse, diamétralement opposée, stipule que l'information sur la qualité des biens n'est pas observable ou est coûteuse à obtenir et que par conséquent les consommateurs effectuent leurs choix de consommation sur

base de la réputation des producteurs. A ce propos, une hypothèse largement présente dans la littérature économique est que la qualité passée des produits peut être une bonne approximation de la réputation des producteurs (Shapiro, 1983). Suivant ce constat, Landon et Smith (1997) estiment un modèle dit de *réputation* au sein duquel la qualité passée des produits apparaît comme une variable explicative.

Notre modèle général englobe les deux hypothèses formulées par Landon et Smith (1997). Il prend la forme suivante³ :

$$\begin{aligned} \log P_i = & \alpha + \beta_1 Q_i + \beta_2 Q_{-1,i} + \sum_j \gamma_j R_{j,i} + \sum_j \delta_j M_{a,j,i} + \sum_j \zeta_j P_{a,j,i} + \\ & \sum_j \iota_j E_{s,j,i} + \sum_j \theta_j J_{u,j,i} + \sum_j \kappa_j H_{m,j,i} + \sum_j \lambda_j E_{m,j,i} + \mu P_{o,i} + \\ & \nu Gr_i + \pi_1 \log B_i + \pi_2 (\log B_i)^2 + \sum_k \rho_k M_{k,i} + \phi Gm_i + \epsilon_i, \end{aligned} \quad (3.1)$$

où le terme d'erreur ϵ_i possède les propriétés classiques de bruit blanc.

Le *modèle de réputation* est obtenu lorsque β_1 est imposé égal à zéro dans le modèle (3.1) alors que le *modèle d'information parfaite* découle des restrictions $\beta_2 = \gamma_1 = \dots = \gamma_5 = 0$.

Dans l'étude de Landon et Smith (1997), le *modèle à information parfaite* est statistiquement rejeté par le *modèle de réputation*. En contraste avec ce résultat, sur base de notre échantillon les deux ensembles de restriction ont été statistiquement rejetés par le test standard du rapport de vraisemblance.

3.3.2 Le modèle final

Etant donné la taille relativement modeste de l'échantillon par rapport au nombre de coefficients à estimer dans le modèle général, plusieurs regroupements ont été opérés. Lorsque cela vérifiait une certaine intuition économique, les variables dont la valeur des coefficients semblait fort proche ont été agrégées et les variables dont les coefficients n'étaient pas estimés avec suffisamment de précision ont été enlevées de la régression. Toutes les restrictions imposées aux paramètres du modèle ont été validées par une batterie de tests statistiques. Ainsi, les restrictions ont été testées une par une à l'aide d'un test de rapport de vraisemblance. A la fin de la procédure les restrictions ont à nouveau été testées afin de minimiser les problèmes dus à la non-unicité des séquences des

3. La forme fonctionnelle log-linéaire a été confrontée aux formes linéaires et semi-log. Le modèle log-linéaire passait mieux les tests traditionnels de mauvaise spécification. De plus, ce modèle a rejeté statistiquement les deux autres sur base d'un J-test pour modèles non emboîtés. Toutefois, les cotes attribuées par Parker (Q et Q_{-1}) apparaissent sous forme linéaire. Cette spécification semble plus intuitive étant donné que le système de cotation prend déjà une forme de type logarithmique. Utiliser une transformation logarithmique pour les cotes ne change pas les principaux résultats. Par ailleurs, la variable $\log(B_i)^2$ est ajoutée afin d'obtenir une élasticité variable du prix par rapport au volume de production.

tests d'hypothèses.

Les catégories R_5 et R_6 ont été regroupées étant donné que leurs coefficients n'apparaissaient pas statistiquement distincts. La même opération est réalisée pour les catégories R_3 et R_4 .

Dans la commune de Margaux, les vins classés parmi les catégories 2^e, 3^e, 4^e et 5^e *Grands Crus Classés du Haut Médoc* sont regroupés dans une seule catégorie appelée Ma_{25} . Il en est de même dans la commune de Pauillac où les 2^e, 4^e et 5^e *Grands Crus Classés* sont regroupés dans la catégorie Pa_{25} et dans la commune de Saint-Julien pour laquelle les 2^e, 3^e et 4^e *Grands Crus Classés* sont repris sous une catégorie unique dénommée Ju_{24} .

Dans la commune de Saint-Estèphe, les coefficients associés aux variables 3^e et 4^e *Grands Crus Classés du Haut Médoc* ne sont pas statistiquement significatifs. Ce manque de précision peut être expliqué par le très petit nombre d'observations pour ces catégories (5 au total). Par contre, les vins de la catégorie 2^e *Grands Crus Classés* (Es_2) ainsi que les vins non-classés (Es_6) gardent un coefficient distinct.

Dans la commune du Haut-Médoc, aucune distinction n'est faite entre les vins, le nombre total d'observations étant très limité.

Concernant le Saint-Emilion, deux grandes catégories de vins subsistent : les Saint-Emilion *Premiers Grands Crus Classés* (Em_1) et les autres Saint-Emilion (Em_{02}).

Après cette procédure, le modèle (3.1) devient :

$$\begin{aligned} \log P_i = & \alpha + \beta_1 Q_i + \beta_2 Q_{-1,i} + \gamma_1 R_{1,i} + \gamma_2 R_{2,i} + \gamma_{34} R_{34,i} + \\ & \delta_1 Ma_{1,i} + \delta_{25} Ma_{25,i} + \zeta_1 Pa_{1,i} + \zeta_{25} Pa_{25,i} + \iota_2 Es_{2,i} + \\ & \iota_6 Es_{6,i} + \theta Ju_{24,i} + \lambda_{02} Em_{02,i} + \lambda_1 Em_{1,i} + \mu Po_i + \\ & \nu Gr_i + \pi_1 \log B_i + \pi_2 (\log B_i)^2 + \sum_k \rho_k M_{k,i} + \phi Gm_i + \epsilon_i. \end{aligned} \quad (3.2)$$

3.3.3 Les résultats

Les résultats de l'estimation du modèle (3.2) sont présentés dans la table 3.2, aux 2^e et 3^e colonnes⁴.

Les coefficients associés aux variables binaires tenant compte de la réputation des producteurs sont estimés avec une très grande précision

4. L'origine est un vin de la commune du Haut-Médoc, non-classé par Parker (R_6), de millésime 1999.

statistique et leur valeur respecte le classement établi par R. Parker.

Il existe une différence substantielle et statistiquement très significative de prix entre les vins appartenant à la catégorie des *Premiers Grands Crus Classés* et les autres vins. Cette remarque est valable quelque soit la commune d'origine (Margaux, Pauillac ou Saint-Emilion). La différence de prix est probablement encore plus grande que ce qui est suggéré par la valeur du coefficient étant donné que la majorité de ces *Premiers Grands Crus Classés* appartient également à l'une des catégories R_1 ou R_2 .

Toutes autres choses étant égales par ailleurs, les *Premiers Grands Crus Classés* de la commune de Margaux sont les vins les plus chers de l'échantillon.

Les *Premiers Grands Crus Classés* originaires des communes de Saint-Emilion et de Pauillac bénéficient d'une prime presque aussi importante que ceux issus de la commune de Margaux.

Les Pomerol, les Margaux 2^e à 5^e *Grands Crus Classés* et les Saint-Emilion autres que les *Premiers Grands Crus Classés* bénéficient d'une prime nettement inférieure à celle des *Premiers Grands Crus Classés*.

De manière étonnante, le coefficient associé aux vins non classés originaires de la commune de Saint-Estèphe est positif et significatif. En fait, une des raisons pourrait être que dans cette catégorie sont rangés plusieurs vins relativement onéreux portant l'appellation *Crus Bourgeois* qui est un classement annexe au classement officiel des *Grands Crus Classés du Haut Médoc*.

En moyenne, les vins de la commune de Saint-Estèphe appartenant à la catégorie 2^e *Grand Crus Classés* s'affichent à un prix inférieur aux autres vins possédant des caractéristiques comparables. Toutefois ce résultat est à interpréter avec prudence étant donné le très petit nombre d'observations dans cette catégorie.

L'élasticité du prix par rapport au volume de production est négative et décroît en valeur absolue avec le niveau de cette variable pour atteindre le seuil de 0 à un niveau de production de 400 000 bouteilles, ce qui est loin au dessus de la production maximale pour les vins composant l'échantillon. Théoriquement, cette relation négative entre volume et prix peut se justifier de deux manières. D'une part, les vins les plus réputés sont généralement produits en très faible quantité (c'est particulièrement le cas dans les communes de Saint-Emilion et de Pomerol). D'autre part, on peut supposer que l'industrie viticole présente des rendements d'échelle croissants.

Les prix affichent une évolution linéaire en fonction du millésime, avec une augmentation d'approximativement 5% par an, ce qui peut raisonnablement être expliqué par les coûts de stockage supportés par les producteurs. Par ailleurs, les vins originaires des communes de Margaux, Pauillac, Saint-Emilion et Pomerol qui ont été produits durant les années 1998 et 2000 bénéficient d'une prime substantielle par rapport aux autres vins affichant des caractéristiques comparables.

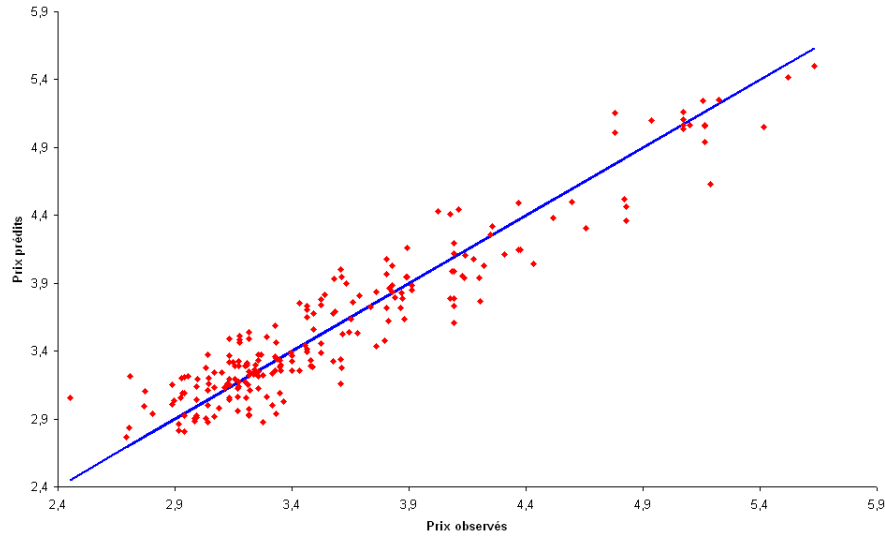
3.3.4 Qualité d'ajustement

En dépit de ces résultats intéressants et très plausibles, plusieurs tests signalent des problèmes de spécification du modèle⁵. On relève en effet que les tests de Ramsey de mauvaise spécification d'ordre 2, 3 et 4 sont tous significatifs.

Afin de visualiser la qualité de l'ajustement, la figure 3.2 présente un graphique mettant en relation les prix observés et les prix prédits (sur une échelle logarithmique) par le modèle (3.2). Il apparaît à la figure 3.2 que le modèle (3.2) affiche des performances médiocres dans les queues de la distribution des prix. En effet, les résidus sont très souvent positifs pour les prix les plus élevés de l'échantillon et systématiquement négatifs pour les prix les plus faibles, ce qui met en évidence les difficultés du modèle (3.2) à rendre compte fidèlement de la nature de la variable dépendante. Cette faiblesse du modèle est particulièrement préjudiciable étant donné que les prédictions pour les vins aux deux extrémités de la distribution des prix sont les plus intéressantes pour bon nombre de consommateurs.

5. Ceux-ci ne sont pas présentés dans la table 3.2 par souci de concision.

FIGURE 3.2 – Prédictions du modèle standard



3.4 Prix hédoniques et distribution GEV

3.4.1 Modèle GEV estimé par la méthode classique

En réponse au problème de spécification identifié dans la section précédente, nous proposons d'utiliser une distribution GEV pour modéliser les prix des grands vins, afin de mieux tenir compte de la nature des observations. En l'occurrence, nous estimons une distribution GEV dans laquelle le paramètre de position (μ) varie en fonction du vecteur des caractéristiques X_i issues du modèle de régression préféré (3.2). Le modèle GEV prend alors la forme :

$$G(\log P_i) = \exp \left\{ - \left[1 + \xi \left(\frac{\log P_i - X_i' \beta}{\sigma} \right) \right]^{-1/\xi} \right\}. \quad (3.3)$$

Toutes les restrictions imposées aux coefficients du modèle général (3.1) ont à nouveau été testées pour vérifier si elles sont toujours valables sous l'hypothèse d'une distribution GEV. Les paramètres du modèle (3.3) sont estimés par la méthode du maximum de vraisemblance. Les résultats sont présentés dans les 4^e et 5^e colonnes de la table 3.2.

Conformément à l'intuition économique, le paramètre de courbure (ξ) prend une valeur négative, impliquant que pour un ensemble de caractéristiques donné le prix du vin ne peut dépasser un certain plafond.

3.4.2 Modèle GEV estimé par la méthode alternative

La méthode AGEV consiste en l'estimation du modèle de régression suivant :

$$\log P_{(i)} = X'_i \beta - \frac{\sigma}{\xi} \left[1 - (-\log(F_{(i)}(\alpha)))^{-\xi} \right] + \epsilon_i, \quad (3.4)$$

dans lequel les observations sont classées par ordre croissant de la valeur de leur résidu découlant de l'estimation du modèle de régression (3.2) et où $F_{(i)}(\alpha)$ représente la fréquence empirique des observations donnée par (1.5). L'estimation des paramètres du modèle (3.4) est réalisée par la méthode du quasi maximum de vraisemblance. Les résultats sont reportés dans les deux dernières colonnes de la table 3.2.

Les différents avantages de la méthode d'estimation alternative de la distribution GEV par rapport à la méthode d'estimation traditionnelle ont été présentés dans le premier chapitre. Dans cette application, l'avantage pratique le plus intéressant est le fait que la condition d'existence (1.2) est automatiquement vérifiée lorsque la méthode AGEV est utilisée, ce qui n'est pas garanti par la méthode classique d'estimation.

Dans le cas présent, étant donné la taille relativement élevée de l'échantillon, le choix du paramètre α dans la fonction (1.5) n'a pas d'influence majeure sur les résultats de l'estimation. Les résultats présentés dans la table 3.2 ont été obtenus en fixant la valeur de ce paramètre égale à zéro.

Le paramètre de courbure (ξ) est négatif et de valeur légèrement inférieure en valeur absolue à la valeur donnée par l'estimation traditionnelle, indiquant un point maximum plus élevé pour la distribution des prix.

3.4.3 Comparaison des résultats

Notons que si les valeurs estimées des coefficients ne sont pas sensiblement différentes dans les trois méthodes, il apparaît que l'ajustement global de la distribution GEV est sensiblement meilleur que celui du modèle de régression classique, avec une légère dominance de la méthode d'estimation alternative.

Les écarts-types des coefficients dans le cas GEV sont également réduits par rapport aux estimations du modèle de régression classique, avec un léger avantage pour la méthode AGEV⁶.

6. Comme expliqué dans le premier chapitre, la présence du terme $F_{(i)}(\alpha)$ dans le modèle (3.4) est un cas typique de ce qui est nommé "régresseur généré" dans la littérature empirique. Bien que les estimateurs des paramètres demeurent consistants, ce n'est plus le cas des estimateurs des écarts-types. Par conséquent, les écarts-types des coefficients ont été estimés par bootstrap (cf. section 1.2.3).

TABLE 3.2 – Résultats des estimations

	OLS		Clas.		Alter.	
	<i>Coef.</i>	<i>Ec.-typ.</i>	<i>Coef.</i>	<i>Ec.-typ.</i>	<i>Coef.</i>	<i>Ec.-typ.</i>
c	3.461	2.110	3.139	2.668	3.199	1.583
Q	.023	.009	.022	.010	.022	.005
Q_{-1}	.021	.006	.023	.007	.022	.003
R_1	1.324	.102	1.325	.094	1.329	.095
R_2	.704	.059	.700	.063	.705	.031
R_{34}	.315	.048	.313	.054	.316	.019
Ma_1	.675	.163	.669	.333	.666	.143
Ma_{25}	.290	.095	.281	.128	.284	.045
Pa_1	.599	.131	.579	.151	.601	.070
Pa_{25}	.108	.083	.097	.112	.102	.070
Es_2	-.317	.134	-.338	.138	-.322	.091
Es_6	.297	.111	.275	.128	.292	.120
Ju_{24}	.101	.079	.100	.111	.101	.074
Em_1	.556	.104	.552	.124	.551	.047
Em_{02}	.242	.092	.235	.120	.237	.053
Po	.358	.096	.354	.120	.357	.081
Gr	.094	.088	.086	.121	.088	.051
$\log B$	-.731	.372	-.699	.466	-.689	.294
$\log B^2$	.029	.016	.027	.021	.027	.013
M_{97}	.146	.095	.157	.253	.151	.052
M_{98}	.109	.065	.109	.070	.106	.036
M_{00}	.047	.075	.049	.079	.039	.029
M_{01}	-.047	.042	-.045	.045	-.047	.022
GM	.175	.070	.183	.073	.188	.035
σ			.203	.012	.204	.011
ξ			-.266	.071	-.252	.046
SSR	9.023		.051		.038	

Par ailleurs, en contraste avec les résultats du modèle de régression classique (3.2), les tests de mauvaise spécification de Ramsey ne sont plus significatifs dans le cas des estimations de la GEV, et ce, quelle que soit la méthode.

Notons que la qualité de l'ajustement de la distribution GEV est particulièrement bonne dans les queues de la distribution. Ceci est illustré dans la table 3.3 où sont reportées les prévisions des différents modèles pour les dix observations présentant les prix les plus élevés de l'échantillon. Le modèle de régression classique (3.2) montre toutes ses limites en sous-estimant quasi systématiquement et parfois largement le prix observé pour ces observations. Quelle que soit la méthode d'estimation utilisée, la distribution GEV témoigne d'une qualité d'ajustement beaucoup plus appréciable pour ces prix extrêmes.

En outre, la méthode AGEV affiche une qualité d'ajustement légèrement meilleure que celle de la méthode classique d'estimation dans les extrémités de la distribution des prix puisque les prévisions du modèle pour les 20 prix les plus élevés de l'échantillon sont plus proches de la valeur observée pour 15 d'entre eux. Ce résultat vaut également pour l'extrémité inférieure de la distribution des prix puisque la méthode AGEV donne des meilleures prévisions pour 16 des 20 vins les moins chers de l'échantillon.

TABLE 3.3 – Prévisions dans les quantiles supérieurs

Observations	OLS	Class.	Alter.
164	150.3	164.9	166.7
173	192.9	170.6	172.1
175	140.8	174.1	176.7
175	159.5	178.4	179.3
175	160.5	178.3	175.9
179	97.8	169.7	171.3
186	182.3	188.4	184.9
225	159.0	222.0	224.4
250	219.8	252.0	254.8
279	240.6	277.6	282.5

3.5 Conclusion

La réputation individuelle des producteurs ainsi que la réputation collective (le classement officiel et la commune d'origine) semblent être les éléments les plus déterminants pour expliquer la variation des prix sur le marché des grands vins de la région de Bordeaux. La qualité, telle que mesurée par la cotation révélée par le guide Parker, a également un impact statistiquement significatif sur les prix, mais quantitativement faible.

Le millésime du vin est un facteur non négligeable. Ainsi, un vin de millésime 2000 issu des communes les plus prestigieuses bénéficie d'une prime d'approximativement 20% par rapport à un vin de 1999 possédant des caractéristiques semblables.

La prime accordée aux vins appartenant à la catégorie des *Premiers Grands Crus Classés* est conséquente et d'ampleur comparable quelle que soit la commune d'origine.

Les vins présentant le meilleur rapport "qualité-prix" ont été identifiés dans la commune de Saint-Estèphe. En effet, les vins de cette commune appartenant à la catégorie des 2^e *Grands Crus Classés du Haut-Médoc* présentent un prix en retrait de près de 30% par rapport aux autres vins possédant des caractéristiques comparables. Toutefois, même si le coefficient est statistiquement significatif, ce résultat est à interpréter avec prudence étant donné qu'il repose sur un nombre très limité d'observations.

En vue d'améliorer l'ajustement par rapport aux estimations par moindres carrés linéaires, le mérite d'une distribution GEV a été testé. Même si les prix implicites des différentes caractéristiques ne sont pas sensiblement différents, une amélioration conséquente de l'ajustement est obtenue lorsque les prix sont modélisés par une GEV.

En outre, nous avons montré dans cette application que la méthode d'estimation alternative (AGEV) offrait un gain supplémentaire en précision par rapport à la méthode classique, en particulier dans les queues de la distribution.

Chapitre 4

Application aux prix hédoniques : les loyers

4.1 Introduction

Le marché locatif constitue un autre cas intéressant de prix hédoniques. La question est ici de savoir quelles sont les caractéristiques qui expliquent les différences de loyers entre les logements. Dans ce chapitre, on construit un indice hédonique des loyers observés entre 1999 et 2005 pour les logements de Louvain-la-Neuve. Dans une première étape, un modèle de régression classique est estimé. La qualité de l'ajustement statistique de ce modèle n'est pas très bonne, en particulier dans les extrémités de la distribution des loyers. Nous montrons que, même si les interprétations économiques sont quasi inchangées par rapport au modèle de régression classique, la qualité d'ajustement d'une distribution GEV est bien meilleure, celle-ci permettant de mieux capturer le comportement de la variable dépendante. En outre, nous comparons les résultats empiriques de la méthode d'estimation alternative de la GEV avec ceux de la méthode classique.

Ce chapitre, tout comme le chapitre précédent, se base sur la théorie des prix hédoniques (Rosen, 1974) selon laquelle le prix d'un bien est fonction de l'ensemble des caractéristiques qu'il contient. Des régressions de prix hédoniques ont été utilisées dans la recherche appliquée pour mesurer différents mécanismes à l'oeuvre sur les marchés immobiliers tels que les externalités environnementales (Clark *et al.*, 1997; Nelson, 1978; Simons *et al.*, 1999), ou la qualité et l'accessibilité des logements (Can, 1990; Dubin, 1988). Cette technique a également été utilisée pour estimer les fonctions de demande pour différentes caractéristiques des logements (Brueckner et Colwell, 1983; Witte *et al.*, 1979).

Les modèles proposés dans la littérature appliquée au marché immobilier sont dans la plupart des cas des modèles de régression classique. Ceux-ci sont utilisés intensivement en économétrie en raison de leur simplicité et des bonnes performances empiriques qu'ils affichent dans la plupart des applications. Néanmoins, ces modèles connaissent deux limites qui prennent toute leur importance dans le cas du marché immobilier.

Premièrement, ils ne tiennent généralement pas suffisamment compte de la dimension spatiale des données.

En effet, d'un côté, ils considèrent les observations comme indépendantes. Pourtant, la présence d'une dépendance spatiale dans ce type de données a été mise en évidence à plusieurs reprises dans la littérature (Dubin, 1988; Can, 1990; Anselin, 1990; 1988). En particulier, Bowen *et al.* (2001) ont montré que si le modèle ne la prend pas en compte, les estimateurs des paramètres deviennent biaisés et inconsistants.

D'un autre côté, les modèles classiques supposent généralement que les paramètres sont constants à travers l'espace ou varient seulement de manière discrète par bloc de logements (par quartier). Cette spécification est probablement trop restrictive dans bon nombre d'applications où la variation est plus locale, tout particulièrement dans les agglomérations où le découpage par quartier n'est pas très fin. Plusieurs techniques statistiques ont été développées pour traiter ce problème d'hétérogénéité spatiale, comme la méthode d'extension spatiale (Casetti, 1972) ou le modèle à coefficients aléatoires (Fotheringham *et al.*, 1996). Plus récemment, la régression spatialement pondérée (*Geographically Weighted Regression*) s'est répandue dans la littérature appliquée (Brunsdon *et al.*, 1996; Fotheringham *et al.*, 1998). Cette méthode statistique permet à la valeur des différents paramètres d'intérêt de varier de manière continue à travers l'espace.

Deuxièmement, les modèles de régression classiques donnent un mauvais ajustement lorsque les données s'écartent trop largement de leur moyenne. Cela se traduit souvent dans les études empiriques par des résultats significatifs des tests traditionnels de mauvaise spécification.

L'utilisation d'une distribution de Weibull à trois paramètres (GEV) pour modéliser les loyers permet de mieux tenir compte, de manière significative, du comportement de la variable dans les queues de la distribution.

En outre, notre approche est relativement à l'abri des critiques touchant les régressions classiques concernant la prise en compte de la dimension spatiale des observations. D'une part, on connaît l'emplacement exact de chaque logement et, dès lors, on peut inclure une variable continue indiquant

la distance du logement par rapport aux principaux biens publics, ce qui permet de réduire la corrélation spatiale des résidus. D'autre part, les limites géographiques de la ville de Louvain-la-Neuve sont bien nettes et celle-ci est découpée en cinq quartiers relativement homogènes. Ceci permet, sans risque d'ambiguïté, d'estimer un modèle avec un ensemble distinct de paramètres par quartier.

Les données sont décrites dans la section 4.2. Dans la section 4.3, les résultats de l'estimation du modèle de régression classique sont présentés et discutés. Les résultats des estimations de la distribution GEV réalisées à la fois par la méthode traditionnelle et par la méthode alternative sont reportés dans la section 4.4.

4.2 Les données

Nous avons obtenu, auprès du service d'information de Louvain-la-Neuve (INFORVILLE), toutes les annonces de biens mis en location entre 1999 et 2005¹. Il y a 781 observations au total, réparties entre les cinq quartiers de la ville, à savoir les Bruyères (346), Lauzelle (198), le Centre (35), l'Hocaille (156) et le Biéreau (46). Les informations disponibles sur chaque logement sont les suivantes.

- *Le Prix (P)* : loyer mensuel réel en euros, hors charges. Les loyers nominaux ont été déflatés par l'indice général des prix à la consommation.
- *L'année de location (Y_{ii})* : variable muette prenant la valeur 1 si l'annonce a été placée durant l'année ii et 0 sinon.
- *La distance par rapport au lac (Lac)* : distance à vol d'oiseau du logement par rapport au centre du lac de Louvain-La-Neuve (en km).
- *La distance par rapport à la gare ($Gare$)* : distance à vol d'oiseau du logement par rapport à la gare de Louvain-la-Neuve (en km).
- *Les salles de bain et de douche* : variables auxiliaires prenant une valeur entre zéro et trois selon le nombre de salles de bain ($0b$ à $3b$) et de salles de douche ($0d$ à $3d$).
- *Les chambres (C)* : variable auxiliaire prenant une valeur comprise entre un et huit selon le nombre de chambres ($1C$ à $8C$).

1. Cette période a l'avantage d'être relativement homogène. En effet, contrairement à la période plus récente, on n'a pas observé de grands chantiers immobiliers à cette époque.

- *Le balcon (Ba), la terrasse (Ter), le jardin (Ja), la cave (Ca) et le garage (Ga)* : sont des variables muettes prenant la valeur 1 si le bien dispose de la commodité et 0 sinon.
- *Les w.c. indépendants (Wc)* : variable auxiliaire prenant une valeur comprise entre zéro et quatre selon le nombre de w.c. indépendants ($0Wc$ à $4Wc$).
- *La surface (Surf)* : superficie habitable du logement en mètres carrés.

4.3 Approche classique

4.3.1 Le modèle

Nous sommes partis d'un modèle général où chaque variable a un coefficient spécifique par quartier. Pour permettre un maximum de flexibilité, le nombre de chambres, de salles de bain, de douche et de toilettes indépendantes apparaissent dans le modèle sous la forme de variables muettes exclusives. Le modèle de départ permet également une élasticité non constante du loyer par rapport à la surface habitable.

Etant donné le très faible nombre d'observations dans certaines catégories, plusieurs simplifications ont été directement opérées avant de procéder à l'estimation du modèle. Ainsi, une variable muette regroupant toutes les observations contenant plus de quatre chambres a été créée pour chaque quartier (5^+R), et il en est de même en ce qui concerne les biens contenant plus de deux toilettes indépendantes (3^+Wc). Pour les salles d'eau, cinq catégories distinctes ont été constituées pour chaque quartier : les biens disposant uniquement d'une salle de douche ($0b1d$), uniquement d'une salle de bain ($1b0d$), d'une salle de bain et d'au moins une salle de douche supplémentaire ($1b1^+d$), de deux salles de bain sans salle de douche ($2b0d$) et de deux salles de bain avec au moins une salle de douche supplémentaire ($2b1^+d$).

L'analyse des éléments diagonaux de la matrice "hat" et des "résidus standardisés" (Belsley *et al.*, 1980) du modèle de départ a permis de découvrir cinq observations aberrantes dans l'échantillon. A défaut de pouvoir les corriger de manière satisfaisante, celles-ci ont été éliminées.

Le modèle général prend la forme suivante² :

$$\begin{aligned} \log P_i = & \Sigma_k \alpha_k Y_{99+k,i} + \Sigma_j \beta_{1,j} Lake_i + \Sigma_j \beta_{2,j} Sta_i + \Sigma_j \beta_{3,j} 1b0s_i + \\ & \Sigma_j \beta_{4,j} 1b1^+ s_i + \Sigma_j \beta_{5,j} 2b0s_i + \Sigma_j \beta_{6,j} 2b1^+ s_i + \Sigma_j \Sigma_k \gamma_{k,j} R_{k,i} + \\ & \Sigma_j \beta_{7,j} Ba_i + \Sigma_j \beta_{8,j} Bu_i + \Sigma_j \beta_{9,j} Ter_i + \Sigma_j \Sigma_k \rho_{k,j} WC_{k,i} + \\ & \Sigma_j \beta_{10} Ca_i + \Sigma_j \beta_{11,j} \log(surf_i) + \Sigma_j \beta_{12,j} \log(surf_i)^2 + \epsilon_i, \end{aligned} \quad (4.1)$$

où l'index (j) pour les cinq différents quartiers est : 1 = Bruyère, 2 = Lauzelle, 3 = Centre, 4 = Hocaille et 5 = Biéreau. Les variables Y_{99+k} (pour $k = 1, 6$) constituent un indice hédonique des loyers entre 1999 et 2005.

Nous avons suivi une approche du général au spécifique afin de dégager le modèle le plus simple pour lequel la perte en vraisemblance n'était pas statistiquement rejetée. Plus précisément, lorsque cela vérifiait une certaine intuition économique, les variables pour lesquelles le coefficient n'était pas statistiquement significatif ont été éliminées et les variables dont les valeurs du coefficient n'étaient pas statistiquement différentes ont été regroupées. Après chaque simplification un test traditionnel du rapport de la vraisemblance est effectué. Les restrictions sont à nouveau testées ensemble à la fin de la procédure pour éviter les problèmes liés à la non-unicité de la séquence des tests d'hypothèse. Les résultats de l'estimation du modèle final sont présentés dans la table 4.1, aux 2^e et 3^e colonnes.

4.3.2 Les résultats

Le modèle final rejette d'autres spécifications plus simples telles qu'un modèle standard avec des coefficients uniques pour l'ensemble des quartiers ainsi qu'un modèle à effet fixe. Ce résultat tend à indiquer la présence de sous-marchés immobiliers au sein de l'espace urbain même dans une ville aussi peu étendue que Louvain-la-Neuve.

Le lac est un bien public portant une externalité positive dans les quartiers des Bruyères, de Lauzelle et de l'Hocaille. Cet effet, renforcé dans le quartier de l'Hocaille, peut probablement s'expliquer en partie par la présence de nombreux logements disposant d'une vue directe sur le lac, ce qui est moins le cas dans les autres quartiers. Au Biéreau et dans le Centre, l'influence du lac n'est pas mesurée précisément. Par contre dans ces quartiers qui sont les plus proches de la gare ferroviaire, la proximité des logements par rapport à celle-ci semble avoir un effet négatif sur les loyers. A l'inverse, la proximité de la gare exerce un effet positif sur le prix dans les quartiers de Lauzelle et de l'Hocaille. Pour ces quartiers, la présence d'un espace vert sur le chemin qui

2. Différentes spécifications de la forme fonctionnelle liant le prix à la surface habitable ont été testées (linéaire, log-linéaire et semi-logarithmique). La forme log-linéaire avec effet quadratique a été retenue sur base des résultats des tests usuels de mauvaise spécification.

mène à la gare explique probablement en partie ce résultat. Notons toutefois que ces effets de distance sont quantitativement très faibles.

Dans les quartiers des Bruyères et de l'Hocaille, la présence d'une salle de bain au lieu d'une salle de douche apporte une prime de 6% au loyer, tandis que l'effet n'est pas significatif dans les autres quartiers. Le loyer est plus élevé lorsque le bien dispose à la fois d'une salle de bain et d'une salle de douche que lorsque celui-ci ne dispose que d'une salle de douche (+7%) ou que d'une salle de bain (+2%). Ceci est vrai dans tous les quartiers excepté à Lauzelle, où l'effet négatif est peu évident à interpréter. Les logements disposant de deux salles de bain sont plus chers que ceux proposant une salle de bain et une salle de douche dans le quartier des Bruyères (+5%) tandis qu'il n'y a pas de différence de prix pour les quartiers de Lauzelle et de l'Hocaille.

Dans tous les quartiers, il existe une relation positive entre le nombre de chambres et le prix mais l'effet d'une chambre supplémentaire est quantitativement plus faible dans les quartiers des Bruyères et de Lauzelle.

La présence d'un balcon apporte une prime statistiquement significative dans les quartiers des Bruyères, du Centre et du Biéreau, avec une augmentation des loyers de 4% en moyenne. Dans les quartiers de l'Hocaille et de Lauzelle l'effet n'est pas significatif.

La prime associée au bureau est comparable dans les quartiers des Bruyères et de l'Hocaille (+5%) tandis qu'elle est significativement plus importante au Biéreau (+12%).

La présence d'une terrasse engendre une augmentation du loyer de 3% aux Bruyères, à l'Hocaille et à Lauzelle, et de 13% au Biéreau. Dans le quartier du Centre le coefficient associé à la terrasse est statistiquement significatif et négatif. On peut trouver différentes explications à ce résultat, par exemple le fait que la terrasse soit une faible compensation de la mauvaise localisation, mais étant donné le faible nombre d'observations toute interprétation est arbitraire.

Les biens disposant d'une cave se louent plus chers dans tous les quartiers hormis aux Bruyères et dans le Centre où l'impact sur le prix n'est pas significatif.

L'effet d'une toilette indépendante supplémentaire est toujours statistiquement significatif et positif dans les quartiers de Lauzelle, de l'Hocaille et des Bruyères. Dans les autres quartiers l'effet n'apparaît pas significatif.

La présence d'un garage apporte une prime de 5% au logement dans les quartiers des Bruyères de l'Hocaille et de Lauzelle. Le coefficient négatif et significatif associé au garage dans le quartier du Centre est peu intuitif. Notons qu'il n'y a que trois observations disposant d'un garage dans ce quartier.

L'effet de la surface habitable sur le loyer est différent en fonction du quartier. Aux Bruyères et à Lauzelle, l'effet est identique avec une élasticité positive et décroissante. Ainsi, l'élasticité est de 0.3 pour une surface habitable de 50 mètres carrés, de 0.2 pour une surface de 120 mètres carrés et atteint une valeur de 0 pour une surface de 500 mètres carrés, soit largement au-delà de la valeur maximale dans l'échantillon. Pour les autres quartiers, par contre, l'élasticité est constante, prenant une valeur de 0.08 au Biéreau, de 0.23 à l'Hocaille et de 0.38 au Centre.

L'indice des prix hédoniques estimés dans le modèle final indique que lorsque l'on contrôle pour l'ensemble des caractéristiques disponibles, les loyers ont augmenté en termes réel d'à peu près 16% au total sur la période 1999-2005, ce qui représente une croissance annuelle d'à peu près 2.5%.

4.4 Distribution GEV

Une particularité de l'échantillon étudié est qu'il possède quelques biens affichant des loyers très élevés. Nous proposons dans cette section d'utiliser une distribution GEV pour modéliser cette variable, ce qui permet de mieux tenir compte de son caractère "extrême".

La distribution GEV (1.1) est utilisée pour modéliser les loyers, exprimés en logarithme, en permettant au paramètre de position (μ) de varier en fonction du vecteur des caractéristiques pertinentes identifiées dans la section précédente. Pour tenir compte de l'hétérogénéité des observations inhérente au quartier, nous permettons au paramètre d'échelle (σ) de prendre une valeur distincte par quartier. Les résultats de l'estimation de la GEV réalisée par la méthode classique sont reportés dans les 4^e et 5^e colonnes de la table 4.1, tandis que ceux de la méthode alternative apparaissent dans les 6^e et 7^e colonnes de cette même table.

D'après les résultats des deux méthodes, la valeur estimée du paramètre de courbure (ξ) de la GEV est négative, ce qui tend à indiquer que les loyers suivent une loi appartenant à la famille de distributions de Weibull. L'égalité du paramètre de dispersion à travers les différents quartiers n'a pas été rejetée par le test traditionnel du rapport de vraisemblance.

Les valeurs estimées des coefficients dans le modèle GEV sont tout à fait comparables à celles obtenues dans l'estimation du modèle de régression clas-

sique (4.1). L'interprétation économique des résultats réalisée dans la section précédente reste par conséquent valable sous l'hypothèse d'une distribution GEV.

Cependant, l'ajustement global de la distribution GEV est nettement meilleur que celui du modèle de régression classique, avec une somme des carrés des résidus et des écarts-types des coefficients beaucoup plus faibles.

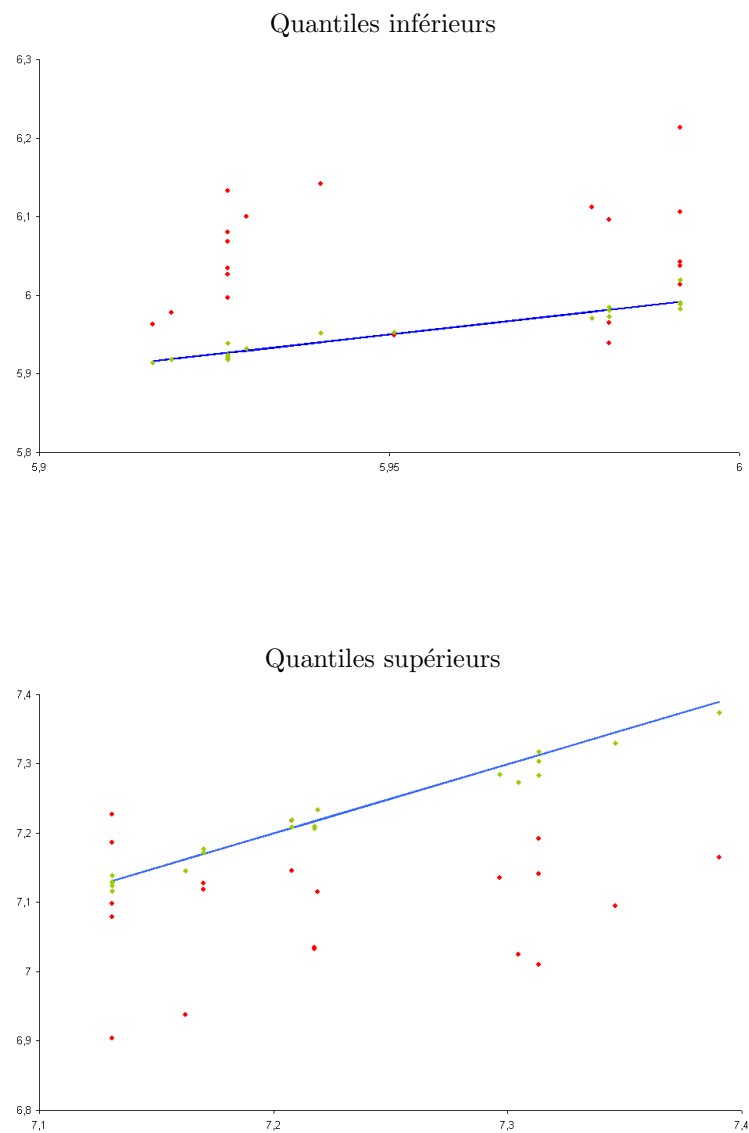
La figure 4.1 illustre les prédictions dans les queues de la distribution des loyers effectuées sur base des résultats du modèle de régression classique (ronds rouges) ainsi que des résultats de l'estimation de la GEV réalisée par la méthode classique (losanges verts) qui ne sont pas sensiblement différents de ceux obtenus par la méthode AGEV. L'exercice est effectué pour les 20 loyers les moins élevés et les 20 loyers les plus élevés de notre échantillon. La figure 4.1 illustre bien les limites du modèle de régression classique qui, systématiquement, sous-estime les loyers les plus élevés et sur-estime les loyers les moins élevés. La distribution GEV donne des prévisions beaucoup plus proches des observations pour ces loyers extrêmes.

Les tests de mauvaise spécification ne figurent pas dans la table 4.1 par souci de concision. Les tests de Ramsey d'ordre 2, 3 et 4 ne sont significatifs pour aucun des modèles estimés. Le test d'hétéroscédasticité de Breusch et Pagan est significatif sauf dans le cas de l'AGEV. En outre, l'hypothèse de normalité des résidus est largement rejetée avec chacune des méthodes.

TABLE 4.1 – Résultats des estimations

	OLS		Clas.		Alter.	
	<i>Coef.</i>	<i>Ec.-typ.</i>	<i>Coef.</i>	<i>Ec.-typ.</i>	<i>Coef.</i>	<i>Ec.-typ.</i>
Y_{00}	.03	.02	.04	.002	.04	.002
Y_{01}	.03	.02	.04	.002	.03	.002
Y_{02}	.09	.02	.08	.002	.09	.002
Y_{03}	.10	.02	.11	.002	.10	.002
Y_{04}	.14	.02	.14	.002	.14	.002
Y_{05}	.15	.02	.15	.002	.15	.002
$Lake_{12}$	-.27	.11	-.23	.010	-.25	.011
$Lake_{45}$	-.89	.20	-.90	.019	-.88	.021
Sta_1	.49	.12	.49	.011	.49	.012
Sta_{35}	-.48	.14	-.44	.013	-.47	.013
Sta_{24}	.52	.79	.29	.081	.26	.081
$1b0s_{14}$	.05	.01	.06	.001	.06	.001
$1b1^+s_{1345}$	.07	.02	.07	.002	.07	.002
$2b0s_1$	.11	.04	.13	.004	.12	.004
$2b0s_{24}$	.06	.03	.06	.003	.06	.003
$2b1^+s_{15}$	.13	.04	.13	.004	.14	.004
$2R_1$	.11	.01	.11	.001	.11	.001
$2R_{23}$	.07	.02	.07	.002	.07	.002
$2R_{45}$	.18	.02	.18	.002	.18	.002
$3R_{123}$	.21	.02	.21	.002	.21	.002
$3R_4$	.30	.03	.30	.002	.30	.003
$3R_5$	.46	.04	.46	.003	.46	.004
$4R_1$	.28	.04	.28	.004	.29	.004
$4R_2$	.36	.03	.37	.002	.36	.003
$4R_{45}$	.46	.03	.45	.003	.46	.003
$5R_1$	.36	.04	.36	.004	.36	.004
$5R_{24}$	.45	.03	.44	.003	.44	.003
Ba_{135}	.04	.01	.04	.001	.04	.001
Bu_{14}	.05	.02	.04	.001	.05	.001
Bu_5	.11	.06	.12	.006	.12	.005
Ter_{124}	.03	.01	.03	.001	.03	.001
Ter_3	-.08	.03	-.09	.003	-.08	.003
Ter_5	.13	.03	.14	.003	.13	.003
Ca_{245}	.05	.01	.05	.001	.05	.001
$2Wc_{24}$	.03	.02	.02	.002	.02	.002
$2Wc_1$	.03	.02	.03	.002	.03	.002
$3Wc_{124}$	.05	.03	.06	.002	.05	.003
Ga_{124}	.05	.01	.05	.001	.05	.001
Ga_3	-.10	.06	-.12	.006	-.10	.006
$Surf_{12}$	.87	.16	.94	.015	.88	.015
$Surf_3$	.38	.07	.38	.007	.38	.008
$Surf_4$	.23	.02	.23	.002	.23	.002
$Surf_5$	.08	.04	.07	.003	.08	.004
$(Surf_{12})^2$	-.07	.01	-.07	.002	-.07	.002
σ			.08	.004	.08	.004
ξ			-.21	.006	-.21	.005

FIGURE 4.1 – Prévisions des quantiles extrêmes



4.5 Conclusion

Dans ce chapitre, nous avons illustré l'intérêt d'utiliser une distribution GEV-Weibull pour modéliser les loyers observés dans la ville de Louvain-la-Neuve entre 1999 et 2005. Nous avons vu qu'un modèle de régression hédonique classique donnait un ajustement assez mauvais dans les queues de la distribution des loyers. La distribution GEV permet de mieux tenir compte du comportement de la variable dépendante. Elle fournit une qualité d'ajustement nettement meilleure, et ce, que les paramètres soient estimés par la méthode classique ou la méthode alternative. Toutefois, les interprétations économiques découlant des valeurs des coefficients estimés par les différentes méthodes ne varient guère.

Dans les milieux politiques belges, on parle régulièrement de la nécessité d'"objectiver" les loyers des biens immobiliers résidentiels. Les résultats de ce chapitre illustrent certaines difficultés qui incomberaient à une telle entreprise.

D'une part, nous mettons en évidence la présence de sous-marchés immobiliers au sein des espaces urbains, même pour des villes très peu étendues telles que la ville de Louvain-la-Neuve. Cette ville possède des quartiers biens délimités et homogènes, ce qui a permis d'utiliser une spécification simple du modèle. Pour d'autres villes, dans lesquelles le découpage en quartiers homogènes n'est pas aussi évident, l'analyse hédonique des loyers sera probablement plus compliquée.

D'autre part, nous avons montré qu'il était judicieux de considérer les caractéristiques de l'environnement des logements. Par exemple, nous avons vu que la proximité du lac est un facteur explicatif des loyers lorsque l'on contrôle pour les autres caractéristiques. Il est très probable que d'autres caractéristiques de l'environnement aient été omises (au même titre que des caractéristiques propres au logement), comme par exemple la pollution sonore qui peut affecter certains lieux traditionnels de réunions estudiantines.

Nous avons montré qu'en termes de prix hédoniques, le loyer réel moyen demandé pour un logement a augmenté de 16% sur la période 1999-2005, soit une croissance annuelle d'à peu près 2.5%. L'augmentation réelle est encore plus sensible si on utilise comme déflateur l'indice santé, qui sert pourtant de base à l'indexation des loyers. Il ne semble pas que ce résultat puisse être attribué à un effet d'amélioration de la qualité dans le temps inhérent à la vétusté relative des logements : une analyse sur un sous-échantillon pour lequel cette information est disponible montre que l'année de construction n'a pas d'effet sur le loyer.

Chapitre 5

Application aux coûts des sinistres automobiles

5.1 Introduction

Il est crucial pour une compagnie d'assurance automobile de pouvoir anticiper les coûts des sinistres qu'elle couvre. Grâce à différentes techniques statistiques, il est possible d'obtenir une estimation du coût moyen attendu en fonction des diverses caractéristiques observables des individus et de leur véhicule. Le coût attendu pour la compagnie de la couverture d'un individu au cours d'un intervalle de temps déterminé (généralement un an) est une variable aléatoire composée de deux éléments qui peuvent être considérés comme indépendants : d'une part la fréquence des sinistres et, d'autre part, le coût par sinistre (Vanderlaan et Louter, 1986).

La fréquence est le plus souvent modélisée par un processus de Poisson ou par des distributions apparentées (binomiale et binomiale inverse) en permettant à la moyenne de varier en fonction des principaux facteurs de risque identifiés parmi les caractéristiques disponibles telles que le sexe, l'âge et l'état civil de l'individu, la vétusté, la puissance et la couleur du véhicule, ou encore la distance parcourue par le véhicule sur l'intervalle de temps (Denuit *et al.*, 2003b).

La modélisation du coût des sinistres est moins évidente. En effet, il n'y a pas de consensus dans la littérature appliquée sur le type de distribution à utiliser. Ainsi, lorsque l'intérêt du chercheur porte uniquement sur la distribution des coûts au dessus d'un certain seuil, la distribution de Pareto généralisée fait preuve d'un meilleur ajustement que des distributions simples telles que la gamma ou la log-normale (Denuit *et al.*, 2003a ; McNeil, 1997). En outre la Pareto généralisée a des fondements théoriques que les autres

n'ont pas. Par contre, lorsqu'il s'agit de modéliser l'ensemble des coûts de sinistres, le choix de la distribution est beaucoup plus équivoque. Vanderlaan et Louter (1986) comparent cinq distributions différentes, parmi lesquelles la gamma et la log-normale, sans pouvoir les départager clairement. En plus des distributions gamma et log-normale, Hogg et Klugman (1983) proposent d'utiliser une distribution de Pareto ou de Weibull pour modéliser les coûts de sinistres. Haberman et Renshaw (1996) proposent d'utiliser une distribution à queue lourde de type Burr, Pareto ou Weibull. Dans le souci d'améliorer l'ajustement, des modèles combinant des distributions à queue épaisse et à queue fine ont également été proposés, notamment une distribution composée log-normale-Pareto (Cooray et Ananda, 2005).

Par ailleurs, utiliser uniquement les caractéristiques des assurés comme facteurs explicatifs des coûts peut paraître réducteur. En effet, le coût pour la compagnie d'assurance dépendra essentiellement des caractéristiques de la tierce partie impliquée dans l'accident qui sont non observables au moment de la signature du contrat d'assurance. Néanmoins, certains facteurs propres à l'assuré tels que sa préférence pour la vitesse, le type de couverture choisie ou encore le degré d'urbanisation de son lieu de résidence (Vanderlaan et Louter, 1986) peuvent avoir une influence significative sur le coût d'un sinistre supporté par la compagnie d'assurances¹.

Dans ce chapitre, nous abordons brièvement la modélisation des fréquences des accidents à l'aide d'une distribution de Poisson ainsi que d'une distribution binomiale inverse. Nous nous concentrons ensuite sur la modélisation des coûts de sinistres en présence d'une série de caractéristiques des assurés. Même si elle a beaucoup de succès dans la littérature empirique, nous n'utilisons pas la distribution Pareto car le seuil varie en fonction des caractéristiques et qu'il est dès lors peu évident de procéder à l'inférence statistique standard à partir des résultats. Nous utilisons plutôt la distribution GEV pour modéliser les coûts de sinistres à partir du premier euro. Néanmoins, comme on le verra par la suite, la nature des données impose de fixer un seuil. En outre, nous estimons les paramètres de la GEV par la méthode d'estimation alternative qui possède une série d'avantages qui ont été exposés dans le premier chapitre. Les résultats sont comparés avec ceux de la méthode d'estimation classique de la distribution GEV.

1. Dans notre échantillon, les coûts observés sont différents des coûts réels des sinistres. En effet, pour certains sinistres, en particulier ceux aux conséquences les plus dramatiques donnant lieu à une enquête judiciaire, le coût renseigné est en réalité une provision effectuée par la compagnie basée sur une estimation du coût attendu de l'accident. Pour d'autres sinistres, il s'agit d'un coût dont sont déduits les remboursements reçus des assurés à la suite de recours ainsi que la franchise en dégâts matériels. En aucun cas il n'est tenu compte d'éventuelles réassurances.

Notre analyse porte sur un échantillon original construit à partir d'un portefeuille d'assurance automobile d'une grande compagnie présente sur le marché belge. Le portefeuille d'assurance utilisé dans cette étude a le mérite d'être de grande taille (plus de 10 000 observations) et de contenir diverses informations sur les assurés ainsi que sur leur véhicule.

La suite de ce chapitre est organisée de la manière suivante : les données sont présentées dans la section 5.2. La section 5.3 présente une analyse des facteurs explicatifs du nombre de sinistres encourus sur un intervalle d'un an. Dans la section 5.4, nous présentons les résultats des estimations de la distribution GEV obtenus par la méthode classique ainsi que par la méthode alternative. Nous comparons également les résultats de la GEV avec ceux d'une distribution de Burr généralisée estimée par la méthode classique ainsi que par la méthode alternative. Dans la section 5.5 nous discutons des résultats en établissant les liens entre l'analyse des coûts et celle des fréquences d'accident.

5.2 Les données

La base de données est issue d'un portefeuille de polices d'assurance automobile d'une des plus grandes compagnies belges. L'analyse porte sur les sinistres survenus en 2007.

Les individus apparaissant dans la base de données ont tous été assurés auprès de la compagnie durant l'ensemble de l'année 2006. Dès lors, on peut tenir compte dans l'analyse, certes imparfaitement, de l'historique de sinistres des individus, ce qui peut être intéressant pour l'estimation du risque (Denuit *et al.*, 2003b). L'inconvénient de l'échantillon est qu'il ne contient pas de conducteur novice, ni d'individus qui auraient été éventuellement exclus d'une autre compagnie dans le courant de l'année 2006 en raison d'un comportement à risque.

A la suite d'une analyse préalable des données, plusieurs observations aberrantes ont été détectées (typiquement des conducteurs de moins de 17 ans ou de plus de 110 ans, des codes postaux fantaisistes et une personne de sexe non identifié). Vu l'impossibilité de corriger ces erreurs de manière satisfaisante, les données concernées ont été retirées de l'échantillon. Par ailleurs, la base de données contient des informations sur tous les types de véhicules assurables. Par souci de cohérence, l'analyse porte uniquement sur les véhicules automobiles de tourisme.

La base de données qui subsiste après avoir opéré ces manipulations se compose de 11638 observations. Parmi celles-ci, le nombre d'assurés ayant encouru au moins un sinistre durant l'année 2007 est de 1180, soit un peu plus de 10% du nombre total d'observations. Parmi ces sinistrés, 916 ont

eu exactement un accident, 210 ont eu deux accidents, 44 en ont eu trois, huit personnes ont eu quatre accidents et deux personnes ont eu cinq accidents.

Les variables utilisées dans l'analyse sont les suivantes :

- *le coût (C)* : il s'agit du coût moyen des sinistres engendrés par l'assuré au cours de l'année 2007 ;
- *l'âge (A)* de l'individu ;
- *le sexe (S)* : variable binaire prenant la valeur 1 si l'individu est une femme et 0 si c'est un homme ;
- *la langue (L)* : variable binaire prenant la valeur 1 si l'individu est francophone et 0 si il est néerlandophone ;
- *la densité de population (D)* : nombre d'habitants par km^2 dans la commune de résidence de l'individu ;
- *la fréquence de paiement* : fréquence avec laquelle l'individu règle sa prime d'assurance. La fréquence est mensuelle (Fm), trimestrielle (Ft), semestrielle (Fs) ou annuelle (Fa) ;
- *l'âge du véhicule (Av)* ;
- *la puissance du véhicule (Pow)* : puissance en kilowatts ;
- *la valeur du véhicule (Val)* : prix catalogue en euros ;
- *le type de contrat* : le contrat de l'assuré peut couvrir la responsabilité civile (Rc), le vol (Vol), les dégâts matériels (Deg), l'incendie (Inc), le bris de vitre (Br) et les risques de poursuite juridique (Jur). Ces variables sont toutes des variables binaires ;
- *le nombre d'accidents encourus durant l'année précédente* : les assurés se trouvent dans l'une des catégories suivantes : pas accidenté ($0Ac$), un accident ($1Ac$), deux accidents ($2Ac$) et trois accidents et plus ($3Ac$). Ces quatre variables sont des variables binaires.

Deux limites de l'échantillon sont à épingle. Premièrement, la distance parcourue annuellement par les assurés n'est pas renseignée, or celle-ci devrait logiquement être une variable explicative de première importance pour l'estimation de la fréquence des sinistres (Vanderlaan et Louter, 1986). Néanmoins, notre intérêt principal porte sur l'analyse du coût des sinistres et on peut

raisonnablement penser que cette variable n'influence pas significativement le coût d'un accident donné. Deuxièmement, les coûts reportés dans la base de données pour chaque accident correspondent au coût net pour la compagnie, c'est à dire au coût total duquel sont déduites les éventuelles sommes récupérées auprès des assurés lors de procédures de recours. Nous reviendrons sur cet inconvénient dans la section 5.4 lors de l'estimation de la distribution des coûts car ceci a une importance capitale sur les résultats.

Dans l'analyse qui suit, selon le type de variables, celles-ci peuvent apparaître en niveau ou en logarithme, l'effet peut être linéaire ou quadratique et différents effets croisés sont également pris en considération.

5.3 Fréquence des accidents

5.3.1 Méthode

Le but de cette analyse est de mesurer les effets des différentes variables disponibles sur la probabilité d'occurrence des accidents². Nous modélisons la fréquence des accidents survenus pendant l'année pour chaque individu i (N_i) à l'aide d'une distribution de Poisson dont la moyenne a une forme exponentielle linéaire, c'est à dire que le modèle prend la forme :

$$N_i \sim \text{Poisson}(\exp(X_i' \beta)), \quad (5.1)$$

où la matrice X contient les variables explicatives d'intérêt et le vecteur β contient les paramètres associés.

L'estimation du modèle (5.1) est réalisée par la méthode du maximum de vraisemblance dont le principe est de déterminer la valeur des coefficients β qui rend les observations recueillies les plus probables.

Les modèles de Poisson supposent implicitement que l'espérance conditionnelle de la variable expliquée ($E(N|x)$) est égale à sa variance conditionnelle ($Var(N|x)$). En pratique, cette hypothèse est rarement vérifiée et il est beaucoup plus fréquent que la variance évolue proportionnellement avec l'espérance. Dans ce cas, les modèles de Poisson ont tendance à exagérer le pouvoir explicatif des covariables. Un modèle binomial négatif permet de tenir compte d'une "sur-dispersion" des observations, en supposant que $Var(N|x) = (1 + \eta^2)E(N|x)$, pour $\eta^2 > 0$. Nous avons également réalisé l'analyse sur base de ce modèle.

2. Nous remercions vivement Michel Denuit pour ses conseils avisés ayant permis d'améliorer sensiblement cette section ainsi que la section suivante consacrée à l'interprétation des résultats

Les résultats qui suivent sont à interpréter avec la plus grande prudence.

D'une part, comme expliqué précédemment la distance parcourue par les individus au cours de la période étudiée, qui constitue une variable primordiale dans ce genre d'analyse, n'est pas observée. Par conséquent, s'il y a la moindre corrélation entre cette variable et une ou plusieurs variables apparaissant dans le modèle de régression, les estimateurs des coefficients s'en trouveront biaisés. Or, il y a de bonnes raisons de penser que la distance parcourue par les individus soit intrinsèquement liée avec des caractéristiques utilisées dans l'analyse telles que le sexe, l'âge des individus et le type de véhicule.

D'autre part, nous nous contentons d'un modèle linéaire dans les paramètres et quadratique dans les variables explicatives. Cette procédure ne garantit aucunement que les effets de ces variables soient correctement pris en compte. Des modèles de régression de type Poisson non-linéaire semblent plus appropriés (voir Denuit et Lang, 2004), mais ceci dépasse le cadre d'analyse de cette recherche qui, rappelons-le, vise essentiellement à modéliser le coût des sinistres.

En outre, l'inclusion du nombre passé de sinistres dans les régresseurs d'un modèle de type (5.1) est proscrit par certains puristes qui préfèrent généralement utiliser des modèles de panel. Ce genre de modèle n'est pas applicable dans notre cas en raison de la nature même des données.

Malgré toutes ces limites, nous avons préféré présenter les résultats de l'estimation d'un modèle de régression de type Poisson et binomiale négative plutôt que de reproduire les corrélations simples entre les différentes variables disponibles. L'analyse présentée dans la section suivante doit être considérée comme une simple analyse exploratoire des données.

5.3.2 Application

Suivant le principe de l'approche "general-to-specific", une succession de restrictions ont été imposées au modèle de départ, qui se voulait le plus général possible, afin de dégager le modèle le plus parcimonieux. Chaque ensemble de restrictions imposées a été testé sur base d'un test traditionnel de rapport de vraisemblance.

Différentes formes fonctionnelles ont été testées et les transformations les plus vraisemblables des variables ont été conservées.

Le modèle final obtenu à la fin de cette procédure de sélection prend la forme : $N_i \sim \text{Poisson}(\exp(\lambda_i))$, avec :

$$\begin{aligned} \lambda_i = & \alpha + \beta_1 \log(A_i) + \beta_2 Av_i + \beta_3 \log(Pow_i) + \beta_4 \log(Pow_i)^2 + \\ & \beta_5 Val_i + \beta_6 D_i + \beta_7 D_i^2 + \beta_8 Fm_i + \beta_9 Fs_i + \beta_{10} Fa_i + \\ & \beta_{11} Vol_i + \beta_{12} 1Ac_i + \beta_{13} 2Ac_i + \beta_{14} 3Ac_i + \\ & \beta_{15}(Fm_i * \log(A_i)) + \beta_{16}(Av_i * Ac06_i) + \epsilon_i, \end{aligned} \quad (5.2)$$

où $Ac06_i$ est une variable binaire prenant la valeur 1 lorsque l'individu a connu au moins un accident durant l'année 2006 et les autres variables ont été décrites dans la section précédente. Les résultats de l'estimation du modèle (5.2) sont reportés dans la table 5.1. Les résultats de l'estimation d'une binomiale négative sont également reportées dans cette même table. Comme attendu, les écarts-types estimés des coefficients sont (très) légèrement plus élevés avec la binomiale. Etant donné que le paramètre tenant compte de la sur-dispersion (η^2) prend une valeur faible, les résultats des deux estimations sont fort proches.

Les résultats indiquent que les variables explicatives sont clairement statistiquement significatives. Néanmoins, le pouvoir explicatif global du modèle (5.2) n'est pas très bon. En effet, le carré de la corrélation entre le nombre de sinistres observés et prédits par la modèle est inférieur à 10%.

Avant toute interprétation des résultats, il est bon de rappeler que l'échantillon analysé est issu d'une sous-population des individus qui étaient déjà assurés auprès de la compagnie durant l'année précédant la période d'observation. Remarquons également que, selon la manière dont la compagnie se comporte face à la segmentation des marchés (notamment si la compagnie est moins pointilleuse que la moyenne quant aux clients à risque ou bien si elle est moins généreuse que les autres pour les clients moins risqués), les résultats peuvent ne pas être représentatifs de la population des conducteurs belges.

L'âge de l'individu influence positivement le nombre moyen de sinistres encourus. Toutefois, s'y ajoute un effet défavorable pour les jeunes selon le mode de financement de leur prime (voir infra).

Le degré d'urbanisation du lieu de résidence de l'assuré a également un effet positif sur le nombre moyen d'accidents. L'effet est quadratique avec un impact marginal du degré de densité de moins en moins prononcé au fur et à mesure que le niveau de la densité de population augmente.

Le nombre d'accidents attendus est plus important pour les individus ayant connu au moins un accident durant l'année précédente. En outre, les individus qui ont connu au moins trois accidents durant l'année 2006 sont plus susceptibles de connaître un nombre élevé d'accidents en 2007 que ceux n'ayant

connu qu'un ou deux accidents durant cette année. Les assurés qui règlent leur facture de prime d'assurance avec une fréquence mensuelle présentent un risque d'accident plus important que les individus qui ont une fréquence de paiement plus faible.

Les individus ayant souscrit une assurance contre le vol ou une garantie en cas de problème juridique présentent un risque accru d'accidents. Cette variable pose un problème de causalité. L'argument classique est celui de l'aléa moral (plus on est assuré plus on prend de risques). Il faut probablement y ajouter le fait que ceux qui, par la nature de leur activité, savent qu'ils courent un risque plus grand, vont se couvrir davantage.

L'âge de la voiture influence négativement le nombre moyen d'accidents tandis que la valeur à l'achat exerce un effet positif (il y a là un effet assez mécanique, les voitures gravement accidentées ne devenant jamais vieilles).

La puissance du véhicule semble jouer un rôle plus complexe, avec un effet quadratique impliquant une moindre exposition au risque d'accident pour les véhicules les moins puissants ainsi que pour les véhicules les plus puissants.

Les effets croisés entre les variables explicatives apparaissant dans le modèle sont peu évidents à interpréter. D'une part, il apparaît que pour les individus réglant leur facture avec une fréquence mensuelle, l'effet de l'âge sur le nombre moyen de sinistres est négatif. Ainsi, les jeunes qui règlent leur facture mensuellement présentent un risque plus élevé que les autres jeunes. D'autre part, pour les individus qui ont connu au moins un accident durant l'année précédente, l'âge de la voiture a un effet positif sur le nombre moyen d'accidents subis durant l'année en cours.

Il est à noter que nous avons tenté d'introduire l'information sur le sexe de l'assuré dans le modèle et que celui-ci n'était pas statistiquement significatif lorsque toutes les variables précitées sont prises en compte. Il est communément admis que les conducteurs jeunes et de sexe masculin sont plus risqués que les autres, en particulier lorsqu'il sont au volant d'un véhicule relativement puissant. Nous avons dès lors testé les effets croisés du sexe, alternativement, avec l'âge et la puissance du véhicule, sans succès. L'absence d'effet du sexe sur la probabilité de sinistre est probablement à mettre en lien avec les nombreuses réserves émises précédemment quant à la portée des résultats.

TABLE 5.1 – Résultats des estimations

	Poisson		Binomiale négative	
	<i>Coef.</i>	<i>Ec.-typ.</i>	<i>Coef.</i>	<i>Ec.-typ.</i>
c	-13.56	3.58	-12.88	3.90
Fm	2.21	.76	2.17	.77
Fs	-.27	.09	-.26	.09
Fa	-.50	.08	-.50	.08
Deg	.61	.07	.57	.07
Jur	.27	.15	.28	.15
$1Ac$	.20	.11	.25	.11
$2Ac$	.21	.13	.26	.13
$3Ac$	.59	.18	.68	.20
$\log(A)$	.32	.10	.33	.10
D	.10	.03	.10	.03
D^2	-.01	.00	-.01	.00
Av	-.11	.01	-.10	.01
$\log(Pow)$	4.60	1.71	4.33	1.86
$\log(Pow)^2$	-.51	.21	-.49	.23
Val	.01	.00	.01	.00
$Fm * \log(A)$	-.55	.20	-.53	.20
$Ac06 * Av$	.09	.02	.08	.02
η^2	*	*	.19	.03

5.4 Coût des sinistres

5.4.1 Sélection de l'échantillon

Rappelons que les coûts renseignés dans notre échantillon correspondent aux coûts nets pour la compagnie, c'est à dire les coûts totaux desquels sont éventuellement enlevées les sommes remboursées par les clients en raison des recours intentés par la compagnie. Cette caractéristique des données perturbe le travail d'analyse. En effet, les individus contre qui la compagnie a introduit un recours sont ceux n'ayant pas respecté les règles du contrat, notamment ceux ayant occasionné un accident en état d'ivresse. On peut raisonnablement penser que ces individus sont probablement parmi les plus couteux pour une compagnie d'assurance et qu'ils possèdent des caractéristiques systématiques. Or, en raison de la nature des données, ces individus n'affichent que des coûts de sinistre très faibles. Malheureusement, l'information concernant le recours éventuel de la compagnie n'est pas disponible. Pour réaliser l'analyse, nous avons fixé, de manière arbitraire, un seuil en dessous duquel les coûts nets observés sont considérés comme suspects. Différentes valeurs ont été testées pour ce seuil (50, 75, 100, 125 et 150 euros). Il apparaît que la fixation d'un seuil améliore sensiblement les résultats des différentes estimations réalisées, mais

que la valeur choisie pour ce seuil entre 50 et 150 euros ne change pratiquement pas les résultats. Dès lors, une valeur faible a été retenue pour ce seuil (50 euros), ce qui n'ampute l'échantillon que de 39 observations, laissant un total de 1141 observations.

5.4.2 Régression classique

L'objet de cette section est d'estimer les effets des différentes variables explicatives disponibles sur la distribution de coûts des sinistres observés durant l'année 2007. La première étape de notre méthode consiste à estimer une régression par moindres carrés ordinaires du coût exprimé en logarithme sur les différentes variables disponibles. Partant d'une spécification aussi générale que possible incluant l'ensemble des variables en niveau, au carré et les effets croisés, nous avons procédé comme d'habitude à la recherche du modèle le plus parcimonieux sous réserve d'une perte en vraisemblance statistiquement significative. Cette procédure nous a amené à sélectionner le modèle de régression suivant :

$$\begin{aligned} \log(C_i) = & \alpha + \beta_1 \log(A_i) + \beta_2 \log(A_i)^2 + \beta_3 Av_i + \beta_4 \log(Val_i) + \\ & \beta_5 D_i + \beta_6 D_i^2 + \beta_7 Deg_i + \beta_8 Br_i + \epsilon_i, \end{aligned} \quad (5.3)$$

où ϵ_i est un terme d'erreur aux propriétés classiques de bruit blanc et les autres variables sont décrites dans la section 5.2.

L'ajustement statistique du modèle (5.3) indique clairement qu'il s'agit d'une mauvaise spécification. Toutefois cette étape est utile pour sélectionner la forme fonctionnelle et pour obtenir des valeurs initiales des paramètres qui seront utilisées pour l'estimation de la GEV dans la section suivante.

5.4.3 Le modèle GEV

Méthode d'estimation classique

La distribution GEV (1.1) est utilisée pour modéliser le coût des sinistres exprimé en logarithme en permettant au paramètre de position (μ) de varier en fonction des différentes caractéristiques pertinentes X_i identifiées dans la section précédente et apparaissant dans le modèle (5.3). Le modèle GEV prend la forme :

$$G(\log C_i) = \exp \left\{ - \left[1 + \xi \left(\frac{\log C_i - X_i' \beta}{\sigma} \right) \right]^{-1/\xi} \right\}. \quad (5.4)$$

Les résultats de l'estimation du modèle (5.4) réalisée par la méthode traditionnelle sont présentés dans les 2^e et 3^e colonnes de la table 5.2.

Méthode d'estimation alternative

La méthode AGEV consiste en l'estimation du système suivant :

$$\log C_{(i)} = X'_{(i)}\beta - \frac{\sigma}{\xi} \left[1 - \left(-\log(F_{(i)}(\alpha)) \right)^{-\xi} \right] + \epsilon_{(i)}, \quad (5.5)$$

avec $F_{(i)}(\alpha)$ donné par la fonction (1.5). Vu le nombre élevé d'observations, le choix du paramètre α dans la fonction (1.5) n'influence pas sensiblement les résultats de l'estimation de (5.5). Les résultats ont été obtenus en fixant la valeur de ce paramètre égale à zéro.

Dans (5.5), les observations sont classées par ordre croissant de la valeur de leur résidu obtenu lors de l'estimation de (5.3). L'estimation des paramètres du modèle (5.5) est réalisée par la méthode du pseudo maximum de vraisemblance. Les résultats sont reportés dans les deux dernières colonnes de la table 5.2.

On notera au passage qu'une valeur de $\sigma = 0$ dans (5.5) renvoie au modèle linéaire de la section précédente tandis que $\xi = 0$ implique que (5.5) devienne une distribution de Gumbel.

Les différents avantages de la méthode d'estimation alternative de la distribution GEV par rapport à la méthode d'estimation traditionnelle ont été présentés dans le premier chapitre. Dans la présente application, un avantage pratique très intéressant est le fait que la condition d'existence (1.2) est automatiquement vérifiée lorsque la méthode AGEV est utilisée, ce qui n'est pas garanti par la méthode classique d'estimation. En outre, la vérification des restrictions imposées dans la section précédente est beaucoup plus facile à réaliser avec le modèle (5.5) qu'avec le modèle (5.4).

Les résultats

Les enseignements tirés des deux méthodes convergent, les coefficients estimés n'étant pas sensiblement différents. On peut en effet espérer qu'avec un échantillon de plus de 1000 observations le biais inhérent à la méthode d'estimation classique soit devenu négligeable. Néanmoins, la qualité de l'ajustement de la méthode alternative est légèrement supérieure à celle de l'estimation classique, avec une somme des carrés des résidus plus faible. Les écarts-types des coefficients sont également généralement réduits par rapport aux résultats de l'estimation traditionnelle du modèle ³.

Conformément à l'intuition économique, le paramètre de courbure (ξ) prend une valeur négative, impliquant que pour un ensemble de caractéristiques

3. Comme expliqué précédemment, en raison de la nature de la variable $F_{(i)}(\alpha)$ dans le modèle (5.5), les écarts-types des coefficients ont été estimés par bootstrap (cf. section 1.2.3, supra).

donné le coût du sinistre ne peut dépasser un certain plafond.

Le sexe de l'assuré a un effet statistiquement significatif sur le coût occasionné suite à un sinistre, avec un coût en moyenne plus élevé pour les femmes. On reviendra ultérieurement sur ce résultat.

L'âge joue un rôle significatif et a un effet non linéaire, avec un coût minimum pour les individus d'une quarantaine d'année, un coût progressant très vite lorsque l'âge descend vers 18 ans et un coût progressant lentement jusque l'âge de 85 ans.

Les personnes qui ont souscrit une assurance contre les dégâts matériels et ceux qui vivent dans une localité à faible densité de population sont associées à un coût moyen des sinistres plus élevé.

Le coût moyen des sinistres occasionnés augmente en même temps que la valeur à l'achat du véhicule et diminue avec son âge.

TABLE 5.2 – Résultats des estimations de la GEV

	Clas.		Alter.	
	<i>Coef.</i>	<i>Ec.-typ.</i>	<i>Coef.</i>	<i>Ec.-typ.</i>
c	13.43	3.94	14.58	2.46
S	.13	.06	.13	.03
Deg	.44	.15	.46	.08
Br	-.37	.17	-.43	.07
$\log(A)$	-5.45	2.11	-6.07	1.52
$\log(A)^2$	.72	.28	.81	.21
D	-.45	.29	-.56	.10
D^2	.33	.20	.42	.11
Av	-.03	.01	-.04	.01
$\log(Val)$	.29	.08	.30	.03
σ	.95	.03	.97	.03
ξ	-.19	.03	-.20	.03
SSR	13.98		13.62	

Aucune des deux méthodes ne passe les tests usuels de mauvaise spécification (Ramsey et hétéroscédasticité des résidus), les résultats étant légèrement moins défavorables pour la méthode alternative. Les Q-Q plot présentés aux figures 5.1 et 5.2 indique que l'ajustement est relativement peu satisfaisant dans les queues de la distribution. La distribution des fréquences des résidus à la figure 5.3, avec des queues particulièrement plates, indique également que cela provient d'un très petit nombre d'observations pour lesquelles nous ne pouvons pas distinguer s'il s'agit effectivement d'une valeur

extrême ou d'une donnée aberrante. Il est aussi fort possible que le choix d'un seuil plus élevé (ici 50 euros) améliore l'ajustement dans l'extrémité inférieure de la distribution.

FIGURE 5.1 – Q-Q plot de l'estimation de la GEV : méthode classique

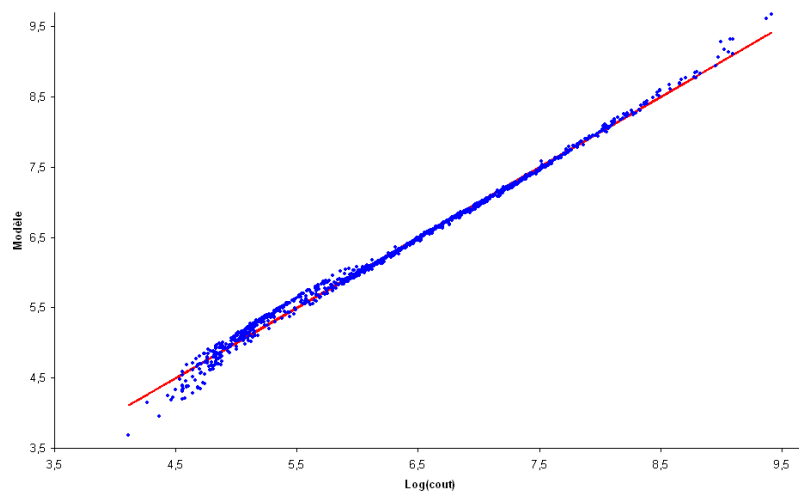


FIGURE 5.2 – Q-Q plot de l'estimation de la GEV : méthode alternative

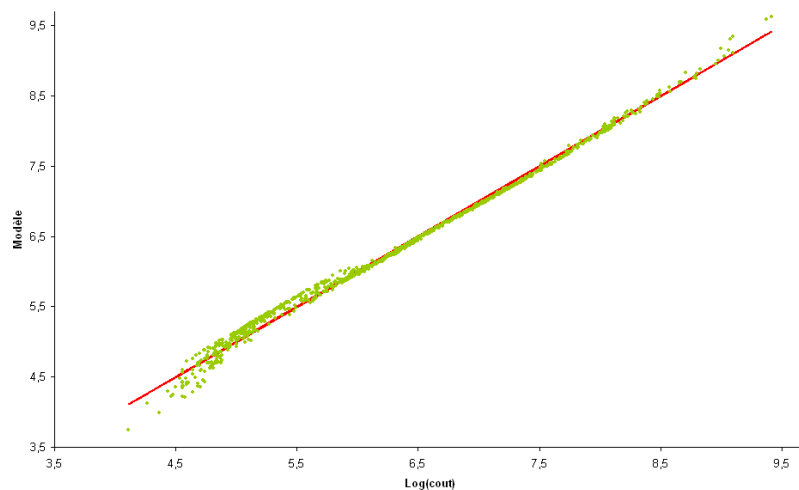
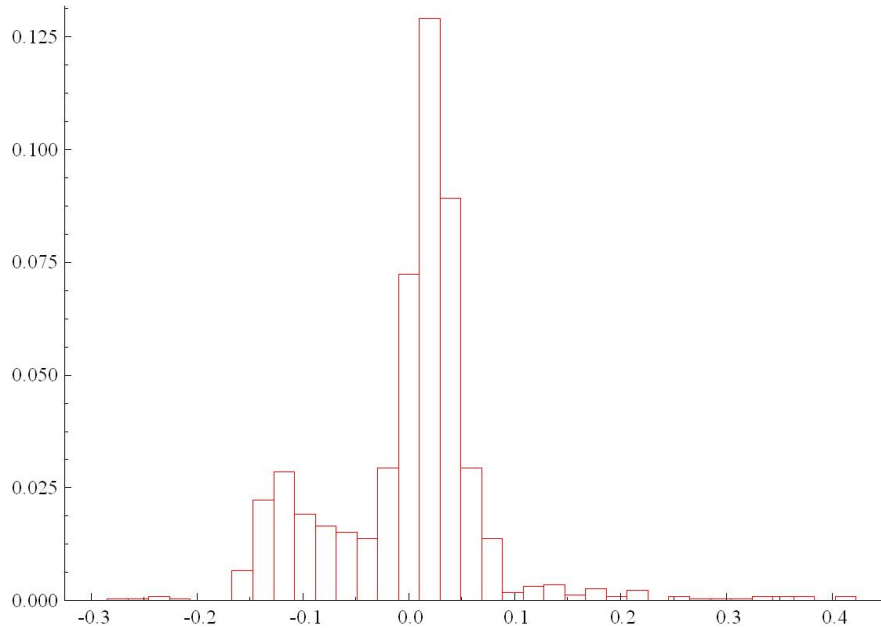


FIGURE 5.3 – Distribution de fréquence des résidus (méthode AGEV)



5.4.4 Distribution de Burr généralisée

Afin de tenter d'améliorer l'ajustement du modèle, nous proposons d'estimer une distribution de Burr généralisée qui est utilisée pour modéliser des variables présentant un comportement largement asymétrique, notamment en hydrologie (Shao *et al.*, 2004) et en sciences actuarielles (Beirlant *et al.*, 1998). Par rapport à la GEV, la distribution de Burr généralisée possède un paramètre de courbure supplémentaire, ce qui la rend plus flexible. Elle prend tout son sens dans notre application dans la mesure où elle englobe comme cas particulier à la fois les distributions de la famille des Weibull à deux paramètres et la Pareto généralisée qui est utilisée pour modéliser des observations extrêmes au-delà d'un certain seuil.

La fonction de répartition de la Burr généralisée prend la forme :

$$F(x) = 1 - \left\{ 1 - k \left(\frac{x - \mu}{\sigma} \right)^\tau \right\}^{1/k}, \quad (5.6)$$

où μ est un paramètre de position, σ un paramètre d'échelle et τ et k sont

deux paramètres de courbure. La fonction de densité de la Burr généralisée est donnée par :

$$f(x) = \tau \lambda^{-1} \left(\frac{x - \mu}{\sigma} \right)^{\tau-1} \left\{ 1 - k \left(\frac{x - \mu}{\sigma} \right)^{\tau} \right\}^{1/k-1}, \quad (5.7)$$

et le support de x est $0 \leq x < \infty$ si $k \leq 0$, et $0 \leq x < \lambda/k^{1/\tau}$ si $k > 0$.

Nous modélisons les coûts de sinistres, exprimés en logarithme, à l'aide de la distribution (5.6) avec un paramètre de position (μ) qui varie en fonction des différentes caractéristiques pertinentes X_i apparaissant dans le modèle (5.3). Les paramètres de la distribution (5.6) ont été estimés par la méthode classique du maximum de vraisemblance ainsi que par la méthode alternative. Dans le cas de la méthode alternative, les écarts-types sont à nouveau obtenus par la méthode du bootstrap. Les résultats des estimations apparaissent dans la table 5.3 et les Q-Q plot découlant des estimations sont reportés aux figures 5.4 et 5.5.

Si la valeur des coefficients, et donc l'interprétation des résultats ne sont que très faiblement modifiées par rapport à celles découlant des résultats des estimations de la GEV, le paramètre de courbure additionnel permet au modèle de mieux capter le comportement de la variable expliquée dans les extrémités de la distribution (figures 5.4 et 5.5 vs. 5.1 et 5.2), si bien que la distribution de Burr rejette la GEV sur base des critères d'information traditionnel (AIC et BIC).

TABLE 5.3 – Résultats des estimations de la Burr généralisée

	Clas.		Alter.	
	<i>Coef.</i>	<i>Ec.-typ.</i>	<i>Coef.</i>	<i>Ec.-typ.</i>
c	10.50	3.18	12.76	5.01
S	.17	.06	.14	.02
Deg	.34	.18	.44	.03
Br	-.33	.19	-.42	.05
$\log(A)$	-4.34	1.72	-5.98	.97
$\log(A)^2$	.57	.23	.79	.13
D	-.02	.29	-.54	.07
D^2	-.09	.19	.40	.08
Av	-.03	.01	-.03	.01
$\log(Val)$	.21	.08	.29	.01
σ	2.57	.07	2.59	.08
τ	1.95	.11	1.99	.15
k	.20	.05	.19	.05

FIGURE 5.4 – Q-Q plot de l'estimation de la Burr généralisée : méthode classique

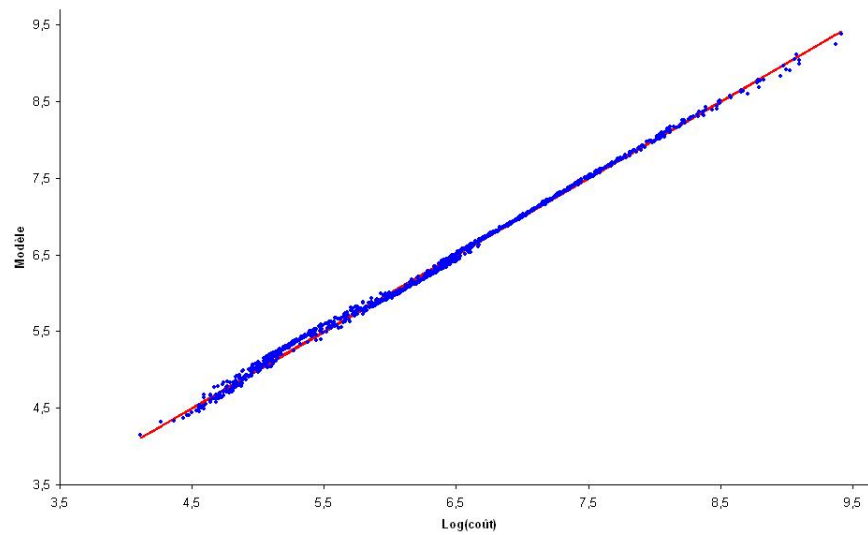
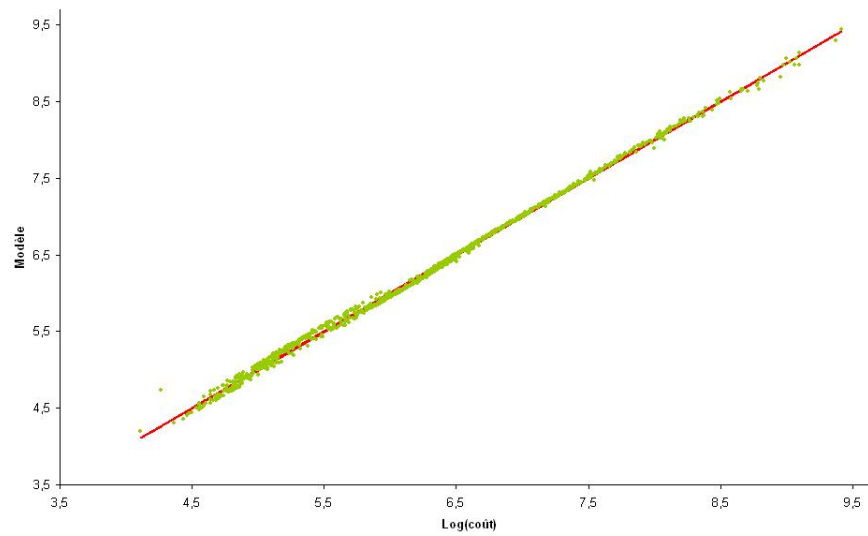


FIGURE 5.5 – Q-Q plot de l'estimation de la Burr généralisée : méthode alternative



5.5 Lien entre l'analyse des coûts et des fréquences

Il est assez intuitif de constater que le passé sinistre des individus est un facteur explicatif du risque d'accident alors qu'il n'influence pas le coût en cas de sinistre.

De même, il y a une certaine logique au fait que les ressortissants des communes à faible densité de population sont moins exposés au risque d'accident, mais que, lorsqu'ils sont impliqués dans un accident, celui-ci coûte relativement cher. En effet, les assurés vivant dans les grandes villes connaissent vraisemblablement plus de petits accrochages réalisés à vitesse faible que les individus résidant dans les communes rurales pour lesquels les accidents sont probablement réalisés à une vitesse plus élevée, impliquant davantage de dommages.

Il y a de bonnes raisons de penser que le choix de la fréquence de paiement de la prime d'assurance dépend de caractéristiques socio-économiques telles que le revenu, la profession ou le milieu social de l'assuré. Dès lors, il paraît sensé de constater que cette variable influence la probabilité d'occurrence des sinistres. Par contre, aucun effet n'a été détecté sur le coût du sinistre.

Il apparaît dans la littérature appliquée dans ce domaine que les conductrices ont une fréquence d'accident moindre que les conducteurs (voir par exemple Denuit *et al.*, 2003b), ce qu'on ne retrouve pas ici. Par contre, et plus curieusement, le fait d'être une femme influence positivement le coût des dommages en cas de sinistre. Une explication plausible pourrait être que les femmes sont plus averses au risque que les hommes et se couvrent davantage contre les dégâts matériels et les poursuites juridiques, impliquant un coût plus élevé pour la compagnie en cas de sinistre. Toutefois, il n'y a pas de différence significative du degré de couverture entre les hommes et les femmes dans notre échantillon.

L'âge de l'assuré apparaît avec un effet positif et linéaire sur la fréquence d'accident, ce qui est étonnant dans la mesure où l'opinion publique considère généralement les jeunes comme étant les conducteurs les plus à risque. Cette affirmation est en fait valable mais seulement pour les jeunes qui règlent leur facture avec une fréquence mensuelle. Au niveau du coût, l'effet de l'âge est non linéaire avec un coût plus élevé des dommages lorsque l'assuré impliqué est très jeune ou très âgé.

La couverture des dégâts matériels et des frais de justice par le contrat d'assurance influence positivement à la fois le risque et le coût des dommages en cas de sinistre. Le fait que ces couvertures supplémentaires augmentent

le risque d'accidents met probablement en évidence le problème souvent évoqué de hasard moral auquel il faudrait, comme mentionné précédemment, ajouter la nature de l'activité du conducteur. L'effet positif de ces couvertures supplémentaires sur le coût net de l'accident pour la compagnie va de soi.

Il apparaît que les voitures plus anciennes ont une fréquence d'accident moindre (avec la réserve mentionnée plus haut) et sont associées à un coût des dommages moins élevé pour la compagnie d'assurance. Ceci est probablement lié en partie au fait que les véhicules plus anciens parcourent moins de km et roulent moins vite.

La puissance du véhicule a un effet sur le risque d'accident puisque celui-ci est plus faible pour les véhicules de faible et de forte puissance. Les véhicules de faible puissance parcourent probablement une distance plus courte que les véhicules de puissance moyenne, tandis que les véhicules de plus forte puissance sont probablement entre les mains de conducteurs plus expérimentés. Curieusement, aucun effet de la puissance n'a pu être détecté pour le coût en cas de sinistre.

La valeur du véhicule a un effet positif évident sur le coût alors que l'effet positif sur le risque est plus difficile à interpréter.

5.6 Conclusion

L'objectif de cette étude était de discuter des techniques de modélisation de la probabilité d'occurrence et du coût des sinistres automobiles supportés par une compagnie d'assurance.

D'après l'analyse des facteurs de risque, il ressort que les individus les plus à risque semblent être les individus d'un âge avancé vivant en milieu urbain qui sont assurés contre les dégâts matériels et les poursuites juridiques, qui ont déjà été accidentés durant l'année précédente et qui conduisent un véhicule cher et récent d'une puissance moyenne.

La distribution GEV a été utilisée pour modéliser les coûts des sinistres en permettant au paramètre de position de varier en fonction d'une série de caractéristiques observables. La méthode d'estimation alternative de cette distribution a été utilisée et les performances empiriques de celle-ci ont été comparées avec celles de la méthode d'estimation classique. Les enseignements que l'on peut tirer des estimations de la GEV réalisées à l'aide des deux méthodes convergent, les coefficients estimés étant pratiquement identiques. Il ressort de l'analyse que les individus occasionnant un coût total par accident plus élevé semblent être les (très) jeunes femmes vivant dans une zone rurale qui ont souscrit une assurance contre les dégâts matériels et qui conduisent un

véhicule récent et relativement cher à l'achat.

Notons tout de même que l'ajustement général de la méthode alternative est légèrement préférable à l'ajustement de l'estimation traditionnelle, lorsque l'on se base sur la valeur de la somme des carrés des résidus. Néanmoins, quelle que soit la méthode d'estimation utilisée, l'ajustement du modèle GEV est peu satisfaisant pour les données les plus extrêmes, en particulier dans le bas de la distribution. En effet, les deux méthodes surestiment systématiquement les coûts les plus élevés et sous-estiment les coûts les plus faibles de l'échantillon. Cet ajustement moins bon du modèle dans les queues de la distribution constitue un problème sérieux. Toutefois, cela ne concerne qu'un très petit nombre d'observations (une petite douzaine) dont la nature exacte n'a pu être validée. Remarquons que ce problème pourrait être dû à la nature de la variable "coût" qui est en réalité un coût net des sinistres incombant à la compagnie d'assurance et non le coût réel des dommages causés par l'assuré.

Nous avons également montré dans ce chapitre que pour des échantillons de grande taille, l'utilisation d'une distribution plus flexible de type Burr généralisée, permet d'obtenir un meilleur ajustement du modèle, en particulier dans les queues de la distribution. Dans cette application, les résultats des estimations de la distribution Burr réalisées par les deux méthodes sont forts proches. Notons que dans des applications à des échantillons de taille beaucoup plus modeste, ce qui est typiquement le cas en réassurance, la méthode d'estimation alternative devrait s'avérer plus performante que la méthode classique pour l'estimation de cette distribution de Burr généralisée.

Alors que le genre de l'assuré n'a pas d'influence significative sur la probabilité d'occurrence de sinistres, cette variable affiche une importance statistiquement significative dans la détermination du coût par accident. Ce résultat suggère, pour cette compagnie d'assurance particulière, qu'appliquer une différenciation des primes en faveur des femmes (ce qui était le cas de certaines compagnies jusqu'à l'interdiction décidée récemment par le législateur belge) ne lui aurait pas été bénéfique.

Bibliographie

Anselin, L. 1990. "Spatial Dependence and Spatial Structural Instability in Applied Regression-Analysis." *Journal of Regional Science*, 30(2), pp. 185-207.

Anselin, Luc. 1988. *Spatial Econometrics : Methods and Models*. Kluwer.

Atkinson, S. E. et R. Halvorsen. 1984. "A New Hedonic Technique for Estimating Attribute Demand - an Application to the Demand for Automobile Fuel Efficiency." *Review of Economics and Statistics*, 66(3), pp. 417-26.

Baker, T. A. 1997. "Quality-Adjusted Price Indexes for Portable Computers." *Applied Economics*, 29(9), pp. 1115-23.

Barnett, William A. et Yul W. Lee. 1985. "The Global Properties of the Minflex Laurent, Generalized Leontief, and Translog Flexible Functional Forms," *Econometrica*, 53, pp. 1421-37.

Beirlant, J. 2004. *Statistics of Extremes : Theory and Applications*. Hoboken, NJ : Wiley.

Beirlant, J. ; Y. Goegebeur ; R. Verlaak et P. Vynckier. 1998. "Burr Regression and Portfolio Segmentation." *Insurance Mathematics and Economics*, 23(3), pp. 231-50.

Belsley, David A. ; Edwin Kuh et Roy E. Welsch. 1980. *Regression Diagnostics : Identifying Influential Data and Sources of Collinearity*. New York ; Chichester : Wiley.

Beron, K. J. ; J. C. Murdoch ; M. A. Thayer et W. P. M. Vijverberg. 1997. "An Analysis of the Housing Market before and after the 1989 Loma Prieta Earthquake." *Land Economics*, 73(1), pp. 101-13.

Bowen, W. M. ; B. A. Mikelbank et D. M. Prestegaard. 2001. "Theoretical and Empirical Considerations Regarding Space in Hedonic Housing Price

Model Applications." *Growth and Change*, 32(4), pp. 466-90.

Breusch, T. S. et A. R. Pagan. 1980. "The Lagrange Multiplier Test and Its Applications to Model-Specification in Econometrics." *Review of Economic Studies*, 47(1), pp. 239-53.

Brueckner, J. K. et P. F. Colwell. 1983. "A Spatial Model of Housing Attributes - Theory and Evidence." *Land Economics*, 59(1), pp. 58-69.

Can, A. 1990. "The Measurement of Neighborhood Dynamics in Urban House Prices." *Economic Geography*, 66(3), pp. 254-72.

Can, A. et I. Megbolugbe. 1997. "Spatial Dependence and House Price Index Construction." *Journal of Real Estate Finance and Economics*, 14, pp. 203-22.

Casetti, E. 1972. "Generating Models by the Expansion Method." *Geographical Analysis*, 4, pp. 81-91.

Chow, Ven Te; David R. Maidment et Larry W. Mays. 1988. *Applied Hydrology*. New York; London : McGraw-Hill.

Clark, D. E. ; L. Michelbrink ; T. Allison et W. C. Metz. 1997. "Nuclear Power Plants and Residential Housing Prices." *Growth and Change*, 28(4), pp. 496-519.

Coles, S. 2001. *An Introduction to Statistical Modeling of Extreme Values*. London ; New York : Springer.

Combris, P. ; S. Lecocq et M. Visser. 1997. "Estimation of a Hedonic Price Equation for Bordeaux Wine : Does Quality Matter?" *Economic Journal*, 107(441), pp. 390-402.

Cooray, A. et M. M. A. Ananda. 2005. "Modeling Actuarial Data with a Composite Lognormal-Pareto Model." *Scandinavian Actuarial Journal*, 5, pp. 321-34.

Cropper, M. L. ; L. B. Deck et K. E. McConnell. 1988. "On the Choice of Functional Form for Hedonic Price Functions." *Review of Economics and Statistics*, 70(4), pp. 668-75.

Cunnane, C. 1978. "Unbiased Plotting Positions - a Review." *Journal of Hydrology*, 37, pp. 205-22.

Davidson, R. et J. G. MacKinnon. 2004. *Econometric Theory and Methods*. New York : Oxford University Press.

Davison, A. C. et R. L. Smith. 1990. "Models for Exceedances over High Thresholds." *Journal of the Royal Statistical Society Series B-Methodological*, 52(3), pp. 393-442.

Denuit, M. ; A. Cebrian et P. Lambert. 2003a. "Generalized Pareto Fit to the Society of Actuaries'large Claims Database." *North Amercian Actuarial Journal*, 7(3), pp. 18-36.

Denuit, M. et S. Lang. 2004. "Non-Life Rate-Making with Bayesian Gams." *Insurance Mathematics and Economics*, 35(3), pp. 627-47.

Denuit, M. ; S. Pitrebois et J.F. Walhin. 2003b. "Tarification Automobile Sur Données De Panel." *Bulletin des Actuairees Suisses*, 1, pp. 51-81.

Di Vittorio, A. et V. Ginsburgh. 1996. "Des Enchères Comme Révélateurs Du Classement Des Vins." *Journal de la Société de Statistique de Paris*, 137(2).

Dubin, R. A. 1988. "Estimation of Regression-Coefficients in the Presence of Spatially Auto-Correlated Error Terms." *Review of Economics and Statistics*, 70(3), pp. 466-74.

Embrechts, P. ; C. Klüppelberg et T. Mikosch. 1997. *Modelling Extremal Events for Insurance and Finance*. New York : Springer.

Fiorentino, M. et S. Gabriele. 1984. "A Correction for the Bias of Maximum-Likelihood Estimators of Gumbel Parameters." *Journal of Hydrology*, 73(1-2), pp. 39-49.

Fisher, R.A. et L.H.C. Tippett. 1928. "Limiting Forms of the Frequency Distribution of the Largest and the Smallest Member of a Sample." *Proc. Cambridge Philos. Soc.*, 24, pp. 180-90.

Fortheringham, A. S. ; M. Charlton et C. Brunsdon. 1996. "The Geography of Parameter Space : An Investigation into Spatial Non-Stationarity." *International Journal of Geographical Information Systems*, 10, pp. 605-27.

Gallant, A. R. 1981. "On the Bias in Flexible Functional Forms and an Essentially Unbiased Form : The Fourier Flexible Form," *Journal of Econometrics*, 15, pp. 211-245.

Golan, A. et H. Shalit. 1993. "Wine Quality Differentials in Hedonic Grape Pricing." *Journal of Agricultural Economics*, 44, pp. 311-21.

Gorman, W.M. . 1980. "The Demand for Related Goods : A Possible Procedure for Analysing Quality Differentials in the Egg Market." *Review of Economic Studies*, 47, pp. 843-56.

Griliches, Z. 1961. "Hedonic Price Indexes for Automobiles : An Econometric Analysis of Quality Change," in *The Price Statistics of the Federal Government*, G.J. Stigler (Chairman), New York : NBER.

Gumbel, E. J. 1958. *Statistics of Extremes*. New York, : Columbia University Press.

Haberman, S. et A. E. Renshaw. 1996. "Generalized Linear Models and Actuarial Science." *Statistician*, 45(4), pp. 407-36.

Hill, B.M. 1975. "A Simple General Approach to Inference About the Tail of a Distribution." *Annal of Statistics*, 3, pp. 1163-74.

Hogg, R. V. et S. A. Klugman. 1983. "On the Estimation of Long Tailed Skewed Distributions with Actuarial Applications." *Journal of Econometrics*, 23(1), pp. 91-102.

Hosking, J. R. M.; J. R. Wallis et E. F. Wood. 1985. "Estimation of the Generalized Extreme-Value Distribution by the Method of Probability-Weighted Moments." *Technometrics*, 27(3), pp. 251-61.

Houthakker, H.S. 1952. "Compensated Change in Quantities and Qualities Consumed." *Review of Economic Studies*, 19, pp. 155-64.

Jenkinson, A.F. 1955. "The Frequency Distribution of the Annual Maximum (or Minimum) of Meteorological Elements." *Quarterly Journal of Royal Meteorological society*, 81, pp. 158-71.

Kuczera, G. 1982. "Combining Site-Specific and Regional Information - an Empirical Bayes Approach." *Water Resources Research*, 18(2), pp. 306-14.

Lancaster, K. 1971. *Consumer Demand : A New Approach*. New York : Columbia University Press.

Landon, S. et C.E. Smith. 1997. "The Use of Quality and Reputation Indicators by Consumers : The Case of Bordeaux Wine." *Journal of Consumer Policy*, 20, pp. 289-323.

Mason, C. et J. M. Quigley. 1996. "Non-Parametric Hedonic Housing Prices." *Housing Studies*, 11(3), pp. 373-85.

McNeil, A.J. 1997. "Estimating the Tails of Loss Severity Distributions Using Extreme Value Theory." *Astin Bulletin*, 27, pp. 117-37.

Nelson, J.P. 1978. "Residential Choice, Hedonic Prices and the Demand for Urban Air Quality." *Journal of Urban Economics*, 5, pp. 357-69.

Nerlove, M. 1995. "Hedonic Price Functions and the Measurement of Preferences : The Case of Swedish Wine Consumers." *European Economic Review*, 39(9), pp. 1697-716.

Northrop, P. J. 2004. "Likelihood-Based Approaches to Flood Frequency Estimation." *Journal of Hydrology*, 292, pp. 96-113.

Oczkowski, E. 1994. "A Hedonic Price Function for Australian Premium Table Wine." *Australian Journal of Agricultural Economics*, 38(1), pp. 93-110.

Oczkowski, E. 2001. "Hedonic Wine Price Functions and Measurement Error." *Economic Record*, 77(239), pp. 374-82.

Pagan, A. 1984. "Econometric Issues in the Analysis of Regressions with Generated Regressors." *International Economic Review*, 25(1), pp. 221-47.

Parker, R. M. 2003. *Bordeaux : A Consumer's Guide to the World's Finest Wines*. New York : Simon and Schuster.

Prescott, P. et A. T. Walden. 1980. "Maximum Likelihood Estimation of the Parameters of the Generalized Extreme-Value Distribution." *Biometrika*, 67(3), pp. 723-24.

Reiss, R.-D. et M. Thomas. 1997. *Statistical Analysis of Extreme Values from Insurance, Finance, Hydrology and Other Fields*. Verlag.

Renshaw, A.E. 1994. "Modelling the Claims Process in the Presence of Covariates." *Astin Bulletin*, 24(2), pp. 265-85.

Rosen, S. 1974. "Hedonic Prices and Implicit Markets : Product Differentiation in Pure Competition." *Journal of Political Economy*, 82, pp. 34-55.

Schamel, G. 2000. "Individual and Collective Reputation Indicators of Wine Quality," In. Centre for International Economic Studies, Adelaide

University.

Schamel, G. et K. Anderson. 2003. "Wine Quality and Varietal, Regional and Winery Reputations : Hedonic Prices for Australia and New Zealand." *Economic Record*, 79(246), pp. 357-69.

Shao, Q. X.; H. Wong; J. Xia et W. C. Ip. 2004. "Models for Extremes Using the Extended Three-Parameter Burr Xii System with Application to Flood Frequency Analysis." *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques*, 49(4), pp. 685-702.

Shapiro, C. 1983. "Premiums for High Quality Products as Returns to Reputation." *The Quarterly Journal of Economics*, 98(4), pp. 659-79.

Simmons, P. et D. Weiserbs. 1979. "Translog Flexible Functional Forms and Associated Demand Systems." *The American Economic Review*, 69(5), pp. 892-902.

Simons, R.A.; W.M. Bowen et A. Sementelli. 1999. "The Price and Liquidity Effects of Ust Leaks from Gas Stations on Adjacent Contaminated Property." *The Appraisal Journal*, (April).

Smith, R. L. 1985. "Maximum-Likelihood Estimation in a Class of Nonregular Cases." *Biometrika*, 72(1), pp. 67-90.

Stanley, L. R. et J. Tschirhart. 1991. "Hedonic Prices for a Nondurable Good - the Case of Breakfast Cereals." *Review of Economics and Statistics*, 73(3), pp. 537-41.

Triplett, J. E. . 1989. "Price and Technological Change in a Capital Good : A Survey of Research on Computers," In *Technology and Capital Formation*, ed. D. W. JORGENSON, LANDAU, R. Cambridge : MIT Press.

Vanderlaan, B. S. et A. S. Louter. 1986. "A Statistical-Model for the Costs of Passenger Car Traffic Accidents." *Statistician*, 35(2), pp. 163-74.

Weibull, G. W. ed. *A Statistical Theory of Strength of Materials*. Stockolm : Handl, 1939.

Witte, A.D. ; H.J. Sumka et H. Erekson. 1979. "An Estimate of a Structural Hedonic Price Model of the Housing Market : An Application of Rosen's Theory of Implicit Markets." *Econometrica*, 47(5), pp. 1151-73.