REPUBLIQUE DU SENEGAL

un peuple, un but, une foi

Université Gaston Berger de Saint-Louis

UFR de Sciences Appliquées et de Technologie

MÉMOIRE DE MASTER

pour l'obtention du diplôme de

Master en Mathématiques Appliquées

Soutenu publiquement le 22 Novembre 2024 par

Abdou-Latifou KEDJERI

Mention : Mathématiques Appliquées

Spécialité : Science de Données et Applications

ESTIMATION DE L'INDICE DES VALEURS EXTRÊMES SOUS TRONCATURE

Encadreur : Prof Aliou DIOP

Composition du Jury

Pr A. Diakhaby	Président du Jury
Pr Aliou Diop	Encadreur
Pr A. Ka Diongue	Examinateur
Pr Mamadou Abdoulaye Konte	Examinateur

Dédicaces

À la mémoire de ma Mère Feue Roukeyatou SAÏBOU,
À mon Père Zibérou KEDJERI TCHAGOLE,
À toute la famille KEDJERI et TCHAGOLE
À toutes mes Sœurs, mes Frères et à tous mes amis de classe que je n'oublierai jamais

Résumé

L'estimation de l'indice des valeurs extrêmes (EVI) est cruciale pour analyser les phénomènes rares, tels que les événements extrêmes dans les domaines de la finance, de l'hydrologie, ou des catastrophes naturelles. Cependant, lorsque les données sont tronquées, les estimateurs classiques peuvent perdre de leur efficacité. Ce mémoire se concentre sur l'adaptation des principaux estimateurs de l'EVI, notamment les estimateurs de Hill, Pickands et de Dekkers-Einmahl-de Haan (DEdH), dans un contexte de troncature.

Après une présentation théorique des estimateurs, nous analysons les biais introduits et les MSE (erreur quadratique moyenne) ou (Mean squared Error en anglais) par la troncature à gauche et proposons des corrections spécifiques. Des simulations de données, ainsi que des applications sur des jeux de données réelles, permettent d'évaluer la performance des estimateurs dans le scénario de troncature à gauche.

Les résultats montrent que, malgré la simplicité de ces méthodes, des ajustements appropriés permettent d'obtenir des estimations plus robustes et précises dans un contexte tronqué. Enfin, ce travail ouvre des perspectives sur l'extension de ces méthodes à des environnements multivariés ou temporels.

Mots-clés : Indice des valeurs extrêmes, troncature, estimateurs de Hill, Pickands, DEdH, phénomènes rares.

Abstract

Estimating the Extreme Value Index (EVI) is crucial for analyzing rare events, such as extreme occurrences in finance, hydrology, or natural disasters. However, when data are truncated, classical estimators may lose their effectiveness. This thesis focuses on adapting the main EVI estimators, including the Hill, Pickands, and Dekkers-Einmahl-de Haan (DEdH) estimators, in the context of truncation.

After a theoretical presentation of the estimators, we analyze the biases introduced and the Mean Squared Error (MSE) by left truncation and propose specific corrections. Data simulations, as well as applications to real datasets, are used to evaluate the performance of the estimators in the left truncation scenario.

The results show that, despite the simplicity of these methods, appropriate adjustments allow for more robust and accurate estimates in a truncated context. Finally, this work opens up prospects for extending these methods to multivariate or temporal environments.

Keywords : Extreme Value Index, truncation, Hill estimator, Pickands estimator, DEdH estimator, rare events.

Table des matières

Table des figures

Liste des tableaux

Notations et abréviations

$v.a$	Variable aléatoire.
fdr	Fonction de distribution cumulative.
F_n	Fonction de distribution empirique.
$F^{\leftarrow}$	Inverse généralisée de F.
TCL	Théorème Central Limite.
EVD	Loi des valeurs extrêmes.
EVI, ξ	Indice des valeurs extrêmes.
GEV	Loi généralisée des valeurs extrêmes.
POT	Dépassements au-delà du seuil.
GPD	Distribution de Pareto généralisée.
$\mathcal{H}_\xi$	Famille des distributions généralisées des valeurs extrêmes.
$\mathbb{P}(A)$	Probabilité de l'événement A.
$\mathbb{E}(X)$	Espérance de la variable aléatoire X.
$i, i.d$	Variables indépendantes et identiquement distribuées.
$X_{1:n}, ..., X_{n,n}$	Statistique d'ordre associée à $X_1, ..., X_n$.
x_F	Point terminal.
$\overset{\text{loi}}{=}$	Égalité en loi.
$:=$	Égalité en définition.
$\xrightarrow{\mathcal{L}}$	Converge en loi.
$\xrightarrow{P}$	Converge en probabilité.
$\xrightarrow{p.s.}$	Converge presque sûrement.
$\mathcal{VR}_\alpha$	Variation régulière d'indice α.
$\sup\{A\}$	Suprémum de l'ensemble A.
$\hat{\xi}^H$	Estimateur de Hill.
$\hat{\xi}^P$	Estimateur de Pickands.
$\hat{\xi}^{DEdH}$	Estimateur de Dekkers-Einmahl-de Haan.
IC	Intervalle de confiance.
MSE	Erreur quadratique moyenne.
$\mathbb{I}_{\{A\}}$	Fonction indicatrice de l'ensemble A.
$L(x)$	Fonction à variation lente.
DA	Domaine d'attraction du maximum.
$M_n = X_{n,n}$	Maximum de $X_1, \ldots, X_n$.
N_u	Nombre d'excès dépassant le seuil u.
$p.s$	Presque sûrement.
al	Autres.
Λ	Loi de Gumbel.
Φ_α	Loi de Fréchet.
Ψ_α	Loi de Weibull.
resp	Respectivement.

1 INTRODUCTION GÉNÉRALE

Généralité

L'étude des valeurs extrêmes est essentielle dans de nombreux domaines, tels que la finance, l'hydrologie, la météorologie et les sciences de l'environnement. L'estimation de l'indice des valeurs extrêmes (EVI) est une composante clé de cette analyse, permettant de caractériser la queue de distribution des données et d'évaluer le risque d'événements rares mais potentiellement dévastateurs.

Les événements rares et catastrophiques, tels que les tremblements de terre, les inondations, les accidents nucléaires, les crises financières, les krachs boursiers et l'apparition de nouveaux phénomènes endémiques, attirent l'attention par leur imprévisibilité. Étant donné l'ampleur des enjeux sociaux et scientifiques qu'ils représentent, il est essentiel de les prendre en considération dans toute discussion sérieuse sur le hasard. Lorsque ces événements sont aléatoires, il est possible d'analyser leur distribution. Ils sont qualifiés d'extrêmes lorsque leurs valeurs sont significativement plus grandes ou plus petites que celles généralement observées.

La théorie statistique traditionnelle, notamment le célèbre théorème de la limite centrale, permet de réaliser des inférences sur les valeurs centrales d'un échantillon. Cependant, elle fournit peu d'informations sur la queue de distribution. La spécificité de la théorie des valeurs extrêmes réside dans son intérêt pour la queue de distribution, qui est à l'origine des phénomènes extrêmes étudiés. Cette théorie est développée pour estimer la probabilité d'événements rares, offrant des estimations fiables des valeurs extrêmes, qui sont peu fréquentes. Le siècle dernier et le début de celui-ci ont été marqués par de nombreux événements extrêmes, et c'est dans ce contexte que la théorie des valeurs extrêmes, développée par Fisher et Tippet en 1928 a pris toute son importance dans plusieurs domaines :

1. Hydrologie : Crues causées par des pluies torrentielles, comme au Togo (Inondations de 2019 endommageant les infrastructures urbaines et perturbant les communautés locales) et au Sénégal (inondations de 1994 et 2003 entraînant l'ouverture de les Brèches importantes dans les infrastructures de protection, entraînant des inondations massives.)
2. Climatologie : Étude des événements climatiques extrêmes tels que les précipitations, les températures, et les chutes de neige, ainsi que la modélisation des grands feux de forêt (voir par exemple Alvarado et al.(1998) [9]).
3. Assurance : Survenue de sinistres d'intensité exceptionnelle, tels que l'ouragan Katrina en 2005, les importants incendies dans les risques industriels, et les sinistres graves en responsabilité civile automobile, qui peuvent affecter négativement la solvabilité des compagnies d'assurance.
4. Finance : Forte volatilité des actifs financiers et gestion des risques opérationnels des banques.

Parmi les pionniers de cette théorie, on peut citer Gnedenko(1943) [5] , Leadbetter(1983)[13],

Resnick(1987) [8], Embrechts et al.(1997) [7], et de Haan et Ferreira (2006) [6].

Cependant, dans de nombreuses applications pratiques, les données peuvent être tronquées, c'est-à-dire que certaines observations ne sont pas disponibles en raison de limitations inhérentes au processus de collecte des données ou à des choix méthodologiques. La troncature pose des défis particuliers à l'estimation de l'indice des valeurs extrêmes, car elle peut biaiser les estimations et réduire leur efficacité.

Ce mémoire se concentre sur l'estimation de l'indice des valeurs extrêmes sous troncature. L'objectif est de développer et d'analyser des méthodes qui prennent en compte cette troncature, afin d'améliorer la précision et la fiabilité des estimations. Les méthodes classiques d'estimation, telles que les estimateurs de Hill, de Pickands et de l'estimateur de Dekkers-Einmahl-de Haan (DEdH), une extension de l'estimateur de Hill classique, seront adaptées et évaluées dans des contextes de données tronquées. En outre, de nouvelles approches seront explorées pour surmonter les défis spécifiques posés par la troncature. L'étude mettra en œuvre des simulations numériques pour comparer la performance des différentes méthodes, et elle utilisera des ensembles de données réels pour illustrer les applications pratiques.

L'intérêt de ce travail réside non seulement dans l'amélioration des techniques d'estimation de l'EVI, mais aussi dans son application potentielle à divers domaines où la compréhension des valeurs extrêmes est cruciale. En fournissant des outils plus robustes pour l'analyse des valeurs extrêmes sous troncature, ce mémoire contribue à une meilleure gestion des risques associés aux événements extrêmes dans divers contextes.

Notre memoire est structuré comme suit :

Le Chapitre 1 est une Introduction Générale de notre travail, le Chapitre 2 présente les concepts et résultats de base nécessaires pour la suite de ce travail. Nous y abordons les distributions asymptotiques en théorie des valeurs extrêmes, en mettant l'accent sur le paramètre réel inconnu ξ, connu sous le nom d'indice de queue ou d'indice des valeurs extrêmes. La connaissance de ce paramètre est cruciale en théorie des valeurs extrêmes, car il détermine le comportement de la queue de la distribution. Plus ξ est élevé, plus la queue est lourde, ce qui rend son estimation particulièrement importante.

Dans le chapitre 3, nous nous intéressons à l'étude de la troncature en examinant les différents types de troncature, des défits supplémentaires qu'elle introduit dans l'estimation de l'EVI.

Dans chapitre 4, nous rappelons les principales méthodes d'estimation du paramètre ξ (EVI) ainsi que les procédures de sélection du nombre de valeurs extrêmes à utiliser pour cette estimation. Nous aborderons ensuite l'estimation des quantiles extrêmes, en mettant en lumière les techniques et résultats clés qui seront essentiels pour le chapitre suivant. Chaque étape est accompagnée des informations les plus pertinentes pour garantir une compréhension approfondie et une application efficace dans le cadre de ce travail.

Dans le chapitre 5 dédié à l'illustration numérique, nous passerons en revue les simulations et l'application des méthodes aux données réelles. Le but est de mettre en lumière l'impact des différentes techniques d'estimation sur les résultats obtenus à partir de données simulées et réelles, en lien avec l'indice des valeurs extrêmes ξ. Nous détaillerons les procédures de simulation de données, en choisissant différentes distributions (comme Pareto ou Fréchet) pour générer des queues lourdes.

2 Généralités sur les Extrêmes

Contents

1 Introduction

La théorie des valeurs extrêmes (TVE) également connue sous le nom d'analyse des valeurs extrêmes ou "Extreme Value Theory" (EVT) en anglais, est une discipline étendue dont l'objectif est d'étudier les événements rares, c'est-à-dire ceux qui ont une faible probabilité de survenir. En d'autres termes, cette théorie offre des outils pour comprendre et prévoir des phénomènes tels que les inondations, les intempéries, les crises financières, les catastrophes naturelles, etc. L'étude des valeurs extrêmes consiste à analyser les queues des distributions de fonctions ou à examiner les plus grandes ou les plus petites observations d'un échantillon donné. Ainsi, la TVE peut être considérée comme un complément à la théorie statistique classique, qui se concentre principalement sur l'étude de la moyenne d'un échantillon plutôt que sur ses valeurs extrêmes. Parmi les ouvrages de référence classiques sur la théorie et les applications des valeurs extrêmes, on peut citer ceux de Coles(2001)[19] , Embrechts(1997)[7], Reiss(2007)[18] et Beirlant et al(2004)[21] qui résument les différentes techniques disponibles.

2 Concepts de la Théorie des Valeurs Extrêmes (TVE)

La théorie des valeurs extrêmes se propose d'analyser la distribution du maximum d'une série de variables aléatoires réelles, en particulier lorsque la loi sous-jacente du phénomène est inconnue. Formellement, supposons que $X_1, X_2, ..., X_n$ soit une série de n variables aléatoires indépendantes et identiquement distribuées $(i.i.d)$ avec une fonction de répartition F donnée par

$$F(x) = Pr(Xi \leq x) \quad pour \quad i = 1, ..., n. \tag{2.1}$$

L'un des aspects fondamentaux de la TVE est l'étude des maxima et minima d'échantillons aléatoires.

3 Distribution des maxima et des minima

3.1 Statistique d'ordre

Dans la théorie des valeurs extrêmes (TVE), les statistiques d'ordre jouent un rôle crucial dans l'analyse des événements rares. Les statistiques d'ordre sont des variables aléatoires qui représentent les rangs des observations dans un échantillon, ordonnées du plus petit au plus grand. Elles fournissent des informations essentielles sur les valeurs extrêmes et permettent de caractériser la distribution des extrêmes.

Définition

Soit $X_1, X_2, ..., X_n$ une séquence de variables aléatoires $va's$ indépendantes et identiquement distribuées (iid) avec fonction commune de distribution F et de densité f. Considerons les $va's$ $X_{1:n}, ..., X_{n:n}$ rangées par ordre croissant. Soit

$$X_{1:n} \leq, \ldots, \leq X_{n:n} \tag{2.2}$$

On appelle les Statistiques d'ordre de l'échantillon $X_1, X_2, ..., X_n$ les $va's$ de l'équation 2.2

Remarque 3.1 *Pour $1 \leq k \leq n$, la variable $X_{k:n}$ est connue sous le nom de la k^{eme} statistiques d'ordre ou statistique d'ordre k.*

Pour analyser le comportement des événements extrêmes. C'est-à-dire les maxima et mes minima.

Le minimum de l'échantillon (statistique d'ordre minimal) est défini par

$$X_{1:n} = min(X_1, ..., X_n)$$

Les statistiques d'ordre intermédiaires : Ce sont les observations situées entre le minimum et le maximum de l'échantillon. Elles sont défini par

$$X_2, ..., X_{n-1}$$

NB : On obtient la correspondance entre minimum et maximum par la relation suivante :

$$min(X_1, ..., X_n) = -max(-X_1, ..., -X_n)$$

Le maximum de l'échantillon, également appelé statistique d'ordre maximal, est défini par $M_n = max(X_1, X_2, ..., X_n)$, représentant le maximum d'un échantillon de taille n. Étant donné que les variables aléatoires sont indépendantes et identiquement distribuées (i.i.d.), la fonction de répartition de M_n est donnée par

$$\begin{aligned} F_{M_n}(x) &= Pr(M_n \leq x) \\ &= P(X_1 < x, \ldots, X_n < x) \\ &= P(X_1 < x) \ldots P(X_n < x) \\ &= [F(x)]^n \end{aligned}$$

La formule ci-dessus présente un intérêt limité, car il est peu fréquent que la distribution exacte de la variable aléatoire sous-jacente X soit parfaitement connue. Même si cette distribution est

bien définie, déterminer la loi du maximum reste souvent une tâche complexe. C'est pourquoi il devient pertinent d'étudier les comportements asymptotiques du maximum M_n après une normalisation appropriée.

L'un des résultats essentiels de la théorie des valeurs extrêmes a été établi en 1928 par Fisher et Tippett.

Définition 3.1 *(lois de même type) : On dit que deux variables aléatoires réelles X et Y sont de même type s'il existe des constantes réelles $a > 0$ et $b \in \mathbb{R}$ telles que Y ait la même loi que $Y \overset{loi}{=} aX + b$. En d'autres termes, si F et H sont les distributions respectives de X et Y, alors $F(ax + b) = H(x)$. Ainsi, les variables de même type partagent la même loi à un facteur de localisation et d'échelle près.*

Soit x_F le point terminal à droite (right-end point) de la fonction de répartition F_X défini par

$$x_F = \sup\{x \in \mathbb{R}, F(x) < 1\}.$$

Ce point peut être infini ou fini Embrechts et al.(1997)[7], (exemple 3.3.22, p.139).

Nous nous intéressons alors à la distribution asymptotique du maximum en faisant tendre n vers l'infini. Nous avons :

$$\lim_{n\to\infty} F_{M_n}(x) = \lim_{n\to\infty} [F_X(x)]^n = \begin{cases} 1 & \text{si } x \geq x_F \\ 0 & \text{si } x < x_F \end{cases} \tag{2.3}$$

On observe que la distribution asymptotique du maximum conduit à une loi dégénérée, représentée par une masse de Dirac en x_F, car pour certaines valeurs de x, la probabilité peut être égale à 1 si x_F est fini et M_n tend vers x_F presque sûrement avec $x_F \leq \infty$. Ce résultat est très peu informatif sur le comportement du maximum, et il est préférable d'obtenir une loi non-dégénérée comme limite de l'équation 2.3. Pour ce faire, l'idée est de considérer non pas le maximum en tant que tel dans l'équation 2.2, mais une renormalisation de ce dernier. La renormalisation la plus simple à laquelle on peut penser consiste en une transformation linéaire, à l'image de celle utilisée dans le Théorème Central Limite (TCL). Nous en rappelons l'énoncé ci-dessous :

Théorème 3.1 *(Théorème Central Limite) Soit $X_1, X_2, \ldots, X_n$ une suite de variables aléatoires indépendantes et identiquement distribuées de fonction de répartition F. Supposons que l'espérance μ et l'écart-type σ de F existent et soient finis avec $\sigma \neq 0$.*
Considérons la somme $S_n = X_1 + X_2 + \ldots + X_n$. Alors

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{loi} \mathcal{N}(0,1), \quad n \to +\infty. \tag{2.4}$$

Le TCL caractérise donc le comportement de la somme linéairement renormalisée de n variables aléatoires iid, en indiquant la loi Normale comme loi limite. La question est de savoir s'il existe un équivalent du TCL non pas pour la somme, mais pour le maximum de n variables aléatoires iid. En 1928, Fisher et Tippett résolvent ce problème avec un théorème portant leur nom, constituant un des piliers de la théorie des valeurs extrêmes.

Théorème 3.2 *Théorème de Fisher-Tippett ou théorème des 3 types extrêmes*
Considérons une suite de n variables aléatoires réelles iid, ayant une loi continue P, notée $X_1, X_2, \ldots, X_n$. *On définit le maximum de cette suite par :*

$$M_n = \max(X_1, X_2, \ldots, X_n)$$

Si deux suites de constantes de normalisation $a_n > 0$ *et* $b \in \mathbb{R}$ *existent, et qu'il existe une loi non-dégénérée H telle que :*

$$\lim_{n\to\infty} \mathbb{P}\left[\frac{M_n - b_n}{a_n} \leq x\right] = \lim_{n\to\infty} [F(a_n x + b_n)]^n \to H(x) \quad , x \in \mathbb{R} \tag{2.5}$$

alors H appartient à l'une des trois lois limites suivantes :

- ***Distribution de Gumbel (pour les variables à queue légère) :***

$$\Lambda(x) = \exp(-\exp^{-x}), \quad x \in \mathbb{R} \tag{2.6}$$

- ***Distribution de Fréchet (pour les variables à queue lourde) :***

$$\Phi_\alpha(x) = \begin{cases} 0 & si\ x \leq 0 \\ \exp(-x^{-\alpha}) & si\ x > 0 \end{cases} \tag{2.7}$$

- ***Distribution de Weibull (pour les variables à queue bornée) :***

$$\Psi_\alpha(x) = \begin{cases} \exp\{-(-x)^\alpha\} & si\ x \leq 0 \\ 1 & si\ x > 0 \end{cases} \tag{2.8}$$

On dit alors que X (ou F_X*) appartient au domaine d'attraction maximum de H ou au max-domaine d'attraction, noté* $(X \in MDA(H))$.

Une démonstration détaillée de ce théorème est fournie dans Resnick(1987)[8], avec des développements supplémentaires dans Embrechts(1997)[7]. Les suites a_n et b_n dépendent des paramètres de la loi de X.

Exemple 1 *(a)* ***Distribution exponentielle*** *:* $F(x) = 1 - \exp(-\lambda x)$, $x \geq 0$, $\lambda > 0$, *appartient à* $MDA(\Lambda)$ *avec* $a_n = \frac{1}{\lambda}$, $d_n = \frac{\ln n}{\lambda}$.

(b) ***Distribution de Pareto*** *:* $F(x) = 1 - \left(\frac{\sigma}{\sigma+x}\right)^\alpha$, $x \geq 0$, $\sigma, \alpha > 0$, *appartient à* $MDA(\Phi_\alpha)$ *avec* $a_n = \left(\frac{n}{\sigma}\right)^{\frac{1}{\beta}}$, $d_n = 0$.

(c) ***Distribution uniforme*** *:* $F(x) = x$, $x \in (0,1)$, *appartient à* $MDA(\Psi_1)$ *avec* $a_n = \frac{1}{n}$, $d_n = 0$.

Le théorème 3.2 présente un résultat fascinant : indépendamment de la distribution de la variable initiale, la distribution des valeurs extrêmes a toujours une forme identique. Bien que le comportement de ces distributions soit très différent, elles peuvent être réunies au sein d'une même paramétrisation. Un seul paramètre ξ, appelé l'indice des valeurs extrêmes (ou indice de queue), permet de contrôler la "lourdeur" de la queue de la distribution :

Définition 3.2 *Représentation de Jenkinson-von Mises*

$$H_\xi(x) = \begin{cases} \exp\left(-(1+\xi x)^{-1/\xi}\right) & ,\xi \neq 0, \quad 1+\xi x > 0 \\ \exp(-\exp(-x)) & ,\xi = 0, \quad -\infty \leq x \leq \infty \end{cases} \tag{2.9}$$

où H est une fonction de répartition non dégénérée.

La distribution GEV (Generalized Extreme Value) représente les trois types de distributions extrêmes :

- $\xi > 0$ *:* ***Fréchet****, avec la fonction quantile* $Q_\xi\left(\frac{x-1}{\xi}\right) = \Phi_{1/\xi}(x)$,
- $\xi = 0$ *:* ***Gumbel****, avec la fonction quantile* $Q_0(x) = \Lambda(x)$,
- $\xi < 0$ *:* ***Weibull****, avec la fonction quantile* $Q_\xi\left(-\frac{x+1}{\xi}\right) = \Psi_{-\frac{1}{\xi}}(x)$.

Cette distribution est connue sous le nom de loi des valeurs extrêmes généralisée (Generalized Extreme Value, GEV).

Nous introduisons les paramètres de position $\mu \in \mathbb{R}$ et d'échelle $\sigma > 0$, puis définissons $Q_{\xi,\mu,\sigma}(x) = Q_\xi\left(\frac{x-\mu}{\sigma}\right)$. Il est important de noter que $Q_{\xi,\mu,\sigma}$ appartient à la même famille que Q_ξ.

Cette formulation est particulièrement utile pour les méthodes statistiques qui s'appuient sur des maxima (*iid*) (indépendants et identiquement distribués). Dans ce cas, ces maxima sont modélisés par la loi GEV, permettant d'ajuster les paramètres afin d'estimer les quantiles et les queues de distribution ; voir Embrechts et al.(1997)[7], section 6.3

En introduisant les paramètres de localisation μ et de dispersion σ dans l'équation 2.9, on obtient la forme la plus générale de la GEV :

$$H_{\xi,\mu,\sigma}(x) = \exp\left(-(1+\xi\frac{x-\mu}{\sigma})^{-1/\xi}\right), \quad \xi \neq 0, \quad 1+\xi\frac{x-\mu}{\sigma} > 0 \tag{2.10}$$

Le théorème de Fisher-Tippett est, en quelque sorte, l'analogue du Théorème Central Limite (TCL) pour les événements extrêmes. Cependant, contrairement au TCL où la distribution normale est la seule distribution limite possible, il existe trois types de distributions limites possibles pour les extrêmes (voir Figure 2.1).

Selon le signe de ξ, on définit trois domaines d'attraction :

- Le cas $\xi = 0$ correspond à la loi de Gumbel, qui est utilisée pour modéliser des valeurs extrêmes lorsque la queue de la distribution décroît exponentiellement.
- Le cas $\xi > 0$ correspond à la loi de Fréchet, qui s'applique aux distributions à queue lourde, caractérisées par une décroissance plus lente des probabilités des événements extrêmes. Le paramètre associé est $\alpha = \frac{1}{\xi}$.
- Le cas $\xi < 0$ représente la loi de Weibull, utilisée pour modéliser des distributions à queue bornée. Cela signifie qu'il existe une limite supérieure aux valeurs extrêmes possibles, avec un paramètre $\alpha = \frac{-1}{\xi}$.

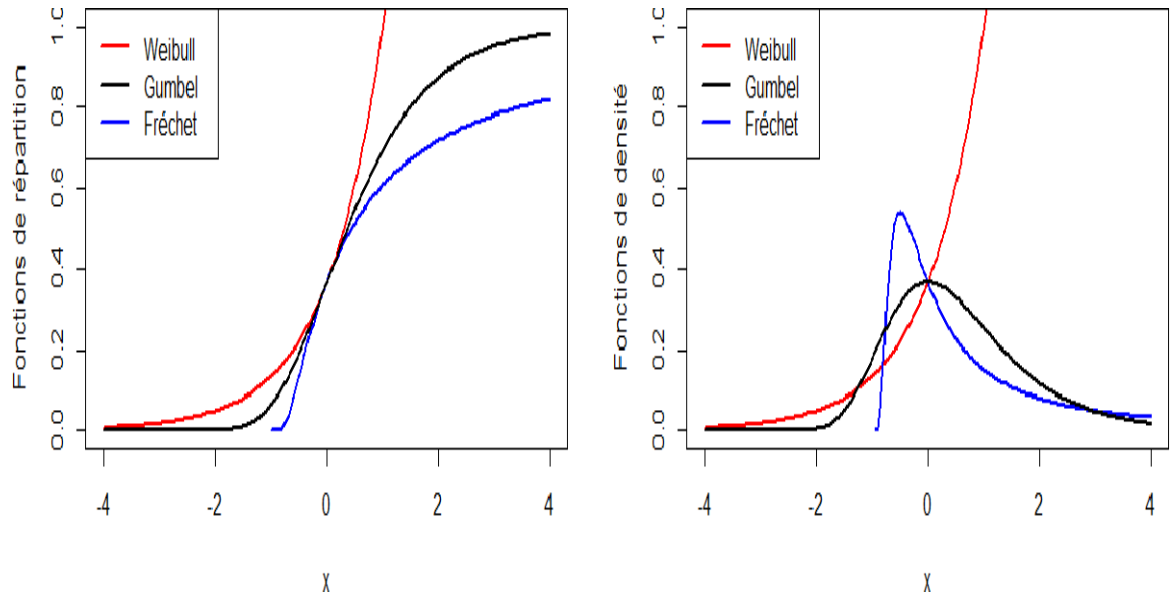

FIGURE 2.1 – Répartition et densité des lois de valeurs extrêmes : $\xi > 1$ pour la loi de Fréchet, $\xi = 0$ pour la loi de Gumbel, et $\xi < 1$ pour la loi de Weibull.

Pour un classement de différentes lois par domaine d'attraction, nous indiquons Embrechts et al.(1997)[7]

Voici ci-dessous un classement de quelques lois par domaine d'attraction dans le tableau 2.1

Domaines d'attraction	**Paramètre ξ**	**Lois associées**
Gumbel	$\xi = 0$	Normale Exponentielle Log normale Gumbel Weibull
Fréchet	$\xi > 0$	Cauchy Student Pareto Généralisée Log-gamma
Weibull	$\xi < 0$	Uniforme Beta Reverse Burr

TABLE 2.1 – Domaines d'attraction et quelques lois associées

La distribution la plus couramment observée dans les données financières ou macroéconomiques est celle de Fréchet. La distribution de Weibull est souvent exclue car elle ne peut pas modéliser des variations rares, alors que les variations en finance ou en économie sont généralement non bornées, permettant la survenue de variations extrêmes. La distribution de Gumbel est également très rare. Cependant, elle est utilisée comme référence pour mesurer, notamment graphiquement, l'écart entre les distributions empiriques et la distribution normale.

3.2 Caractérisation des Domaines d'Attraction

Un enjeu majeur est de déterminer les conditions nécessaires et suffisantes pour qu'une distribution appartienne à un domaine d'attraction. Cette recherche implique d'analyser la distribution des valeurs extrêmes associées à une distribution donnée. La question est : étant donné une loi H de type extrême (Fréchet, Gumbel ou Weibull), quels critères doivent être remplis pour que la loi du maximum d'une suite de variables aléatoires i.i.d. de loi F converge vers H ? Diverses caractérisations des domaines d'attraction de Fréchet, Gumbel et Weibull ont été proposées par Resnick(1987)[8], Embrechts et al.(1997)[7], et de Haan(2006)[6], utilisant les classes de fonctions à variation régulière. L'inverse généralisée d'une fonction monotone U est notée $U^{\leftarrow}$ et définie par

$$U^{\leftarrow}(s) = \inf\{x \in \mathbb{R}; U(x) \geq s\} \quad \text{, pour} \quad 0 < s \leq 1.$$

Dans ce qui suit, nous allons présenter quelques théorèmes de caractérisation des trois domaines d'attraction.

3.3 Fonction à variations régulières

Le concept de variation régulière est largement appliqué dans la théorie des valeurs extrêmes. Dans cette partie nous présenterons que quelques résultats clés (définitions, extensions et propriétés) de la théorie de la variation régulière pertinents pour notre étude. Pour plus de détails, on peut consulter Bingham et al.(1987)[37].

Définition 3.3 *Une fonction $U(\cdot) : \mathbb{R}_+ \to \mathbb{R}_+$ mésurable est dite à variations régulières à l'infini d'indice $\rho \in \mathbb{R}$, notée $U(\cdot) \in \mathcal{RV}_\rho$, si U est positive à l'infini (c'est-à-dire s'il existe une constante A telle que pour tout $x > A$, $U(x) > 0$) et si pour tout $\lambda > 0$, on a :*

$$\lim_{\lambda \to \infty} \frac{U(\lambda x)}{U(\lambda)} = x^\rho$$

On appelle ρ l'indice de la fonction à variation régulière.

Remarque 3.2
— *Si $\rho = 0$, c'est-à-dire $U(\cdot) \in \mathcal{RV}_0$, alors la fonction $U(\cdot)$ est appelée fonction à variations lentes à l'infini et est généralement notée $\ell(\cdot)$. C'est-à-dire on a*

$$\lim_{x \to \infty} \frac{U(\lambda x)}{U(\lambda)} = 1$$

avec $l(x) = 1$
— *Si $\rho = \infty$, on parle de fonction à variations rapides à l'infini.*

Conséquence 3.1 *Une fonction mesurable $U : \mathbb{R}_+ \to \mathbb{R}_+$ est dite à variation régulière d'indice ρ près de 0 (notée $U \in \mathcal{RV}_\rho^0$) si, pour tout $x > 0$,*

$$\lim_{\lambda \to 0} \frac{U(\lambda x)}{U(\lambda)} = x^\rho.$$

Cela signifie que $U(1/x)$ est à variation régulière d'indice $(-\rho)$ à l'infini.

Lemme 3.1 *Inverse d'une fonction à variation régulière*
— *Si U est à variation régulière d'indice $\rho > 0$ à l'infini, alors $U^{\leftarrow}(x)$ est à variation régulière d'indice $1/\rho > 0$ à l'infini.*
— *Si U est à variation régulière d'indice $\rho < 0$ à l'infini, alors $U^{\leftarrow}(1/x)$ est à variation régulière d'indice $-1/\rho > 0$ à l'infini.*

Pour la preuve de ce lemme, on peut consulter Bingham et al.(1987)[37], Théorème 1.5.12, ou Resnick(1987)[8]. Il est évident que si ℓ est une fonction à variation lente et $\rho \in \mathbb{R}$, alors la fonction $U(x) = x^\rho \ell(x)$ pour tout $x > 0$ appartient à $\mathcal{RV}^\rho$.

Proposition 3.1 *Soient $\rho \in \mathbb{R}$ et $U \in \mathcal{RV}^\rho$. Il existe alors une fonction à variation lente ℓ à l'infini telle que :*

$$\forall x > 0, U(x) = x^\rho \ell(x). \tag{2.11}$$

Ce résultat indique que l'étude des fonctions à variation régulière à l'infini peut être réduite à l'analyse des fonctions à variation lente. Parmi ces dernières, on trouve :

- Les fonctions ayant une limite strictement positive à l'infini.
- Les fonctions de la forme $U(x) = |\log x|^\beta$, où $\beta \in \mathbb{R}$.
- Les fonctions U telles que

$$\exists M > 0, \forall x \geq M, U(x) = c + dx^{-\beta}(1 + o(1)) \tag{2.12}$$

où $c, \beta > 0$ et $d \in \mathbb{R}$. En d'autres termes, pour x suffisamment grand, $U(x)$ se comporte asymptotiquement comme $c + dx^{-\beta}$.

Ces fonctions appartiennent à la classe de Hall.

Théorème 3.3 *(Représentation de Karamata, Resnick, 1987) Toute fonction à variation lente ℓ à l'infini peut s'exprimer sous la forme*

$$\ell(x) = c(x) \exp\left(\int_1^x \Delta(t) t^{-1} dt \right) \tag{2.13}$$

où $c(\cdot) > 0$ et $\Delta(\cdot)$ sont deux fonctions mesurables telles que

$$\lim_{x\to\infty} c(x) = c_0 \in]0, \infty[\quad et \quad \lim_{x\to\infty} \Delta(x) = 0.$$

Proposition 3.2 *Pour toute fonction à variation lente ℓ à l'infini, on a :*

$$\lim_{x\to\infty} \frac{\log(\ell(x))}{\log(x)} = 0. \tag{2.14}$$

Pour davantage de détails sur la théorie des fonctions à variation régulière, le lecteur peut consulter les ouvrages de de Haan(1985)[23], Bingham et al(1987)[37] ainsi que Lo(1986)[16].

3.4 Domaine d'attraction de Gumbel

La caractérisation des fonctions de répartition appartenant au domaine de Gumbel est plus complexe. Le théorème suivant, démontré notamment dans Resnick(1987)[8], (Proposition 1.4), décrit les conditions sous lesquelles F appartient à $D(H_\xi)$ avec $\xi = 0$.

Théorème 3.4 *Soit F une fonction de distribution ayant un point terminal x_F. Alors, F appartient au domaine d'attraction de Gumbel $D(H_\xi)$, $\xi = 0$, si et seulement s'il existe un réel z, tel que $z < x_F \leq \infty$ et :*

$$1 - F(x) = c(x) \exp\left(- \int_z^x \frac{g(t)}{a(t)} \, dt \right), \quad z < x < x_F, \tag{2.15}$$

où $c(\cdot)$ et $g(\cdot)$ sont des fonctions mesurables positives vérifiant $c(x) \to c > 0$ et $g(x) \to 1$ lorsque $x \uparrow x_F$.

La fonction $a(\cdot)$ est positive et absolument continue par rapport à la mesure de Lebesgue, avec une densité $a'(x)$ telle que $\lim_{x \uparrow x_F} a'(x) = 0$. Dans ce cas, une possible sélection pour les suites a_n et b_n est $a_n = F^{\leftarrow}(1 - 1/n) = U(n)$ et $b_n = a(a_n)$. Un choix possible pour la fonction $a(\cdot)$ est :

$$a(x) = \int_x^{x_F} \frac{F(t)}{F(x)}\, dt, \quad x < x_F. \tag{2.16}$$

La fonction $a(\cdot)$ est appelée fonction auxiliaire.

Remarque 3.3 *La représentation donnée par l'équation 2.15 n'est pas unique, car elle dépend des fonctions $c(\cdot)$ et $g(\cdot)$. Une représentation alternative avec une fonction $g = 1$ peut être utilisée (voir Resnick(1987) [8], Proposition 1.4).*

3.5 Domaine d'attraction de Fréchet

Dans ce contexte, nous désignons par $\bar{F}(\cdot) = 1 - F(\cdot)$ la fonction de survie associée à F, et par $Q(\cdot)$, l'inverse généralisée de F, définie comme $Q(s) = F^{\leftarrow}(s) = \inf\{x \in \mathbb{R} : F(x) \geq s\}$ pour $0 < s \leq 1$. Cette fonction $Q(\cdot)$ est également connue sous le nom de fonction quantile de la distribution F. De plus, la fonction quantile de queue est définie par $U(x) = Q(1 - 1/x)$ pour $x > 1$.

Le théorème 3.5, mentionné initialement par Gnedenko(1943) [5] et démontré dans Resnick(1987) [8], Proposition 1.13, établit que chaque fonction se trouvant dans le domaine d'attraction de Fréchet possède une variation régulière.

Théorème 3.5 *Une fonction de distribution F avec un point terminal x_F est dans le domaine d'attraction de Fréchet $D(H_\xi)$, avec $\xi > 0$, si et seulement si $x_F = +\infty$ et F a une variation régulière d'indice $-1/\xi$ à l'infini, c'est-à-dire :*

$$\lim_{t \to \infty} \frac{\bar{F}(tx)}{\bar{F}(x)} = x^{-1/\xi}. \tag{2.17}$$

Dans ce contexte, les suites de normalisation a_n et b_n sont données par :

$$a_n = F^{\leftarrow}(1 - 1/n) = U(n) \quad et \quad b_n = 0, \quad \forall n > 0. \tag{2.18}$$

Remarque 3.4 *D'après la proposition 3.1, la fonction F appartient au domaine $D(H_\xi)$, $\xi > 0$, si et seulement si $x_F = \infty$ et $F(x) = x^{-1/\xi} \ell_F(x)$ où ℓ_F est une fonction de variation lente à l'infini.*

Remarque 3.5 *Il est évident que pour tout $s \in (0, 1)$, on a $Q(1 - s) = F(1/s)$. Ainsi, d'après la conséquence 3.1, l'équation 2.17 montre que $Q(1 - s)$ a une variation régulière d'indice $-\xi$ en 0, notée $Q(1 - \cdot) \in \mathcal{RV}_0^{-\xi}$. Cela signifie que $Q(1 - s) = s^{-\xi} \ell(1/s)$, avec $\ell \in \mathcal{RV}_0$. Par conséquent, la fonction de queue $U(\cdot)$ a une variation régulière d'indice ξ à l'infini ($U \in \mathcal{RV}_\xi$).*

Remarque 3.6 *Selon l'équation 2.17, $F \in \mathcal{RV}^{-1/\xi}$, $\xi > 0$. Ainsi, les équations 2.2 et 2.13 permettent la représentation suivante :*

$$F(x) = c(x)x^{-1/\xi} \exp\left(\int_1^x \Delta(t)t^{-1}\,dt\right), \quad x < x_F \tag{2.19}$$

avec $\lim_{x\to\infty} c(x) = c_0 \in (0,\infty)$ *et* $\lim_{x\to\infty} \Delta(x) = 0$.

3.6 Domaine d'attraction de Weibull

Le résultat suivant, détaillé dans les travaux de Gnedenko(1943[5] et Resnick(1987) [8], Proposition 1.13, illustre comment une simple transformation de la variable dans la fonction de répartition permet de passer du domaine d'attraction de Fréchet à celui de Weibull.

Théorème 3.6 *Une fonction de distribution F avec un point terminal x_F appartient au domaine d'attraction de Weibull $D(H_\xi)$, avec $\xi < 0$, si et seulement si x_F est fini. La fonction de répartition transformée F^* est définie comme suit :*

$$F^*(x) = \begin{cases} F(x_F - 1/x) & si\ x > 0 \\ 0 & si\ x \le 0. \end{cases} \tag{2.20}$$

Cette fonction F^* se trouve alors dans le domaine d'attraction de Fréchet avec un indice des valeurs extrêmes $-\xi > 0$, signifiant que $\bar{F}^*$ possède une variation régulière d'indice $1/\xi$ à l'infini, notée $\overline{F}^* \in \mathcal{RV}_{1/\xi}$. Dans ce cadre, les suites de normalisation peuvent être choisies comme suit :

$$a_n = x_F - F^{\leftarrow}(1 - 1/n) \quad et \quad b_n = x_F. \tag{2.21}$$

Pour plus de détails sur la démonstration du Théorème [3.6], le lecteur peut se référer à Resnick(1987) [8], Proposition 1.13, ou à Embrechts(1997)[7], Théorème 3.3.12.

Remarque 3.7 *Il découle du Théorème [3.6] que F appartient au domaine d'attraction de Weibull $D(H_\xi)$, avec $\xi < 0$, si et seulement si x_F est fini et que $F(x) = (x_F - x)^{-1/\xi}\ell((x_F - x)^{-1})$, où ℓ est une fonction de variation lente ($\ell \in \mathcal{RV}_0$). De manière équivalente, la fonction quantile $Q(1-s)$ peut s'exprimer sous la forme :*

$$Q(1-s) = x_F - s^{-\xi}\ell(1/s), \quad \ell \in \mathcal{RV}_0. \tag{2.22}$$

3.7 Description générale

Dans la section précédente, nous avons séparément exposé les propriétés de caractérisation ou de description pour les trois domaines d'attraction. En les unifiant, il est possible de développer une propriété de caractérisation "globale" qui englobe les trois lois limites pour les maxima. Il convient également de noter que le domaine d'attraction de H_ξ est construit de manière similaire aux domaines d'attraction des trois distributions des valeurs extrêmes. Les résultats ci-dessous fournissent une caractérisation unifiée de ces trois domaines d'attraction, comme discuté dans de de Haan(2006)[6].

Théorème 3.7 *Soit $\xi \in \mathbb{R}$, les assertions suivantes sont équivalentes :*

1. *La fonction de répartition F appartient au domaine d'attraction de H_ξ.*
2. *Il existe une fonction positive $a(\cdot)$ telle que pour tout $x > 0$,*

$$\lim_{t\to\infty} \frac{U(tx) - U(t)}{a(t)} = D_\xi(x) = \begin{cases} \frac{x^\xi - 1}{\xi}, & si\ \xi \neq 0, \\ \log x, & si\ \xi = 0. \end{cases} \tag{2.23}$$

3. Il existe une fonction positive $f(\cdot)$ telle que pour tout x vérifiant $1+\xi x > 0$,

$$\lim_{t\uparrow x_F} \frac{\overline{F}(t+xf(t))}{\overline{F}(t)} = \begin{cases} (1+\xi x)^{-1/\xi}, & si\ \xi \neq 0, \\ e^{-x}, & si\ \xi = 0. \end{cases} \tag{2.24}$$

Un choix possible pour la fonction $f(\cdot)$ est $f(t) = a(1/(\overline{F}(t)))$, où la fonction $a(\cdot)$ est celle utilisée dans l'équation [2.23].

En supposant que $U(\infty) > 0$, la condition [2.23] implique que

$$\lim_{t\to\infty} \frac{U(tx)-U(t)}{a(t)/U(t)} = \begin{cases} \log x, & \text{si } \xi \geq 0, \\ \frac{x^\xi - 1}{\xi}, & \text{si } \xi < 0. \end{cases} \tag{2.25}$$

Dekkers et al.(1987)[42] ont utilisé ce résultat pour développer l'estimateur des moments du paramètre ξ.

Proposition 3.3 *Soit $\xi \in \mathbb{R}$, $F \in D(H_\xi)$ si et seulement si pour tous $x > 0$, $y > 0$, $y \neq 1$,*

$$\lim_{t\to\infty} \frac{U(tx)-U(t)}{U(ty)-U(t)} = \begin{cases} \frac{x^\xi - 1}{y^\xi - 1}, & si\ \xi \neq 0, \\ \frac{\log x}{\log y}, & si\ \xi = 0. \end{cases} \tag{2.26}$$

Les preuves concernant ces résultats de caractérisation générale peuvent être trouvées dans Embrechts(1997)[7]. La Proposition ci-dessus est utilisée pour construire l'estimateur de Pickands pour ξ.

La méthode classique de la théorie des valeurs extrêmes, connue sous le nom de « block component-wise », a été critiquée en raison de la perte d'informations qu'elle engendre. Elle consiste à estimer les paramètres de la distribution GEV en utilisant uniquement les maxima de blocs de données de même taille, ce qui peut conduire à ignorer certaines valeurs extrêmes ou à inclure des blocs qui n'en contiennent aucune.

Pour surmonter les limites des méthodes basées uniquement sur les maxima annuels, une approche alternative consiste à exploiter toutes les valeurs dépassant un certain seuil. Cette approche est connue sous le nom de méthode POT (Peaks-over-Threshold), ou méthode des excès au-delà d'un seuil élevé, permettant ainsi une meilleure utilisation des données disponibles.

4 Théorie des excédents au-delà de seuils élevés

L'approche Peaks Over Threshold (POT) est une méthode statistique dans la théorie des valeurs extrêmes, qui se concentre sur les observations dépassant un certain seuil plutôt que sur les maximums d'échantillons. Cela permet d'analyser un plus grand nombre de données extrêmes et d'obtenir des estimations plus précises des probabilités et impacts des événements rares. L'approche est particulièrement utile dans des domaines comme la finance, l'hydrologie, et la gestion des risques environnementaux, pour comprendre et gérer les événements extrêmes.

4.1 Excès au-delà d'un seuil

La méthode des excès au-delà d'un seuil, également connue sous le nom de Peak Over Threshold (POT), examine les valeurs qui dépassent un certain seuil. Plutôt que de se concentrer uniquement sur les valeurs maximales, cette méthode prend en compte toutes les valeurs qui excèdent un

seuil prédéfini. Le principe de base est de choisir un seuil suffisamment élevé et d'étudier les valeurs qui le dépassent.

Cette approche a été initialement développée par Pickands en 1975 et a été largement explorée par d'autres chercheurs, tels que Smith(1987)[30], Davison et Smith(1990)[34], ainsi que Reiss et Thomas (2001)[33].

Pour formaliser cette méthode, on définit un seuil $u \in \mathbb{R}$. Le nombre de dépassements du seuil u est donné par $N_u = \text{card}\{i : i = 1, \ldots, n, X_i > u\}$, et les excès correspondants sont $Y_i = X_i - u > 0$ pour $1 \leq j \leq N_u$, où N_u représente le nombre de dépassements du seuil u par les X_i (pour $1 \leq i \leq n$), et $Y_1, \ldots, Y_{N_u}$ sont les excès correspondants (voir Figure 2.2).

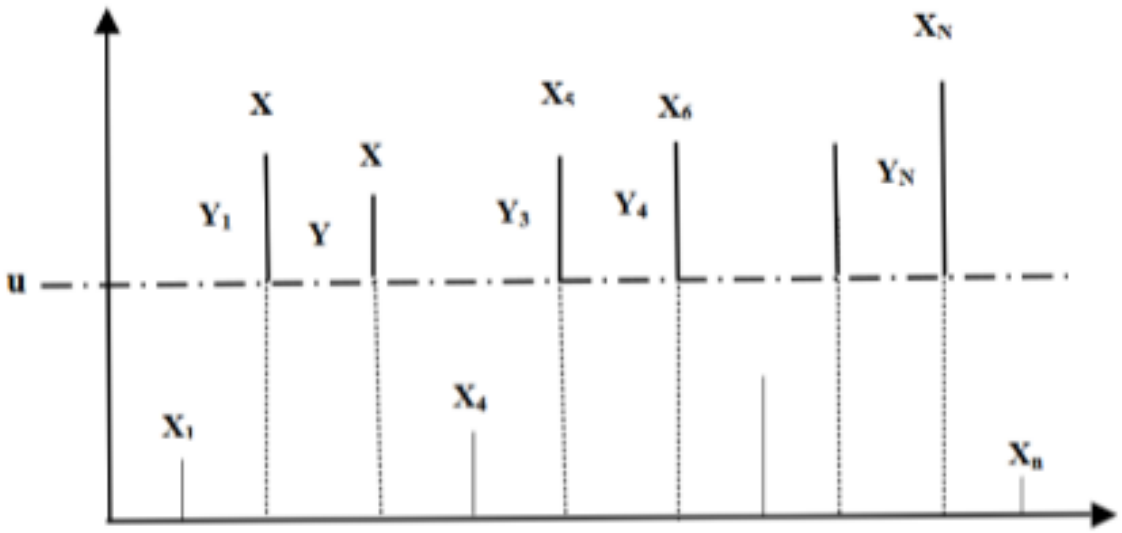

FIGURE 2.2 – Méthode des excès

Nous cherchons à partir de la fonction de répartition F de X à définir une loi conditionnelle F_u par rapport au seuil u pour les variables aléatoires dépassant ce seuil.

Le théorème de Pickands-Balkema-de Haan énonce la forme limite des valeurs extrêmes : sous certaines conditions de convergence, la loi limite est une loi de Pareto généralisée, notée GPD.

La théorie des valeurs extrêmes peut également être formulée en examinant le comportement asymptotique des excès $X - u$ au-dessus d'un seuil u, plutôt que de se concentrer uniquement sur les maximums. Cette approche est fondée sur la loi des excès, qui est définie comme suit :

Définition 4.1 *Soit X une variable aléatoire avec une fonction de répartition F. Pour un seuil u inférieur à la valeur quantile x_F, la loi des excès de X au-dessus du seuil u est :*

$$F_u(x) = P(X - u \leq x \mid X > u)$$

En développant l'expression de la fonction de répartition des excès au-delà du seuil u, on obtient :

$$F_u(x) = \frac{F(x+u) - F(u)}{1 - F(u)} \tag{2.27}$$

Cela revient à $F_u(y) = P(X \leq y | X > u) = \frac{F(y)-F(u)}{1-F(u)}$ pour $y \geq u$. F_u représente la distribution de X conditionnée par le fait que X est supérieur à u.

Si la distribution parente F était connue, il serait possible de déterminer la distribution des dépassements de seuil comme dans l'équation 2.27. Cependant, dans les applications pratiques, cette distribution n'est généralement pas connue. Par conséquent, on utilise des approximations qui sont généralement applicables pour des seuils élevés. Cela est similaire à l'utilisation de la distribution GEV pour approximer la distribution des maxima de longues séquences lorsque la distribution parente est inconnue.

En supposant que la distribution des maxima peut être approximée par une loi GEV (Généralisée des Valeurs Extrêmes) avec la fonction de répartition $H_{\xi,\mu,\beta}$ selon le théorème de Fisher-Tippett,pour certains $\mu, \beta > 0 \quad$ et $\quad \xi$, on peut approximer :

$$F^n(x) \approx \exp\left[-\left(1+\xi\frac{x-\mu}{\beta}\right)^{-\frac{1}{\xi}}\right]$$

Ainsi, en utilisant une approximation pour des valeurs élevées de x, on obtient :

$$\ln(F(x)) \approx -\left(1-F(x)\right)$$

D'où :

$$1-F(x) \approx \frac{1}{n}\left(1+\xi\frac{x-\mu}{\beta}\right)^{-\frac{1}{\xi}}$$

En revenant à l'expression de $F_u(x)$, on trouve :

$$F_u(x) \approx 1-\left(1+\xi\frac{x+u-\mu}{1+\xi(\frac{u-\mu}{\beta})}\right)^{-\frac{1}{\xi}} \approx 1-\left(1+\xi\frac{x}{\tilde{\beta}}\right)^{-\frac{1}{\xi}} \tag{2.28}$$

où $\tilde{\beta} = \beta + \xi(u-\mu)$. Cette expression représente la fonction de répartition d'une loi de Pareto Généralisée, indiquant que la distribution de Pareto Généralisée est une bonne approximation de la loi des excès $X-u \mid X > u$.

4.2 Distribution de Pareto Généralisée (GPD)

La distribution de Pareto généralisée (GPD) est cruciale pour modéliser les excès au-delà d'un seuil.

Théorème 4.1 *(Pickands-Balkema-de Haan)*
Une fonction de répartition F de point terminal x_F appartient au domaine d'attraction maximale de H_ξ, si et seulement si, il existe une fonction positive $\beta(u)$ telle que :

$$\lim_{u\to x_F}\sup_{0\le y\le x_F-u}|F_u(y)-G_{\xi,\beta(u)}(y)| = 0$$

dont $F_u(y)$ représente la fonction de répartition conditionnelle des excès pour un seuil u élevé et $G_{\xi,\beta(u)}(y)$ est la fonction de répartition de la loi de Paréto Généralisée.

La fonction $G_{\xi,\beta(u)}(y)$ définie comme suit :

$$G_{\xi,\beta(u)}(y) = \begin{cases} 1-\left(1+\xi\frac{y}{\beta(u)}\right)^{-1/\xi}, & \xi \neq 0 \\ 1-\exp\left(-\frac{y}{\beta(u)}\right), & \xi = 0 \end{cases}, \quad \beta(u) \ge 0 \tag{2.29}$$

Pour

$$\begin{cases} y \in [0, (x_F - u)] & \text{si } \xi \geq 0 \\ y \in [0, \min(\frac{-\beta}{\xi}, x_F - u)] & \text{si } \xi < 0 \end{cases}$$

Les conditions sur y dépendent de ξ, avec $y \geq 0$ lorsque $\xi \geq 0$, et $0 \leq y \leq -\frac{\beta(u)}{\xi}$ lorsque $\xi < 0$.

Ce théorème 4.1 indique que pour un seuil u suffisamment élevé, la loi de Pareto Généralisée constitue une approximation adéquate de la loi des excès. La fonction de répartition $F_u(x-u)$ peut ainsi être approximée par une loi de Pareto Généralisée avec les paramètres ξ et $\beta(u)$ adaptés à u.

Le paramètre ξ de la distribution de Pareto généralisée est identique à celui utilisé dans la distribution des valeurs extrêmes généralisée (GVE).

— ξ : paramètre de forme qui détermine la queue de la distribution.
— β : paramètre d'échelle qui détermine l'étendue des valeurs au-dessus du seuil.

Remarque 4.1 *En fonction du signe de ξ, nous distinguons les cas suivants :*

— *$\xi > 0$: La distribution est de type Pareto, caractérisée par une queue lourde.*
— *$\xi = 0$: La distribution est de type exponentielle, avec une queue légère.*
— *$\xi < 0$: La distribution est de type Beta, bornée au-dessus de $u - \frac{\beta}{\xi}$.*

Remarque 4.2

— *On peut étendre cette distribution en introduisant un paramètre de position μ. La forme généralisée s'écrit alors $G_{\xi,\beta,\mu}(y) = G_{\xi,\beta}(y-\mu)$.*
— *Lorsque $\xi = 0$, on obtient une distribution exponentielle avec paramètre $\frac{1}{\beta}$.*
— *La distribution Pareto Généralisée peut être reliée à la distribution des valeurs extrêmes par la relation : $G_{\xi,\beta}(y) = 1 + \ln[H_\xi\left(\frac{y}{\beta}\right)]$.*

Illustration des fonctions de répartition et de densité des lois GPD

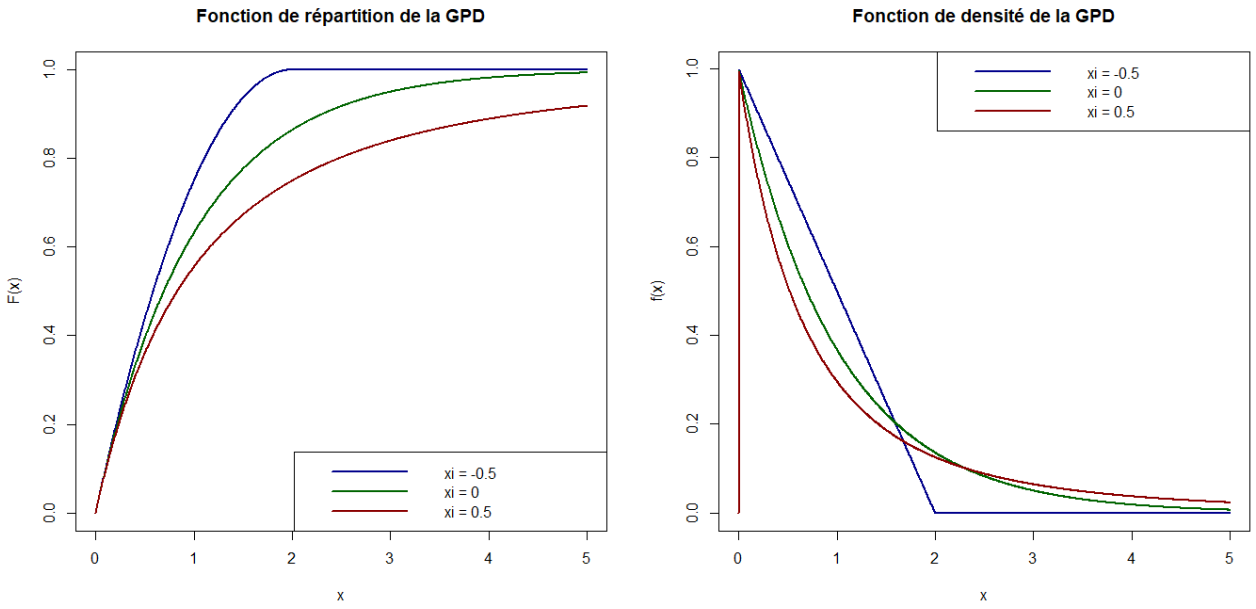

FIGURE 2.3 – Fonction de Repartition et de densité pour $\xi = -0.5$, $\xi = 0$, $\xi = 0.5$

La figure 2.3 illustre les fonctions de répartition et de densité des distributions de Pareto pour un paramètre d'échelle fixé à 1. En variant le paramètre de forme entre -0.5 et 0.5, on obtient différentes lois GPD comme suit :

— $\xi = -0.5$ pour la loi de Pareto II (en rouge) :
— $\xi = 0$ pour la loi exponentielle (en vert) :
— $\xi = 0.5$ pour la loi de Pareto (en bleu) :

Interprétation

Courbes de densité (PDF) :

— Pour $\xi = -0.5$, la densité tombe rapidement à zéro, indiquant des valeurs maximales finies.
— Pour $\xi = 0$, la densité décroît exponentiellement, typique d'une loi exponentielle.
— Pour $\xi = 0.5$, la densité décroît lentement, indiquant des valeurs extrêmes plus probables.

Courbes de répartition (CDF) :

— Pour $\xi = -0.5$, la courbe atteint 1 rapidement, signifiant que les grandes valeurs sont rares.
— Pour $\xi = 0$, la courbe suit une distribution exponentielle.
— Pour $\xi = 0.5$, la courbe atteint 1 plus lentement, signifiant que les grandes valeurs sont plus probables.

En sélectionnant un seuil approprié, il est possible d'ajuster une loi de Pareto Généralisée aux données qui dépassent ce seuil. Cependant, la détermination d'un seuil optimal est un défi pratique. Nous veront dans le Chapitre 4 les differentes méthodes du choix de ce seuil.

3 Troncature

Contents

1 Introduction

En théorie des valeurs extrêmes, l'estimation de l'indice des valeurs extrêmes (EVI, pour Extreme Value Index) joue un rôle central dans la caractérisation des queues de distribution. Cet indice est essentiel pour diverses applications, allant de la gestion des risques financiers à la prévision des événements climatiques extrêmes. Toutefois, dans de nombreux contextes pratiques, les données disponibles sont tronquées, c'est-à-dire qu'elles sont limitées à un certain intervalle en raison de contraintes de mesure ou de sélection. La troncature des données introduit des défis supplémentaires dans l'estimation de l'indice des valeurs extrêmes. La troncature peut être à gauche, à droite, ou les deux, et elle modifie la distribution observée des données, rendant les méthodes d'estimation traditionnelles inadaptées ou biaisées. Par exemple, dans les études de mortalité humaine, les données peuvent être tronquées à gauche si seules les personnes d'un certain âge minimum sont incluses dans l'étude. Dans les domaines financiers, les pertes extrêmes peuvent être tronquées à droite en raison de limites de déclaration.

Dans ce chapitre, nous examinerons les différents types de troncature et leurs effets sur les distributions de données.

2 Types de Troncature et leurs effets

Dans l'analyse statistique des données, la troncature survient lorsque les observations sont limitées à un certain intervalle, ce qui signifie que les valeurs en dehors de cet intervalle ne sont pas disponibles pour l'analyse. La troncature peut se produire de différentes manières, chacune ayant des effets distincts sur la distribution des données observées et, par conséquent, sur les méthodes d'estimation utilisées pour analyser ces données. Dans cette section, nous examinerons en détail les types de troncature les plus courants : la troncature à gauche, la troncature à droite, et la troncature double. Nous discuterons également des effets de ces types de troncature sur les distributions de données et les implications pour l'estimation de l'indice des valeurs extrêmes.

De nombreuses études ont été menées sur l'analyse des données tronquées (voir Klein et Moeschberger(1997)[36] et Klein et Moeschberger(2003)[35]).

2.1 Troncature à Gauche

Définition 2.1 *La troncature à gauche se produit lorsque les observations en dessous d'un certain seuil T_L ne sont pas incluses dans l'échantillon. En d'autres termes, seules les valeurs $X_i \geq T_L$ sont observées.*

2.2 Exemple de troncature à gauche.

Exemple 2.1 *Une étude sur la survie des résidents d'une maison de retraite en Californie est décrite. L'âge au décès de chaque résident est enregistré pour indiquer le risque encouru, garantissant ainsi que seuls ceux ayant un risque actuel sont inclus. Les résidents qui ont intégré la maison de retraite après l'âge initial prévu pour l'étude sont inclus dans cette analyse. C'est un exemple typique de troncature à gauche.*

Pour cet exemple, le lecteur peur voir textcolorblueKlein et Moeschberger(1997)[36]

Exemple 2.2 *Supposons qu'une étude soit menée pour examiner les caractéristiques des personnes atteintes du SIDA. Dans cette étude, seules les personnes qui ont développé le SIDA (la phase avancée de l'infection par le VIH) sont incluses dans l'échantillon. Les personnes qui sont séropositives mais n'ont pas encore développé le SIDA ne sont pas observées ni incluses dans l'étude.*

Cela signifie que les cas moins graves (c'est-à-dire les personnes infectées par le VIH mais n'ayant pas encore développé le SIDA) sont systématiquement exclus de l'échantillon. Le "seuil" de troncature ici est le développement du SIDA. Seules les observations des personnes ayant dépassé ce seuil (c'est-à-dire ayant développé le SIDA) sont prises en compte dans l'étude.

NB : Dans le contexte des pertes d'assurance, l'utilisation d'une troncature à gauche est souvent pratiquée pour se concentrer sur les réclamations plus importantes, qui ont un impact financier plus significatif pour la compagnie d'assurance.

Effets sur la Distribution :

La troncature à gauche modifie la distribution des données observées en éliminant les valeurs plus petites que T_L. Cela peut biaiser les estimations des paramètres de la distribution originale si les méthodes traditionnelles d'estimation sont utilisées sans ajustement pour la troncature.

Implications pour l'Estimation de l'Indice des Valeurs Extrêmes :

Pour l'indice des valeurs extrêmes, la troncature à gauche peut rendre difficile l'estimation précise des queues de distribution inférieures. Des méthodes spécifiques doivent être employées pour ajuster cette troncature afin d'obtenir des estimations non biaisées de l'indice des valeurs extrêmes.

2.3 Troncature à Droite

Définition 2.2 *La troncature à droite se produit lorsque les observations au-dessus d'un certain seuil T_U ne sont pas incluses dans l'échantillon. En d'autres termes, seules les valeurs $X_i \leq T_U$ sont observées.*

2.4 Exemple de troncature à droite.

Exemple 2.3 *voir Lagakos et al.(1988)[28] et Klein et Moeschberger(1997)[36]*

Cet exemple présente des données sur l'infection et le temps d'induction pour 258 adultes et 37 enfants infectés par le virus du sida et ayant développé la maladie avant le 30 juin 1986. Les données sont exprimées en termes d'années, en commençant à partir du 1er avril 1978, date à laquelle les adultes ont été contaminés par une transfusion sanguine infectée. Le temps écoulé jusqu'à l'apparition de la maladie est mesuré depuis la date d'infection. Concernant la population pédiatrique, les enfants ont été infectés dès la naissance, et la période d'infection est calculée à partir du 1er avril 1978 jusqu'à leur naissance.

Dans cette étude d'échantillonnage, seules les personnes ayant développé le SIDA avant la fin de la période d'étude sont incluses. Les personnes infectées mais n'ayant pas encore développé la maladie ne sont pas retenues dans l'échantillon. Ce type de données est connu sous le nom de données tronquées à droite.

Exemple 2.4 *Considérons une étude sur la durée de vie des ampoules. Nous testons 10 ampoules en les allumant simultanément et en enregistrant leur temps de défaillance. Cependant, après 1000 heures, l'expérience est arrêtée, et il reste 2 ampoules encore allumées. Les données sont alors tronquées à droite, car la durée de vie exacte des ampoules restantes au-delà de 1000 heures n'est pas connue.*

Effets sur la Distribution :

La troncature à droite modifie la distribution des données observées en éliminant les valeurs plus grandes que T_U. Cela affecte particulièrement l'estimation des queues de distribution supérieures, qui sont cruciales pour l'analyse des valeurs extrêmes.

Implications pour l'Estimation de l'Indice des Valeurs Extrêmes :

La troncature à droite complique l'estimation des queues de distribution supérieures, ce qui est directement pertinent pour l'indice des valeurs extrêmes. Les méthodes d'estimation doivent être adaptées pour tenir compte de cette troncature, sinon les résultats peuvent être gravement biaisés.

2.5 Troncature Double

Définition 2.3 *La troncature double survient lorsque les observations sont limitées à un intervalle $[T_L, T_U]$, ce qui signifie que seules les valeurs X_i telles que $T_L \leq X_i \leq T_U$ sont observées.*

Exemple 2.5 *En Médecine Les essais cliniques pour un nouveau médicament contre l'hypertension peuvent exclure les patients qui ont commencé le traitement avant le début de l'étude (troncature à gauche) et ceux dont les résultats ne sont pas disponibles à la fin de l'étude (troncature à droite). Par exemple, une étude menée par Smith et al.(2015)[32] a examiné 500 patients traités pour hypertension entre 2010 et 2015. Les patients ayant commencé le traitement avant 2010 ou après 2015 ont été exclus des données.*

Exemple 2.6 *On suppose qu'une étude épidémiologique soit menée pour estimer la distribution des âges au diagnostic d'une certaine maladie. Cependant, les données disponibles proviennent uniquement des patients diagnostiqués entre 30 et 70 ans. Les patients diagnostiqués avant*

30 ans ou après 70 ans ne sont pas inclus dans l'étude pour diverses raisons, telles que des restrictions sur les données de l'hôpital ou des critères de l'étude.

Dans ce cas, on dit qu' il y a une double troncature :

- *Troncature à gauche : Les patients diagnostiqués avant 30 ans ne sont pas inclus dans l'échantillon.*
- *Troncature à droite : Les patients diagnostiqués après 70 ans ne sont pas inclus non plus.*

Ainsi, l'étude ne prend en compte que les patients dont l'âge au moment du diagnostic se situe entre 30 et 70 ans, excluant les cas en dehors de cet intervalle.

Effets sur la Distribution :

La troncature double combine les effets des troncatues à gauche et à droite, en éliminant à la fois les valeurs inférieures à T_L et supérieures à T_U. Cela peut rendre l'estimation des paramètres de la distribution originale encore plus complexe.

Implications pour l'Estimation de l'Indice des Valeurs Extrêmes :

L'estimation de l'indice des valeurs extrêmes sous troncature double nécessite des techniques spécifiques qui peuvent ajuster les effets des deux types de troncature. Les méthodes d'estimation doivent être capables de reconstruire les queues de la distribution en utilisant uniquement les données tronquées disponibles.

Effets Spécifiques de la Troncature sur les Distributions de Données

La troncature modifie la forme des distributions de données observées. Voici quelques effets spécifiques de chaque type de troncature :

- Réduction de la Variance : la troncature à gauche et à droite réduit la variance observable, car les valeurs extrêmes (soit très petites, soit très grandes) sont exclues.
- Biais des Estimations : les estimateurs de la moyenne, de la variance et des autres paramètres peuvent être biaisés si la troncature n'est pas correctement prise en compte.
- Distribution Troncaturée : la distribution originale des données est transformée en une distribution tronquée, où les densités de probabilité sont ajustées en conséquence. Par exemple, la densité de probabilité d'une variable aléatoire tronquée à gauche est ajustée pour être nulle en dessous du seuil de troncature.
- Effet sur les Quantiles : les quantiles de la distribution observée ne correspondent pas directement aux quantiles de la distribution originale, nécessitant des ajustements pour une estimation correcte.

3 Conclusion

La troncature des données a des effets significatifs sur la distribution des données observées et sur les méthodes d'estimation utilisées pour analyser ces données. En comprenant les types de troncature et leurs effets, nous pouvons développer et appliquer des méthodes d'estimation adaptées qui prennent en compte ces modifications, garantissant ainsi des estimations précises et non biaisées des indices des valeurs extrêmes. Dans la suite de notre travail, nous nous concentreront uniquement sur le modèle de troncature à gauche, offrant des solutions pratiques pour surmonter les défis posés par ce modèle de troncature sur les données.

4 Estimation des paramètres sous Troncature

Contents

1 Introduction

Cette section examine diverses méthodes d'estimation des paramètres ξ ou σ dans la distribution asymptotique des valeurs extrêmes. Parmi les méthodes notables figurent celles non-paramétriques telles que les estimateurs de Pickands(1975)[15], Hill(1975)[11], Dekkers et al.(1987)[26], Peng(1998)[10], et Danielsson et al.(1996)[31]. On trouve également des estimateurs basés sur le QQ-plot, le graphique de la moyenne des excès, et des techniques de régression. Les méthodes des moments, des moments pondérés, et du maximum de vraisemblance sont aussi fréquemment utilisées. Ces méthodes sont théoriquement cohérentes et asymptotiquement normales, mais présentent des différences significatives dans les simulations. Aucune méthode n'est universellement supérieure, bien que celles de Hill, Pickands et des moments soient les plus couramment employées, avec des comparaisons détaillées dans la littérature.

Tsourti et Panaretos(2003)[38] soutiennent que l'efficacité d'une méthode d'estimation est influencée par la distribution des données et la véritable valeur de l'indice de queue. Ils préconisent l'emploi de techniques telles que le graphique log-log, la moyenne empirique des excès, et la

statistique de Jackson pour identifier le domaine d'attraction de la loi des valeurs extrêmes. Pour plus de détails, voir El-Adlouni et al.(2007)[39]. La loi des valeurs extrêmes, lorsqu'elle est définie, est caractérisée par l'indice des valeurs extrêmes, ainsi que par des paramètres d'échelle et de position, essentiels pour déterminer l'épaisseur de la queue de la distribution.

2 Estimation de l'indice de queue dans un modèle de type Pareto

La question cruciale est comment estimer cet indice à partir d'un échantillon fini $(X_1, ..., X_n)$. Deux approches principales sont utilisées : l'approche EVT, basée sur la GEVD, et l'approche POT, utilisant la GPD. Chacune de ces approches propose des méthodes paramétriques et non paramétriques pour l'estimation. Les méthodes paramétriques incluent le maximum de vraisemblance, les moments, les moments de probabilités pondérées Hosking et Wallis(1987)[40], et les méthodes de régression Beirlant et Goegebeur(2003)[44]. Les méthodes non paramétriques comprennent les estimateurs de Pickands (1975)[15], Hill(1975)[11], et les moments DEdH Dekkers, Einmahl et de Hann(1989)[42].

Dans la suite de notre travail,nous allons nous concentrer principalement sur les méthodes non paramétriques et leurs propriétés asymptotiques. Nous présenterons et analyserons aussi les méthodes statistiques adaptées à l'estimation de l'indice des valeurs extrêmes dans le contexte de données tronquées, en discutant leurs propriétés théoriques et pratiques, ainsi que leurs performances à travers des simulations et des études de cas.

3 Estimateur Non-pamétrique

La loi des valeurs extrêmes est une description fondamentale dans la théorie des valeurs extrêmes, spécifiant la forme asymptotique des distributions des valeurs extrêmes. Selon le théorème des valeurs extrêmes, pour un ensemble de variables aléatoires $X_1, X_2, \ldots, X_n$, les valeurs extrêmes suivent, asymptotiquement, l'une des trois types de distribution de valeurs extrêmes : la distribution de Gumbel, la distribution de Fréchet, ou la distribution de Weibull. Ces trois types peuvent être unifiés dans une distribution appelée la distribution de valeurs extrêmes généralisée (GEV). La fonction de répartition de la GEV est :

$$H_\theta(x) = \begin{cases} \exp\left\{-\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-1/\xi}\right\} & \text{si } \xi \neq 0 \text{ et } \left(1+\xi\frac{x-\mu}{\sigma}\right) > 0 \\ \exp\left\{-\exp\left(-\frac{x-\mu}{\sigma}\right)\right\} & \text{si } \xi = 0 \text{ et } x \in \mathbb{R} \end{cases} \tag{4.1}$$

où le vecteur $\theta = (\xi, \mu, \sigma)$ appartient à l'ensemble $\Theta \subseteq \mathbb{R}^2 \times \mathbb{R}^+$. Ce vecteur inclut un paramètre de forme (ξ), un paramètre de localisation (μ), et un paramètre d'échelle (σ). Ces paramètres doivent être estimés à partir d'un échantillon $(X_1, \ldots, X_n)$ composé de n variables aléatoires indépendantes et identiquement distribuées selon une fonction de répartition F.

Dans la littérature, deux estimateurs les plus couramment utilisés sont celui de Hill(1975)[11] et de Pickands(1975)[15]. On désigne par $X_{1,n}, \ldots, X_{n,n}$ les statistiques d'ordre associées à l'échantillon $X_1, \ldots, X_n$. Autrement dit, les valeurs de $X_1, \ldots, X_n$ sont classées par ordre croissant de telle sorte que

$$X_{1,n} \leq X_{2,n} \leq \ldots \leq X_{n,n}$$

.

On s'intéresse aux k plus grandes (ou plus petites) valeurs de cet échantillon, où k dépend a priori de n, bien que cette dépendance ne soit pas explicitement indiquée dans la notation. L'idée est d'avoir $k \to \infty$ lorsque $n \to \infty$, mais sans choisir "trop" de valeurs, ce qui impose

$k/n \to 0$. Cela pose naturellement la question du choix optimal de k. En effet, pour calculer ces estimateurs sur les queues de distribution, un k trop élevé pourrait inclure des valeurs qui ne sont pas vraiment extrêmes, tandis qu'un k trop faible ne permettrait pas aux estimateurs d'atteindre une stabilité suffisante.

Il est également important de noter que pour les échantillons de petite taille, il est préférable de se tourner vers une approche paramétrique. En revanche, l'approche non paramétrique n'est envisageable que si l'on dispose d'un grand nombre d'observations. Ce qui est le cas de notre étude.

3.1 Estimateur de Hill

Une grande partie de la théorie concernant l'estimation de l'indice de valeur extrême est axée sur les indices de valeurs extrêmes positifs. L'estimateur de Hill(1975)[11] est l'estimateur le plus couramment utilisé pour estimer l'indice des valeurs extrêmes ξ. Toutefois, il est applicable uniquement aux distributions appartenant au domaine d'attraction de Fréchet, c'est-à-dire lorsque $\xi > 0$. Il repose sur les observations les plus grandes de l'échantillon et se définit en fonction du nombre k_n d'excès considéré par :

$$\boxed{\hat{\xi}^H(k_n) = \frac{1}{k_n}\sum_{i=1}^{k_n} \ln(X_{n-i+1,n}) - \ln(X_{n-k_n,n})} \tag{4.2}$$

Si k_n est choisi de telle sorte que $\lim\limits_{n\to\infty} k_n = \infty$ et $\lim\limits_{n\to\infty} \frac{k_n}{n} = 0$, on peut démontrer que l'estimateur de Hill suit une distribution gaussienne asymptotique :

$$\sqrt{k_n}\left(\hat{\xi}^H(k_n) - \xi\right) \xrightarrow{L} N(0, \xi^2) \quad \text{quand} \quad n \to \infty$$

En outre, il est possible de montrer que l'estimateur de Hill est dérivé de l'estimateur du maximum de vraisemblance pour une distribution de Pareto avec un paramètre α tel que $\alpha = \frac{1}{\xi}$.

Lemme 3.1 *Soit L une fonction à variation lente. Alors, pour tout $p > 0$, les assertions suivantes sont vraies :*

1. $L(x) = O(x^p)$ *lorsque* $x \to \infty$.
2. $\int_x^\infty t^{-p-1}\,dt \sim \frac{1}{p}x^{-p}L(x)$ *lorsque* $x \to \infty$.

Lemme 3.2 *Soit $F \in D(\Phi_\xi)$. On a :*

$$\frac{1}{\bar{F}(t)} E\left[(\log(X) - \log(t))\,1_{\{X>t\}}\right] \xrightarrow[t\to\infty]{} \frac{1}{\alpha} = \xi$$

Propriétés asymptotiques de $\hat{\xi}^H$

(a) Mason(1982)[45] a établi les propriétés asymptotiques de l'estimateur de Hill, démontrant sa consistance faible pour toute suite vérifiant :

$$k = k_n \to \infty \quad et \quad k_n/n \to 0 \quad lorsque \quad n \to \infty.$$

(b) La consistance forte a été prouvée par Deheuvels, Haeusler et Mason(1987)[14] sous les conditions que

$$k/\log\log n \to \infty \quad et \quad k_n/n \to 0 \quad lorsque \quad n \to \infty.$$

(c) Sous certaines conditions de second ordre, la normalité asymptotique de l'estimateur de Hill a été démontrée par Davis et Resnick(1984)[46], Haeusler et Teugels(1985)[47], Goldie et Smith(1987)[48], Dekkers et al.(1989)[26], montrant que

$$\sqrt{k}(\hat{\xi}_n^{(H)} - \xi) \sim N(0, \xi^2).$$

On peut associer à l'estimateur de Hill un intervalle de confiance asymptotique $I_n(\alpha)$ de niveau α :

$$I_n(\alpha) = \left[\hat{\xi} - z_{\alpha/2}\frac{\hat{\xi}}{\sqrt{k}}, \hat{\xi} + z_{\alpha/2}\frac{\hat{\xi}}{\sqrt{k}}\right],$$

où $z_{\alpha/2}$ est le quantile d'ordre $1 - \alpha/2$ de la loi normale centrée réduite.

La convergence se fait en loi. Dans le domaine de Fréchet, la fonction de survie est généralement exprimée sous la forme

$$1 - F(x) = x^{-1/\xi}\ell(x),$$

où ℓ est une fonction à variation lente. Cette formulation engendre un biais considérable pour l'estimateur de Hill, rendant son utilisation en pratique assez complexe. Dans ce contexte, la fonction ℓ agit comme un paramètre de nuisance de dimension infinie, ce qui complique l'estimation (voir Bertail(2002)[49]). Pour plus de détails sur la consistance de $\hat{\xi}^H$, voir Beirlant et al.(2007)[51]. Nous commencerons par examiner les conditions de premier et second ordre.

Proposition 3.1 *(Première condition, selon De Haan et Ferreira(2006)[6] Les déclarations suivantes sont équivalentes :*

1. *La distribution F est caractérisée par une queue lourde.*
2. *La fonction $1 - F$ montre une variation régulière à l'infini avec un indice $-1/\xi$, exprimé par :*
$$\lim_{t\to\infty}\frac{1 - F(tx)}{1 - F(t)} = x^{-1/\xi}, \quad x > 0$$
3. *La fonction $Q(1 - s)$ présente une variation régulière à zéro, avec un indice $-\xi$, tel que :*
$$\lim_{x\to 0}\frac{Q(1 - sx)}{Q(1 - s)} = x^{-\xi}, \quad x > 0$$
4. *La fonction U montre une variation régulière à l'infini avec un indice ξ, comme suit :*
$$\lim_{t\to\infty}\frac{U(tx)}{U(t)} = x^{\xi}, \quad x > 0$$

Proposition 3.2 *(Seconde condition, selon De Haan et Ferreira(2006)[6] Une fonction de répartition $F(\cdot)$ appartenant à $\mathcal{D}$(Fréchet), où $\xi > 0$, satisfait une condition de second ordre à l'infini si elle respecte l'une des conditions suivantes :*

1. *Il existe un paramètre $\rho \leq 0$ et une fonction $A_1(\cdot)$ qui approche 0 sans changer de signe à l'infini, définie par : $\forall x > 0$,*
$$\lim_{t\to\infty}\frac{1 - F(tx)}{1 - F(t)} = x^{-1/\xi}\left(1 + A_1(t)x^{\rho}\frac{x^{\rho} - 1}{\rho}\right)$$

2. *Un paramètre $\rho \leq 0$ et une fonction $A_2(\cdot)$ qui tendent vers 0 sans changer de signe à zéro existent, définis par : $\forall x > 0$,*

$$Q(1-x) = x^{-\xi}\left(1 + A_2(x)x^{\rho}\frac{x^{\rho}-1}{\rho}\right)$$

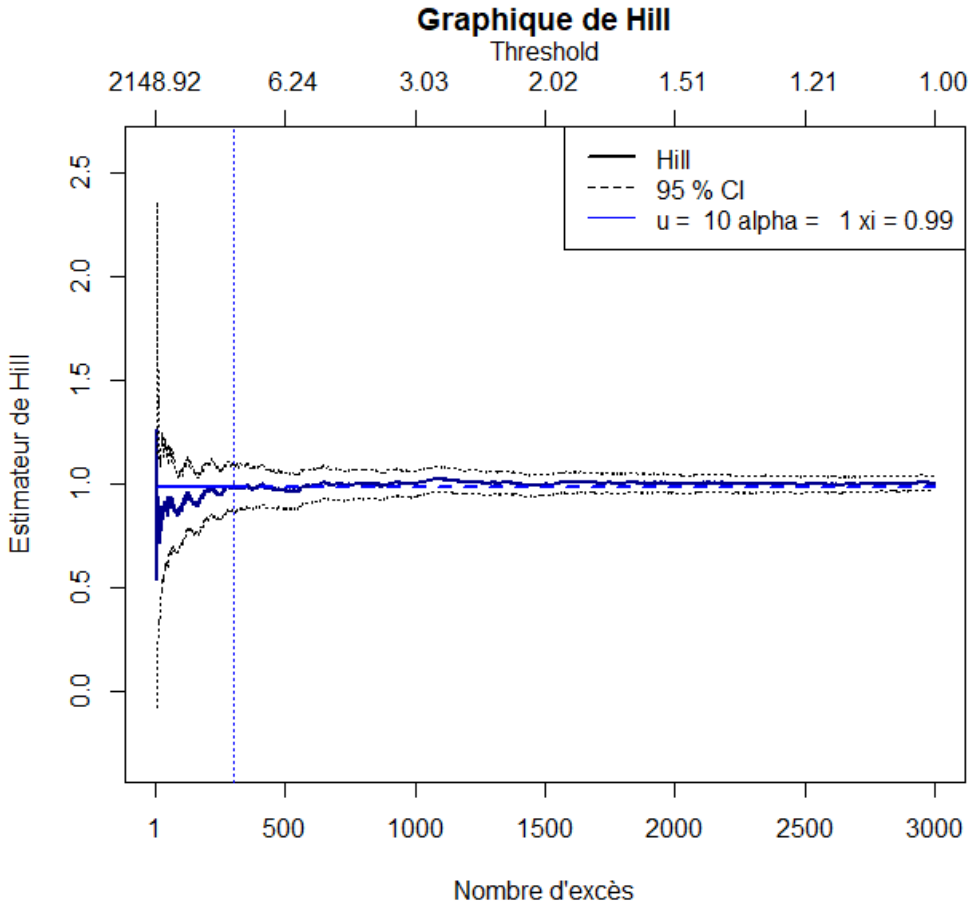

FIGURE 4.1 – L'estimateur de Hill en fonction des valeurs de k_n pour un échantion de taille 3000 de la de pareto avec un IC de 95%

3.2 Estimateur de Pickands

L'estimateur de Pickands a été introduit en 1975 par James Pickands dans Pickands(1975)[15] pour toute valeur de $\xi \in \mathbb{R}$. Il a l'avantage d'être applicable à ces trois domaines d'attraction sans aucune restriction sur la valeur de l'indice des valeurs extrêmes ξ.

Définition 3.1 *(Estimateur de Pickands)*
Soit $X_1, \ldots, X_n$ une suite de variables indépendantes et identiquement distribuées de fonction de répartition $F \in DA(H_\xi)$, où $\xi \in \mathbb{R}$. Soit $k = k_n$ une suite d'entiers avec $1 < k < n$. L'estimateur de Pickands est défini pour tout $\xi \in \mathbb{R}$ par :

$$\boxed{\hat{\xi}^P(k_n) = \frac{1}{\log 2}\log\left(\frac{X_{n-\lceil\frac{k_n}{4}\rceil+1,n} - X_{n-\lceil\frac{k_n}{2}\rceil+1,n}}{X_{n-\lceil\frac{k_n}{2}\rceil+1,n} - X_{n-k_n+1,n}}\right)} \tag{4.3}$$

L'auteur démontre la consistance faible de cet estimateur. La convergence forte ainsi que la normalité asymptotique ont été démontrées par Dekkers et De Haan(1989)[52]. Drees(1995)[53] et Dekkers et De Haan(1989)[52] ont proposé des améliorations de cet estimateur. En particulier, ils ont démontré les résultats suivants :

Théorème 3.1 *Dekkers et De Haan(1989)[52]*

Soit $X_1, \ldots, X_n$ une suite de variables aléatoires iid suivant F. Les statistiques d'ordre sont notées

$$X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$$

.

Supposons que F appartient au domaine d'attraction de la loi H_ξ, notée $DA(H_\xi)$, où $\xi \in R$.

a) Consistance faible :

$$\hat{\xi}_P \xrightarrow{P} \xi \text{ lorsque } n \to \infty$$

Pickands démontre que si $k/n \to 0$ et $k \to \infty$ alors :

$$\frac{X_{n-i+1:n} - X_{n-j+1:n}}{X_{n-j+1:n} - X_{n-l+1:n}} \approx 2^\xi.$$

En utilisant les propriétés asymptotiques des statistiques d'ordre et des quantiles, Pickands montre que l'estimateur converge en probabilité vers ξ.

b) Consistance forte : Supposons que $k/\log\log n \to \infty$ lorsque $n \to \infty$,

$$\hat{\xi}^P_{k,n} \xrightarrow{p.s.} \xi, \text{ lorsque } n \to \infty$$

c) Normalité asymptotique :

$$\sqrt{k}(\hat{\xi}^P_{k,n} - \xi) \xrightarrow{D} N\left(0, \xi^2(\xi)\right), \text{ lorsque } n \to \infty$$

où

$$\xi^2(\xi) = \frac{\xi^2(2^{2\xi+1}+1)}{(2(2^\xi-1)\log 2)^2}.$$

Notons que l'estimateur de Pickands ξ^P, qui repose sur les distances entre deux statistiques ordonnées, ne prend pas en compte le maximum de l'échantillon $X_{n,n}$. Cela entraîne une perte d'information concernant la queue de la distribution. Une justification plus formelle de cet estimateur a été apportée par Embrechts et al.(1997)[7].

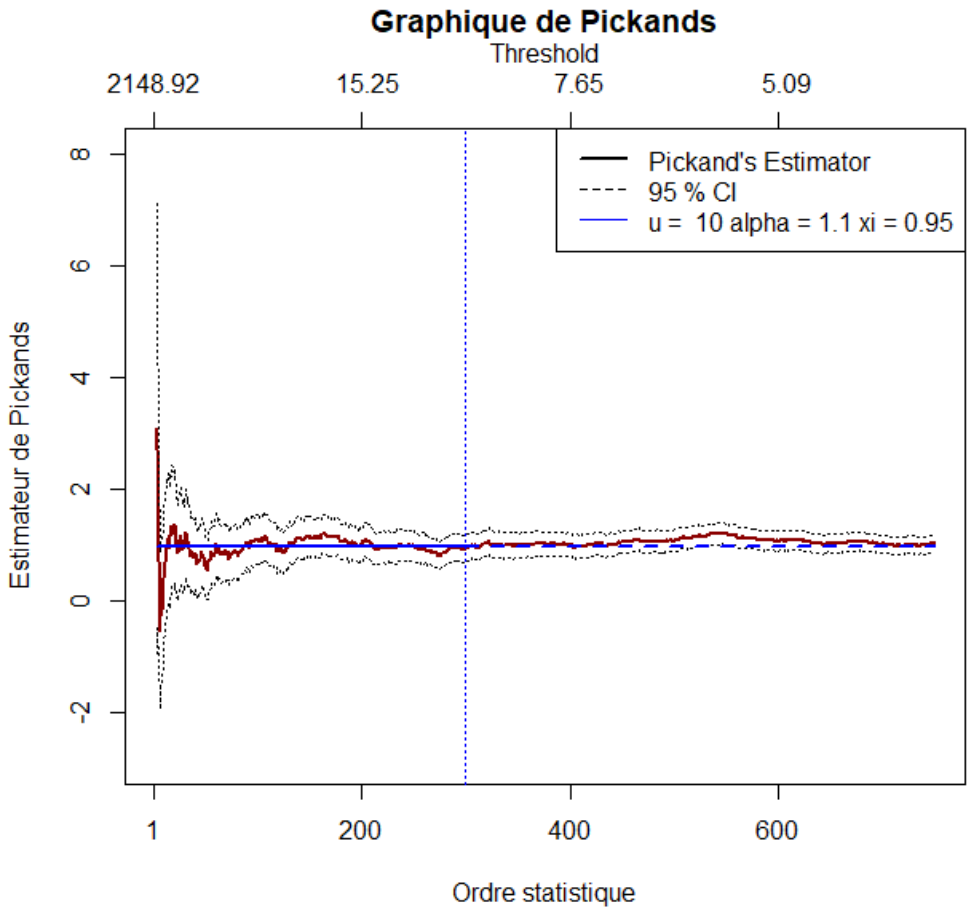

FIGURE 4.2 – L'estimateur de Pickands d'un échantillon de 3000 suivant la loi de pareto en fonction des valeurs de k_n

3.3 Estimateur DEdH

Le troisième estimateur de l'indice de queue, proposé par Dekkers, Einmahl et De Haan, est une généralisation de l'estimateur de Hill applicable à tous les domaines d'attraction. Il se définit comme suit :

$$\boxed{\hat{\xi}^{\mathrm{DEdH}}(k_n) = M_{k_n}^{(1)} + 1 - \frac{1}{2}\left(1 - \frac{(M_{k_n}^{(1)})^2}{M_{k_n}^{(2)}}\right)^{-1}} \tag{4.4}$$

où $M_{k_n}^{(r)} = \frac{1}{k_n}\sum_{i=1}^{k_n}(\ln(X_{n-i+1}) - \ln(X_{n-k_n}))^r$. La valeur de $M_{k_n}^{(1)}$ correspond à l'estimateur de Hill.

On peut démontrer que cet estimateur suit la convergence en loi suivante :

$$\sqrt{k_n}\left(\hat{\xi}^{\mathrm{DEdH}}(k_n) - \xi\right) \xrightarrow{L} N(0, \sigma_M^2) \quad \text{quand} \quad n \to \infty$$

où :

$$\sigma_M^2 = \begin{cases} 1+\xi^2 & \text{si}\xi \geq 0 \\ (1-\xi)^2(1-2\xi)\left(4 - \frac{8(1-2\xi)}{1-3\xi} + \frac{(5-11\xi)(1-2\xi)}{(1-3\xi)(1-4\xi)}\right) & \text{si } \xi < 0 \end{cases}$$

En pratique, il peut être difficile de classer ces estimateurs. Cependant, l'estimateur de Hill est souvent préféré en raison de sa variance asymptotique plus faible. Il est important de noter que l'estimateur de Hill n'est valide que pour les indices de queue $\xi > 0$, il convient donc de vérifier cette condition avant de l'utiliser. On trouve le comportement asymptotique de ce estimateur dans Deheuvels et al.(1988)[43]

Remarque :

- Estimateur de Hill : Utilise uniquement les k_n plus grandes valeurs de l'échantillon, ce qui le rend sensible au choix de k_n et moins robuste pour des données extrêmement variables.
- Estimateur de Pickands : Utilise des ratios de maximums et peut être plus robuste que Hill dans certaines situations, mais il peut encore être sensible au choix de k_n
- Estimateur DEdH : Introduit un facteur de correction basé sur le deuxième moment empirique des ratios logarithmiques, ce qui améliore la performance pour les queues très lourdes et réduit le biais de l'estimation.

Pour chacun de ces estimateurs, il est nécessaire de choisir une valeur appropriée pour k_n, qui représente le nombre de données utilisées dans l'estimation. Les méthodes pour déterminer la valeur optimale de k_n seront abordées dans la suite de notre travail.

4 Estimation des quantiles extrêmes et des périodes de retour

Souvent, l'objectif principal ne se limite pas à choisir la loi des valeurs extrêmes ou à estimer la fonction de répartition en identifiant l'indice de queue de la distribution. En réalité, le but ultime est généralement d'estimer un quantile extrême ou un niveau de retour, ainsi que de déterminer une période de retour.

4.1 Estimation de quantiles extrêmes et de niveaux de retour

Dans l'étude des événements extrêmes, on se focalise souvent sur l'estimation de la fréquence des occurrences rares. Dans divers domaines d'application tels que l'hydrologie, l'assurance, la finance ou le contrôle statistique de la qualité, l'un des objectifs principaux est de déterminer les valeurs élevées pour lesquelles la probabilité de dépassement est extrêmement faible, proche de zéro.

Nous visons à estimer le quantile d'ordre $\alpha_n = 1 - \frac{1}{t}$ de F, soit q défini par

$$\boxed{q(\alpha_n) = F^{-1}(1-\alpha_n).} \tag{4.5}$$

où F^{-1} est la fonction réciproque de la fonction de répartition. Pour des valeurs élevées de t, ce quantile est considéré comme extrême. Cela correspond au niveau qui, en moyenne, sera dépassé une seule fois tous les t périodes (ou ans). Le niveau de retour est défini comme la valeur $q(\alpha_n)$ que l'on s'attend à dépasser en moyenne une seule fois sur t périodes, soit

$$E\left(\sum_{j=1}^{t} \mathbb{I}_{(X_j > q(\alpha_n))}\right) = 1 \quad \Longleftrightarrow \quad P(X_j > q(\alpha_n)) = \frac{1}{t} \quad \text{pour } j = 1, \ldots, t$$

$$\iff 1 - F(q(\alpha_n)) = \tfrac{1}{t},$$

où $\mathbb{I}$ est la fonction indicatrice. Ainsi, estimer un niveau de retour d'ordre t équivaut à estimer un quantile extrême d'ordre $\alpha_n = 1 - \frac{1}{t}$. Étant donné que ce quantile se situe au-dessus de l'observation maximale avec une probabilité tendant vers 1 lorsque $t \to \infty$, il est impossible de simplement inverser la fonction de répartition empirique comme pour les quantiles classiques. Par conséquent, il est nécessaire d'utiliser les plus grandes observations pour estimer la fonction de répartition au-delà de l'observation maximale.

Pour estimer un quantile extrême $q(\alpha_n)$ d'ordre $\alpha_n = 1 - \frac{1}{t}$, on utilise les méthodes suivantes :

4.2 Méthode des Valeurs Extrêmes (GEV)

Cette méthode se base sur la loi asymptotique du maximum d'un échantillon, qui suit la loi des valeurs extrêmes (GEV).

Si les maxima de blocs de longueur n suivent une distribution des valeurs extrêmes généralisées (GEV) avec une fonction de répartition $H_{\xi,\mu,\beta}$, alors le niveau de retour x_{pt} est un quantile de cette distribution, défini comme suit :

$$q^{GEV}(\alpha_n) = H^{-1}_{\xi,\mu,\beta}(\alpha_n) = \mu - \frac{\beta}{\xi}\left[1 - (-\log(\alpha_n))^{-\xi}\right] \quad \text{pour } \xi \neq 0.$$

Lorsque $\xi = 0$, la distribution GEV se simplifie en distribution de Gumbel, et l'expression du quantile devient :

$$q^{GEV}(\alpha_n) = \mu - \beta \log(-\log(\alpha_n))$$

Nous pouvons estimer $q^{GEV}(\alpha_n)$ à l'aide des estimations par maximum de vraisemblance des paramètres (ξ, μ, β). Il est également possible de construire des intervalles de confiance asymétriques en utilisant la méthode du profil de vraisemblance.

Le quantile extrême est estimé en inversant la fonction de répartition de la GEV et en estimant ses paramètres. L'estimateur du quantile extrême prend une forme spécifique selon la valeur du paramètre ξ.

Le niveau de retour correspond à un quantile particulier de la distribution marginale, associé à une probabilité spécifique. Lorsque les données sont indépendantes et identiquement distribuées (iid), il est simple de calculer cette probabilité. En effet, à partir de l'équations 4.5 , on peut déduire que :

$$(\alpha_n)^{1/n} = P^{1/n}\{X_j \leq q(\alpha_n)\} = F(q(\alpha_n)),$$

ce qui nous permet d'obtenir :

$$q(\alpha_n) = F^{-1}\left((\alpha_n)^{1/n}\right).$$

L'estimateur de quantile extrême obtenu par cette méthode s'écrit sous la forme :

$$\hat{q}^{\text{GEV}}(\alpha_n) = \begin{cases} \hat{\mu} - \frac{\hat{\beta}}{\hat{\xi}}\left\{1 - [-\log(\alpha_n)]^{-\hat{\xi}}\right\} & \text{si } \hat{\xi} \neq 0 \\ \hat{\mu} - \hat{\beta}\log[-\log(\alpha_n)] & \text{si } \hat{\xi} = 0 \end{cases} \tag{4.6}$$

Pour plus de détails, nous proposons de voir McNeil(1998)[12]

4.3 Méthode des Excès (GPD)

Cette méthode, appelée méthode des excès (P.O.T.), estime le quantile extrême en utilisant la loi de Pareto généralisée (GPD). Cette loi modélise les observations qui dépassent un certain seuil élevé, et le quantile est obtenu en inversant la fonction de répartition de la GPD et en estimant ses paramètres avec les données au-dessus du seuil.

Comme défini en 2.27, la fonction de répartition au-delà du seuil u est donnée par :

$$F_u(x) = \frac{F(u+x) - F(u)}{1 - F(u)}, \quad \text{si } x \geq 0$$

Cela équivaut à :

$$F_u(x-u) = 1 - \frac{1 - F(x)}{1 - F(u)}, \quad \text{si } x \geq u$$

Ainsi, la fonction de répartition $F(x)$ peut être réécrite comme :

$$F(x) = F(u)F_u(x-u), \quad \text{si } x \geq u$$

Avec $F = 1 - F$, sachant que :

$$F(u) = 1 - P[X \leq u] = P[X > u] = \frac{N_u}{n}$$

et que $F_u \approx 1 - G_{\hat{\xi},\hat{\beta}}$ pour u assez grand, où $G_{\xi,\beta}$ est la loi GPD définie en 2.29, N_u est le nombre d'observations au-dessus du seuil u, et $\hat{\xi}$ et $\hat{\beta}$ sont les estimateurs des paramètres de la GPD. L'estimateur de $F(x)$ peut alors être écrit comme :

$$\hat{F}(x) = \frac{N_u}{n}\left(1 + \hat{\xi}\frac{x-u}{\hat{\beta}}\right)^{-\frac{1}{\hat{\xi}}}, \quad \forall\, \hat{\xi} \neq 0.$$

Par inversion, l'estimateur du quantile extrême obtenu par cette méthode s'écrit sous la forme :

$$\hat{q}^{\text{GPD}}(\alpha_n) = u + \frac{\hat{\beta}}{\hat{\xi}}\left[\left(\frac{n}{N_u}(1-\alpha_n)\right)^{-\hat{\xi}} - 1\right] \tag{4.7}$$

4.4 Détermination du seuil des extrêmes

Le choix du seuil u (ou du nombre d'excès k_n, avec n étant le nombre d'observations disponibles) est crucial non seulement dans la modélisation des montants de sinistres, mais également en météorologie et en hydrologie. En météorologie, ce seuil permet d'identifier les événements climatiques extrêmes, tels que les tempêtes ou vagues de chaleur. En hydrologie, il est essentiel pour évaluer les crues ou les sécheresses extrêmes, influençant ainsi les stratégies de gestion des ressources en eau et des infrastructures. Un seuil trop bas conduit à une mauvaise approximation par une loi de Pareto Généralisée, tandis qu'un seuil trop élevé réduit le nombre d'observations disponibles, augmentant ainsi la variance des estimateurs. Le seuil optimal est donc un compromis entre biais et variance.

En pratique, plusieurs méthodes sont disponibles pour sélectionner ce seuil optimal, y compris des méthodes graphiques et numériques. Les méthodes graphiques impliquent une certaine subjectivité de la part de l'utilisateur, contrairement aux méthodes numériques qui fournissent une estimation directe du seuil.

4.5 La fonction des excès moyens

La fonction des excès moyens, appelée "mean excess function" en anglais, représente l'espérance de l'excès d'une variable aléatoire X au-delà d'un seuil u, à condition que ce seuil soit dépassé. Elle est définie comme suit :

$$e(u) = E[X - u \mid X > u] = \frac{\int_u^{+\infty}(1 - F(x))\,dx}{1 - F(u)} \tag{4.8}$$

Pour un échantillon de variables aléatoires indépendantes et identiquement distribuées $X_1, \ldots, X_n$, une estimation empirique de cette fonction, notée $\hat{e}_n(u)$, est obtenue par :

$$\hat{e}_n(u) = \frac{\sum_{i=1}^n \max(0, x_i - u)}{\sum_{i=1}^n 1_{x_i > u}} \tag{4.9}$$

Cette estimation est simplement la somme des excès au-delà du seuil u divisée par le nombre d'observations dépassant ce seuil. Pour tracer le graphique des excès moyens, on utilise les valeurs des observations $(x_i)_{1 \leq i \leq n}$ comme seuils. Ainsi, en ordonnant les observations par ordre croissant $x_{(1)}, \ldots, x_{(n)}$, l'estimation empirique de la fonction des excès moyens est donnée par :

$$\hat{e}_n(x_{(k)}) = \frac{1}{n-k}\sum_{j=1}^{n-k}(x_{(k+j)} - x_{(k)}) \tag{4.10}$$

Le "mean excess plot" est le graphique qui représente les points $\{x_{(k)}, \hat{e}_n(x_{(k)})\}$ pour tous les $k \in [1, n]$.

Le "mean excess plot" est un outil graphique utilisé pour identifier le seuil u à partir duquel les données suivent une distribution de Pareto Généralisée. Selon une proposition, il existe une relation entre la forme du "mean excess plot" et le comportement d'une distribution de Pareto Généralisée : en effet, le graphique des excès moyens pour une telle distribution est linéaire par rapport à u.

Proposition 1 :

Soit X une variable aléatoire suivant une distribution de Pareto Généralisée (GPD) avec paramètres γ et σ. La fonction des excès moyens de X est donnée par :

$$e(u) = \frac{\gamma}{1-\gamma}u + \frac{\sigma}{1-\gamma}$$

Démonstration :

Supposons que $\gamma < 1$. La fonction des excès moyens $e(u)$ est définie comme :

$$e(u) = \frac{1}{P(X > u)}\int_u^{+\infty} P(X > x)\,dx$$

On a :

$$P(X > u) = \left(1 + \frac{\gamma u}{\sigma}\right)^{-\frac{1}{\gamma}}$$

Et :

$$\int_u^{+\infty} P(X > x)\, dx = \int_u^{+\infty} \left(1 + \frac{\gamma x}{\sigma}\right)^{-\frac{1}{\gamma}} dx$$

En utilisant le changement de variable $y = 1 + \frac{\gamma x}{\sigma}$, on obtient :

$$\int_u^{+\infty} P(X > x)\, dx = \frac{\sigma}{1-\gamma}\left[\left(1 + \frac{\gamma u}{\sigma}\right)^{1-\frac{1}{\gamma}}\right]$$

Ainsi, la fonction des excès moyens est :

$$e(u) = \frac{\sigma}{1-\gamma}\left[\left(1 + \frac{\gamma u}{\sigma}\right)^{1-\frac{1}{\gamma}}\right] \times \left(1 + \frac{\gamma u}{\sigma}\right)^{\frac{1}{\gamma}} = \frac{\sigma}{1-\gamma}\left(1 + \frac{\gamma u}{\sigma}\right)$$

Ce qui conduit à l'expression finale :

$$e(u) = \frac{\gamma}{1-\gamma}u + \frac{\sigma}{1-\gamma}$$

NB : Si la fonction empirique des excès moyens montre un comportement linéaire à partir d'un certain seuil positif, cela indique que les données au-delà de ce seuil suivent une distribution de Pareto Généralisée. Il est alors important d'identifier le seuil à partir duquel le "mean excess plot" devient approximativement linéaire.

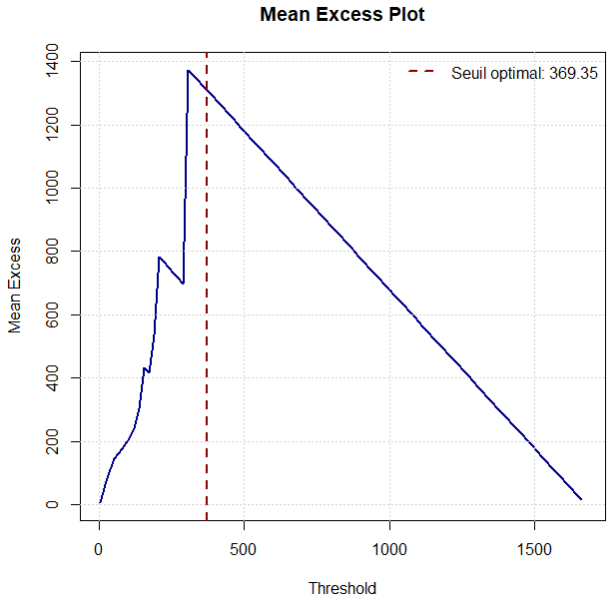

FIGURE 4.3 – Identification du Seuil Optimal pour la Distribution de Pareto Généralisée à l'aide du Mean Excess Plot

Remarque :

Retenons que la fonction des excès moyens pour

une distribution de Pareto avec un paramètre $\alpha > 0$, qui appartient au domaine d'attraction de Fréchet avec un indice des valeurs extrêmes $\xi = \frac{1}{\alpha}$, est exprimée comme suit : $e(u) = \frac{k+u}{\alpha-1}$.

En revanche, pour une distribution de Weibull avec des paramètres $\lambda > 0$ et $\tau > 0$, appartenant au domaine d'attraction de Gumbel, la fonction des excès moyens est donnée par : $e(u) = \frac{u^{1-\tau}}{\lambda\tau}$.

5 Sélection du nombre d'observations extrêmes k_n

L'estimation de l'indice des valeurs extrêmes pour une distribution à queue lourde dépend du choix du nombre de statistiques d'ordre extrêmes, noté k. Ce choix est complexe car un k trop petit augmente la variance de l'estimateur, tandis qu'un k trop grand introduit du biais. Un bon équilibre entre variance et biais est essentiel pour minimiser l'erreur quadratique moyenne. Cette partie se concentre sur l'estimateur de Hill et explore des méthodes pour déterminer un k optimal, afin d'identifier précisément où commence la queue de la distribution.

5.1 Méthode Graphique

Estimer l'indice de queue ξ des distributions à queue lourde est crucial dans de nombreuses applications. Avant de procéder à des analyses numériques, une méthode graphique universelle est recommandée. Cette approche vise à sélectionner un nombre approprié de statistiques d'ordre extrêmes.

La méthode consiste à tracer un graphique avec les points $(k, \hat{\xi}_n(k))$ pour $k = 1, \ldots, n$, où $\hat{\xi}_n(k)$ est un estimateur discuté dans le chapitre précédent. Le choix de k est délicat : il doit être suffisamment grand pour éviter de fortes fluctuations dues à un échantillon trop réduit, mais pas trop grand pour éviter un biais causé par les valeurs centrales de l'échantillon. Un graphique stable aide à déterminer cette valeur optimale de k, comme illustré dans la figure ci-dessous.

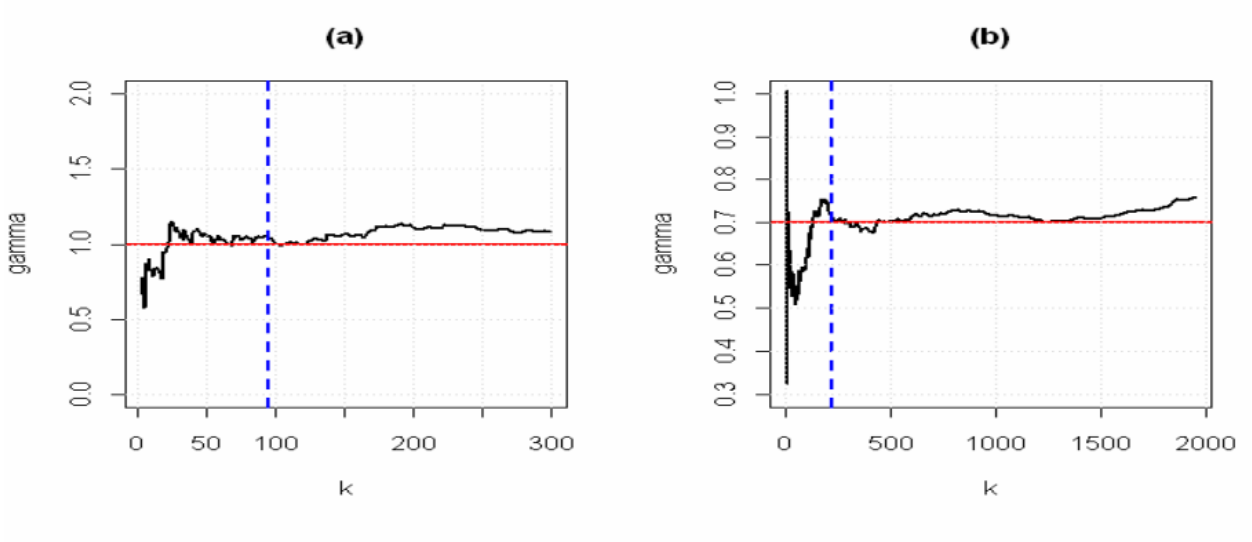

FIGURE 4.4 – Estimateur de Hill de l'indice de valeur extrême (IVE) pour (a) une distribution Pareto standard basée sur 300 échantillons de taille 3000 et (b) les données de Danish Fire. La ligne horizontale représente la valeur estimée de l'indice de queue, tandis que la ligne verticale indique le nombre optimal de statistiques d'ordre.

Sur la Figure 4.4, on observe que pour la distribution Pareto standard, le nombre optimal de statistiques d'ordre k semble se stabiliser autour de 80, tandis que pour les données de réclamations danoises, k est autour de 220. En utilisant l'estimateur de Hill, nous obtenons $\hat{\xi}_{97,3000} = 0.9874$ pour la distribution Pareto standard et $\hat{\xi}_{220,2167} = 0.6899$ pour les données danoises. Nous introduisons également une nouvelle méthode graphique, appelée sum-plot, proposée par Sousa(2002)[50].

5.2 Méthode du Sum-Plot

La méthode du sum-plot, proposée par Sousa dans sa thèse de doctorat en 2002, est une approche graphique pour déterminer la valeur optimale de k. Théoriquement, le graphe des points (k, S_k) pour $1 \leq k < n$ devrait être une ligne droite. Une fois k déterminé, cette valeur est utilisée dans l'estimateur de Hill pour calculer l'indice de queue ξ. Sousa a conclu que, indépendamment de la valeur de l'indice de queue, qu'elle soit $0 < \xi < 1/2$ ou autre, la méthode reste efficace.

Soit la variable aléatoire

$$S_k = \sum_{j=1}^{k} jV_j = \sum_{j=1}^{k} j(\log X_{j,n}).$$

Où $X_{1,n} \leq X_{2,n} \leq \ldots \leq X_{k+1,n}$ sont les statistiques d'ordre correspondantes.

Si nous choisissons k de sorte que $X_{k+1,n}$ soit suffisamment grand pour satisfaire la condition suivante

$$\bar{F}(x) = 1 - F(x) = x^{-1/\xi}\ell(x), \tag{4.11}$$

où ℓ est une fonction à variation lente, pour x $> X_{k+1,n}$ nous avons $S_k \sim \xi k$. L'approximation a montré que la pente de la ligne sur le graphique est ξ. Sousa a démontré que ξ peut être estimé à l'aide d'un modèle de régression linéaire.

$$S_i = \beta_0 + \beta_1 i + \xi_i, \quad i = 1, \ldots, k$$

Il est évident que la valeur estimée de $\hat{\xi}$ est la pente $\hat{\beta}_1$ du modèle de régression, déterminée par l'estimation des moindres carrés généralisés :

$$\hat{\xi} = \hat{\beta}_1 = \frac{k}{k+1}\hat{\xi}_n^H + \frac{k}{k+1}\log X_{1,n} \tag{4.12}$$

Si β_0, alors $\hat{\xi} = \hat{\xi}_n^H$.

La méthode proposée a montré son efficacité pour les distributions Pareto et les inverses de gamma inversés. Cependant, bien que la méthode graphique soit utile pour choisir k, elle est souvent subjective. Pour obtenir une sélection plus précise et automatique de k, nous présentons des méthodes alternatives qui visent à minimiser l'erreur quadratique moyenne dans l'estimation de l'indice des valeurs extrêmes.

5.3 Procédures Adaptatives

Lorsqu'on travaille avec un échantillon aléatoire de taille finie, choisir le nombre optimal d'extrêmes supérieures n'est pas simple. Le processus de détermination de k optimal est complexe

car il ne dépend pas uniquement de la taille de l'échantillon et de l'indice des valeurs extrêmes, mais aussi de paramètres inconnus qui caractérisent F, tels que le paramètre de second ordre.

Pour surmonter ce défi, de nombreux algorithmes et procédures adaptatives ont été proposés pour calculer $\hat{k}_{\text{opt}}$, avec l'objectif que

$$\frac{\hat{k}_{\text{opt}}}{k_{\text{opt}}} \to 1 \quad \text{lorsque} \quad n \to \infty.$$

L'estimation associée, $\hat{\xi}_n(\hat{k}_{\text{opt}})$, est asymptotiquement aussi efficace que $\hat{\xi}_n(k_{\text{opt}})$. Nous allons décrire certaines des méthodes les plus reconnues pour sélectionner le nombre approprié de statistiques maximales afin d'assurer une estimation précise.

5.4 Minimisation de l'erreur quadratique moyenne asymptotique

Pour assurer la précision de l'estimateur $\hat{\xi}^H_{(k(n),n)}$, il est important de calculer l'erreur quadratique moyenne (AMSE), qui est fonction de k :

$$\text{AMSE}(\hat{\xi}) = \mathbb{E}_\infty[(\hat{\xi} - \xi)^2] = \text{Var}(\hat{\xi}) + \text{Biais}^2(\hat{\xi}), \tag{4.13}$$

où $\mathbb{E}_\infty$ représente l'espérance mathématique sous la distribution asymptotique. Le k_{opt} est donc déterminé par :

$$k_{\text{opt}} = \arg\min_k \text{AMSE}(\hat{\xi}).$$

Le choix optimal de k est celui qui minimise la MSE. Pour l'estimateur de Hill appliqué aux fonctions appartenant au domaine d'attraction maximal de Fréchet, $\xi > 0, \quad (\rho = max(-1, -\xi))$, De Haan et al.(1998)[23] ont proposé de sélectionner le nombre d'observations k_{opt} qui minimise l'erreur quadratique moyenne de l'estimateur de Hill selon la formule suivante :

$$k_{opt} \approx \begin{cases} \left\{\frac{(1+\xi)^2}{2\xi}\right\}^{\frac{1}{(2\xi+1)}} n^{\frac{2\xi}{(2\xi+1)}} & \text{si } 0 < \xi < 1 \\ 2\left(\frac{n}{3}\right)^{\frac{2}{3}} & \text{si } \xi = 1 \\ 2n^{\frac{2}{3}} & \text{si } \xi > 1 \end{cases} \tag{4.14}$$

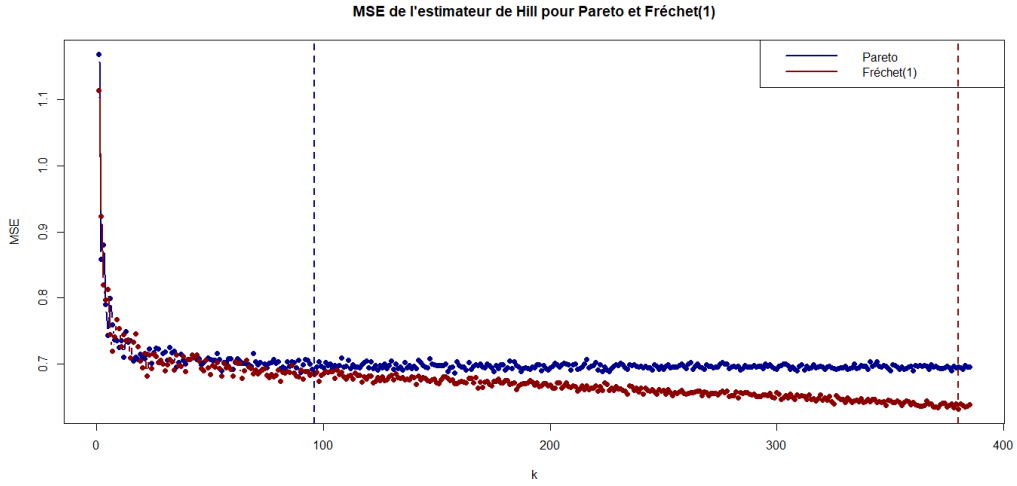

FIGURE 4.5 – MSE de l'estimateur de Hill pour l'IVE distribution de Pareto standard et de Fréchet(1), baée sur 300 Èchantillons de 3000 observations. Les lignes verticales correspondent au minimum du MSE

6 Conclusion

Dans ce chapitre, nous avons exploré diverses méthodes d'estimation de l'indice des valeurs extrêmes (EVI). En complément, des techniques pour le choix optimal de k , essentiel à la robustesse des estimations, ont été examinées, mettant en lumière l'impact sur la précision des estimateurs. L'estimation des périodes de retour a également été abordée, soulignant l'importance d'une bonne maîtrise de ces méthodes pour des applications pratiques, notamment dans les domaines de l'hydrologie et de la gestion des risques.

5 Illustration numérique

Contents

1 Introduction

Dans ce chapitre, nous avons entrepris une simulation extensive pour évaluer la performance de nos modèles (4.2, 4.3 et 4.4) en fonction de divers paramètres. Avec le logiciel RStudio, nous avons réalisé les simulations pour des échantillons de tailles différentes (n = 200, 500 et 800) avec 500 replications (R = 500) en utilisant une distribution de Pareto, connue pour sa pertinence dans l'analyse des valeurs extrêmes en varient l'index des valeurs extrêmes $\xi = 0.5, 1, 1.5$, qui détermine l'épaisseur de la queue de la distribution.Nous avons introduit la troncature (gauche) de $(10\%, 20\% \quad \text{et} \quad 90\%)$ et du nombres $k = 10, 20, 30$ de valeurs extrêmes afin d'analyser leur impact sur les performances des estimateurs.Dans le cas de troncature, cela éauivaut à dire que nous ignorons les $(10\%, 20\%, 30\% \quad \text{et} \quad 90\%)$ des valeurs les plus faibles de l'échantillon, en ne conservant que les valeurs au-dessus du (10e, 20e, 30e et 90e) percentile respectivement. Cela simule la situation où nous n'avons accès qu'à la queue supérieure des données.
Rappelons que le k_{opt} est défini par :

$$k_{\text{opt}} = \arg\min_{k} \text{MSE}(\hat{\xi}).$$

Où

$$\text{MSE}(\hat{\xi}) = \mathbb{E}\left[(\hat{\xi}_n - \xi)^2\right] = \text{Var}(\hat{\xi}) + (\text{Bias}(\hat{\xi}))^2 \tag{5.1}$$

Avec $\text{Bias}(\hat{\xi}) = \mathbb{E}[\hat{\xi}] - \xi$ et $\text{Var}(\hat{\xi}) = \mathbb{E}[\hat{\xi}^2] - (\mathbb{E}[\hat{\xi}])^2$

Les biais et les erreurs quadratiques moyennes (MSE) définis dans l'équation 5.1 des estimateurs ont été calculés pour chaque poucentage de troncature et du nombre k de valeurs extrêmes, permettant une évaluation comparative de leur précision et de leur robustesse. L'objectif de cette étude est de fournir des recommandations sur le choix des estimateurs et des paramètres optimaux pour les analyses de valeurs extrêmes dans des conditions variées, contribuant ainsi à une meilleure compréhension et gestion des risques associés aux événements rares et extrêmes.

2 Présentation des résultats numériques par simulation

ξ		1								
$\%Tronc$		10%								
k		10			20			30		
		$\hat{\xi}^H$	$\hat{\xi}^P$	$\hat{\xi}^{DEdH}$	$\hat{\xi}^H$	$\hat{\xi}^P$	$\hat{\xi}^{DEdH}$	$\hat{\xi}^H$	$\hat{\xi}^P$	$\hat{\xi}^{DEdH}$
200		1.205	3.603	1.191	1.284	-0.235	1.269	1.531	1.275	1.380
	Biais	-0.02	-0.52	-0.23	-0.01	-0.02	-0.14	-0.00	-0.23	-0.10
	MSE	0.10	2.50	0.21	0.05	1.07	0.11	0.04	0.68	0.07
500		0.829	0.368	0.687	0.942	1.496	0.717	1.029	-0.530	0.780
	Biais	0.00	-0.63	-0.20	0.01	-0.00	-0.11	-0.01	-0.15	-0.08
	MSE	0.10	3.04	0.20	0.05	1.04	0.10	0.04	0.66	0.06
800		0.942	-0.82	0.593	0.804	2.018	0.815	0.958	1.007	0.776
	Biais	-0.01	-0.62	-0.21	-0.01	0.02	-0.12	0.00	-0.19	-0.09
	MSE	0.10	2.93	0.22	0.05	0.98	0.11	0.03	0.76	0.07
$\%Tronc$		20%								
200		1.205	3.603	1.191	1.284	-0.235	1.269	1.531	1.275	1.380
	Biais	-0.02	-0.52	-0.23	-0.01	-0.02	-0.14	0.00	-0.23	-0.10
	MSE	0.10	2.50	0.21	0.05	1.07	0.11	0.04	0.68	0.07
500		0.829	0.368	0.687	0.942	1.496	0.717	1.029	-0.530	0.780
	Biais	0.00	-0.63	-0.20	0.01	-0.00	-0.11	-0.01	-0.15	-0.08
	MSE	0.10	3.04	0.20	0.05	1.04	0.10	0.04	0.66	0.06
800		0.942	-0.82	0.593	0.804	2.018	0.815	0.958	1.007	0.776
	Biais	-0.01	-0.62	-0.21	-0.01	0.02	-0.12	-0.00	-0.19	-0.09
	MSE	0.10	2.93	0.22	0.05	0.98	0.11	0.03	0.76	0.07

TABLE 5.1 – Résultats de simulations pour différentes tailles d'échantillon de la loi de pareto d'IVE $\xi = 1$, pourcentages de troncature $\%Tronc$ et de valeurs de k

ξ		1								
$\%Tronc$		90%								
k		10			20			30		
		$\hat{\xi}^H$	$\hat{\xi}^P$	$\hat{\xi}^{DEdH}$	$\hat{\xi}^H$	$\hat{\xi}^P$	$\hat{\xi}^{DEdH}$	$\hat{\xi}^H$	$\hat{\xi}^P$	$\hat{\xi}^{DEdH}$
200		1.205	3.603	1.191	0	0	0	0	0	0
	Biais	-0.02	-0.52	-0.23	0.00	0.00	0.00	0.00	0.00	0.00
	MSE	0.10	2.50	0.21	0.00	0.00	0.00	0.00	0.00	0.00
500		0.829	0.368	0.687	0.942	1.496	0.717	1.029	-0.530	0.780
	Biais	0.00	-0.63	-0.20	0.01	-0.00	-0.11	-0.01	-0.15	-0.08
	MSE	0.10	3.04	0.20	0.05	1.04	0.10	0.04	0.66	0.06

TABLE 5.2 – Résultats de simulations pour différentes tailles d'échantillon de la loi de pareto d'IVE $\xi = 1$, de valeurs de k et pourcentage de troncature $\%Tronc = 90$

ξ		0.5								
$\%Tronc$		10%								
k		10			20			30		
		$\hat{\xi}^{H}$	$\hat{\xi}^{P}$	$\hat{\xi}^{DEdH}$	$\hat{\xi}^{H}$	$\hat{\xi}^{P}$	$\hat{\xi}^{DEdH}$	$\hat{\xi}^{H}$	$\hat{\xi}^{P}$	$\hat{\xi}^{DEdH}$
200		0.602	2.724	0.641	0.642	-0.739	0.635	0.765	0.530	0.642
	Biais	-0.01	-0.49	-0.17	0.00	-0.04	-0.11	0.00	-0.22	-0.08
	MSE	0.03	2.12	0.12	0.01	0.86	0.07	0.01	0.57	0.04
500		0.414	-0.113	0.285	0.471	0.955	0.276	0.514	-0.956	0.2798
	Biais	0.00	-0.60	-0.16	0.00	-0.03	-0.09	0.00	-0.14	-0.06
	MSE	0.02	2.62	0.12	0.01	0.83	0.06	0.01	0.55	0.04
800		0.471	-1.243	0.147	0.402	1.444	0.421	0.479	0.546	0.319
	Biais	-0.01	-0.59	-0.16	0.00	-0.01	-0.09	0.00	-0.18	-0.07
	MSE	0.02	2.52	0.13	0.01	0.76	0.07	0.01	0.64	0.04
ξ		1.5								
$\%Tronc$		60%								
200		1.807	4.550	1.740	1.926	0.240	1.903	2.296	2.010	2.117
	Biais	-0.02	-0.54	-0.28	-0.01	0.01	-0.17	0.00	-0.24	-0.11
	MSE	0.23	3.06	0.34	0.12	1.37	0.18	0.08	0.84	0.12
500		1.243	0.839	1.088	1.414	2.046	1.158	1.544	-0.126	1.281
	Biais	0.01	-0.66	-0.25	0.01	0.03	-0.13	-0.01	-0.16	-0.10
	MSE	0.22	3.62	0.33	0.12	1.36	0.17	0.08	0.83	0.11
800		1.413	-0.421	1.039	1.207	2.610	1.208	1.437	1.469	1.234
	Biais	-0.02	-0.65	-0.26	-0.01	0.05	-0.15	-0.01	-0.21	-0.10
	MSE	0.22	3.51	0.36	0.12	1.30	0.19	0.07	0.94	0.12

TABLE 5.3 – Résultats de simulations pour différentes tailles d'échantillon de la loi de pareto d'IVE $\xi = (0.5, 1.5)$, pourcentages de troncature $\%Tronc$ et de valeurs de k

2.1 Interprétation des résultats

— Précision des Estimations : Les estimateurs Hill et DEdH tendent à être plus précis avec des échantillons plus grands et des valeurs plus élevées de k. L'estimateur Pickands peut montrer des biais et des MSE plus élevés, surtout avec une troncature importante.
— Impact de la Troncature : Une troncature plus élevée tend à augmenter la MSE et dégrader la performance des estimateurs, particulièrement pour l'estimateur Pickands.
— Optimisation des Paramètres : L'augmentation de la taille de l'échantillon et le choix de valeurs plus élevées pour k améliorent généralement la précision des estimateurs.

Nous résumons ceci pour retenir que le choix de l'estimateur dépendra des conditions spécifiques des données (taille de l'échantillon, pourcentage de troncation) et de l'importance du biais et de la MSE dans l'application pratique de l'analyse des valeurs extrêmes.

En considérant les performances des différents estimateurs sous diverses conditions, il est maintenant pertinent d'appliquer ces résultats à des cas concrets. Dans la suite de notre travail, nous allons utiliser ces trois modèles ($\hat{\xi}^H$, $\hat{\xi}^P$ et $\hat{\xi}^{DEdH}$) pour analyser des données réelles, afin de valider leur efficacité et leur robustesse dans des situations pratiques et réelles.

3 Application à des Données réelles

Pour notre étude hydrologique, nous avons utilisé des données de débit provenant du site "01594440" du USGS (United States Geological Survey). Les données ont été téléchargées via le package dataRetrieval de R, qui permet un accès direct aux bases de données du National Water Information System (NWIS).

Les paramètres de recherche étaient les suivants :

1. Numéro de site (siteNumber) : "01594440", identifiant unique de la station de mesure.
2. Code de paramètre (parameterCd) : "00060", correspondant aux mesures de débit (discharge) en pieds cubes par seconde (cfs).
3. Période d'étude : du 1er janvier 2010 au 1er janvier 2020.

Les données récupérées incluent des valeurs moyennes journalières de débit, codées sous la forme **X_00060_00003**, et des codes de qualification indiquant la qualité des données.

Nous donnons un exemple des cinq premières lignes des données collectées :

Code_Agence	Num_Site	Date	Débits	Approbation
USGS	1594440	2010-01-01	678	A
USGS	1594440	2010-01-02	596	A
USGS	1594440	2010-01-03	455	A
USGS	1594440	2010-01-04	428	A
USGS	1594440	2010-01-05	437	A
USGS	1594440	2010-01-06	418	A

TABLE 5.4 – Aperçu des données de débits (discharge_data_read)

Nous avons visualisé les données de débit pour observer les variations temporelles en utilisant la bibliothèque **ggplot2** de R, ce qui nous a permis de mieux comprendre les tendances et les comportements du débit au cours de la période d'étude.

Cette analyse des données de débit est essentielle pour comprendre les variations hydrologiques et pour appliquer les méthodes d'extrêmes pour l'estimation des indices d'extrême et des niveaux

de retour. Le tableau 5.5 et 5.6 nous donnent respectivement les statistiques descriptives de nos données et le résumé de notre variable d'interet (Débits)

Mean	**Std. Dev**	**Min**	**Q1**
13756.03	15924.74	1120.00	3990.00
Median	**Q3**	**Max**	**MAD**
8140.00	17200.00	148000.00	7694.69
IQR	**CV**	**Skewness**	**SE.Skewness**
13155.00	1.16	3.06	0.06
Kurtosis	**N.Valid**	**Pct.Valid**	
14.25	1827.00	100.00	

TABLE 5.5 – Statistiques descriptives des débits (en unités appropriées)

	Min.	**1er Qu.**	**Médiane**	**Moyenne**	**3e Qu.**	**Max.**
Débits	1120	3995	8140	13756	17150	148000

TABLE 5.6 – Résumé statistique des débits (en unités appropriées)

Pour ces données réelles, il est important de verifier que nos données suivent la loi de pareto nous permettant d'assurer que les outils et modèles statistiques appropriés sont utilisés pour l'analyse, la modélisation, la gestion des risques, et la prise de décision. Cela conduit à des résultats plus précis et fiables, particulièrement dans des domaines où les événements extrêmes jouent un rôle crucial exactement comme dans notre cas.

Nous allons considérer à la fois les méthodes graphiques et les tests formels pour obtenir une conclusion robuste.

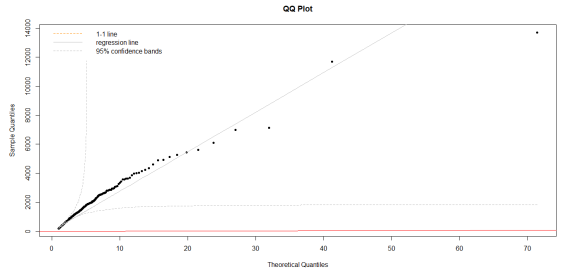

FIGURE 5.1 – Quantile-Quantile Plot des Débits Observés

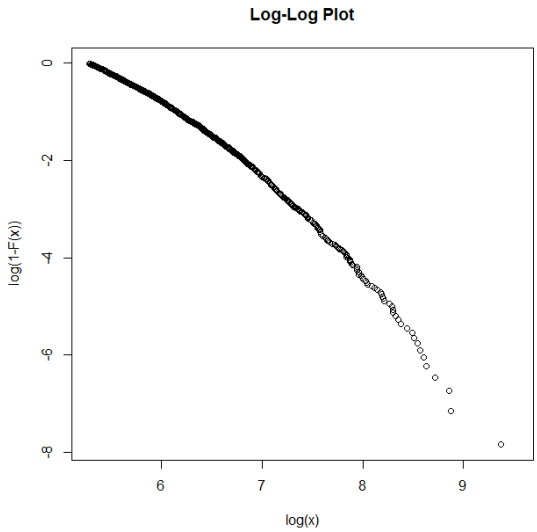

FIGURE 5.2 – Graphique Log-Log des Débits Observés

Dans les deux graphiques, on peut remarquer que nos données semblent suivre une distribution à queue lourde comme la loi de Pareto sur une large plage, mais les valeurs extrêmes dévient

dans les deux cas (écart aux queues de la distribution dans le QQ plot, déviation aux grandes valeurs de x dans le log-log plot).

Les deux graphiques suggèrent que la loi de Pareto est une bonne approximation pour une grande partie de vos données. Cependant, des écarts apparaissent aux queues de la distribution.

Les deux méthodes (QQ plot et Log-log plot) montrent que les valeurs extrêmes dans nos données ne sont pas bien modélisées par la distribution de Pareto avec le paramètre de forme actuel.

Dans la suite du travail nous allons explorer d'autres distributions et ajustement comme des critères statistiques comme l'AIC (Akaike Information Criterion) pour mieux capturer le comportement des queues de la distribution. Il s'agira des distributions de :

- Fréchet (souvent utilisée pour les queues lourdes, comme la Pareto),
- Weibull (pour des queues plus légères),
- Gumbel (pour des queues exponentiellement décroissantes).

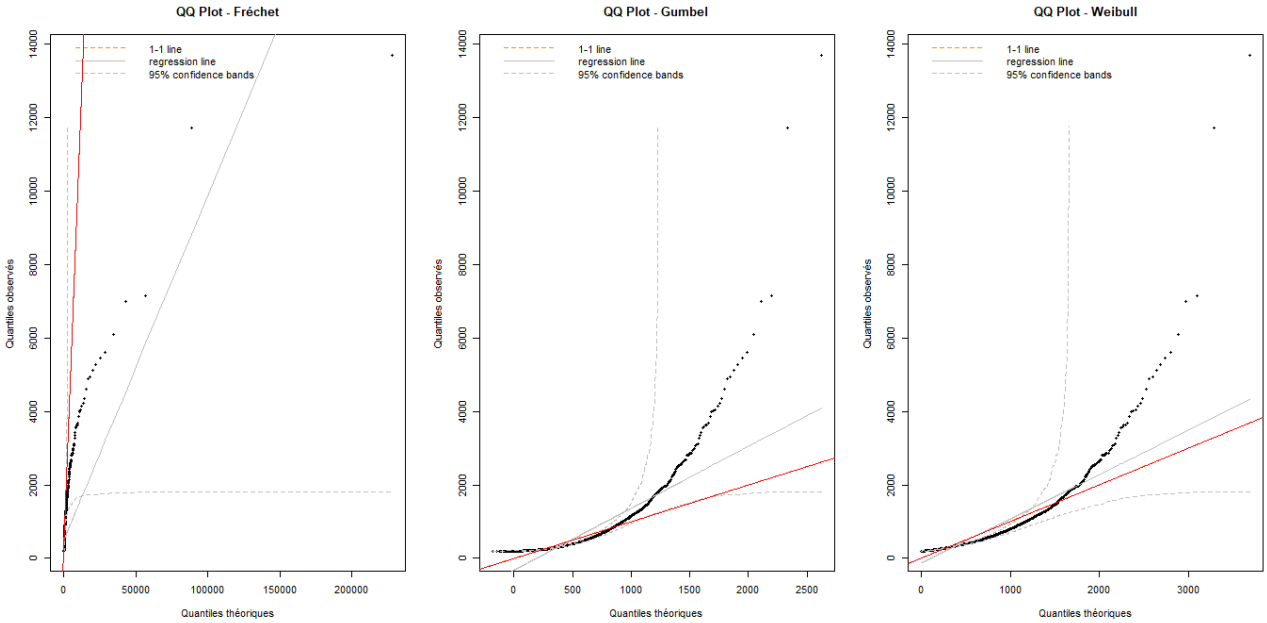

FIGURE 5.3 – QQ plot des Débits des lois de Fréchet, Gumbel et de Weibull

Model	AIC
Fréchet	35033.83
Gumbel	37237.91
Weibull	37286.81

TABLE 5.7 – AIC des modèles ajustés

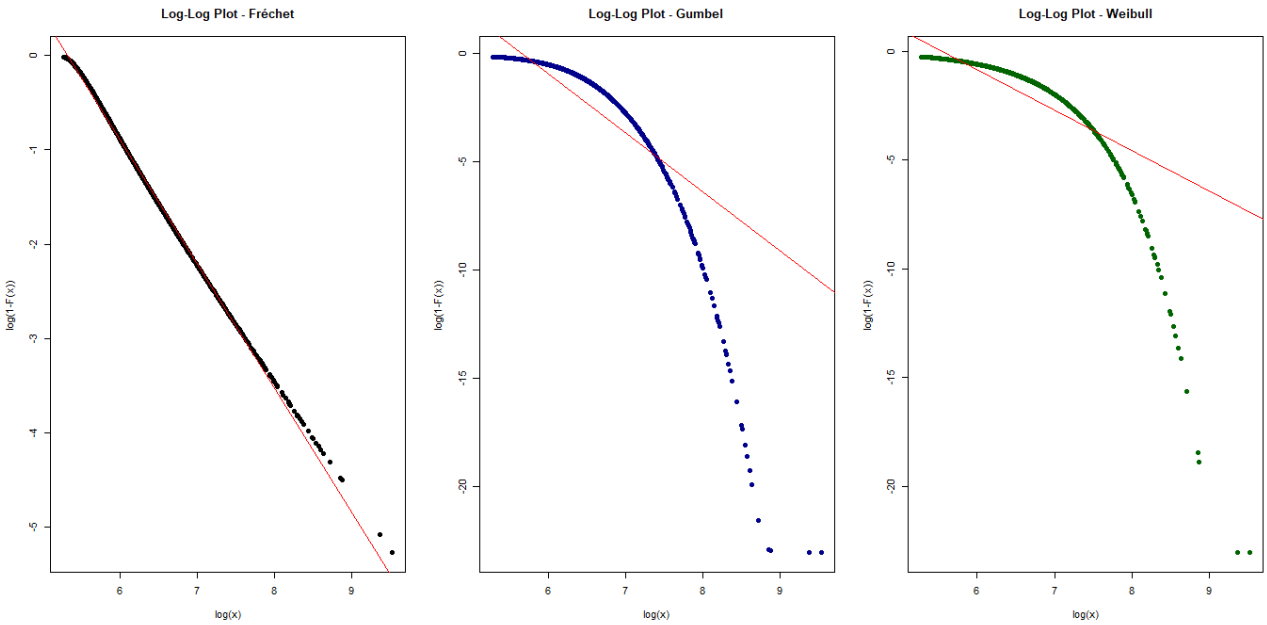

FIGURE 5.4 – Log-log plot des Débits des lois de Fréchet, Gumbel et de Weibull

Sur la Figure 5.5 nous pouvons observr que les données suivent correctement la droite ainsi que dans la Figure 5.4 et que l'AIC voir le tableau 5.7 est le plus bas pour le modèle de Fréchet. Cela indique que le modèle de Fréchet est le meilleur ajustement parmi les distributions testées pour nos données. Les bons résultats dans le Log-Log plot et le plus bas AIC suggèrent que le modèle de Fréchet décrit le mieux la distribution des extrêmes dans nos données et que les modèles de Gumbel et Weibull sont moins appropriés.

Choix du seuil optimal

Dans cette partie nous allons utiser le Mean Excess Plot pour identifier le seuil optimal de nos données. Le graphe 5.5 indique que le seuil optimal est de 7946.48 après une troncature à gauche de 20%. Donc seuil au-delà duquel les valeurs observées peuvent être modélisées.

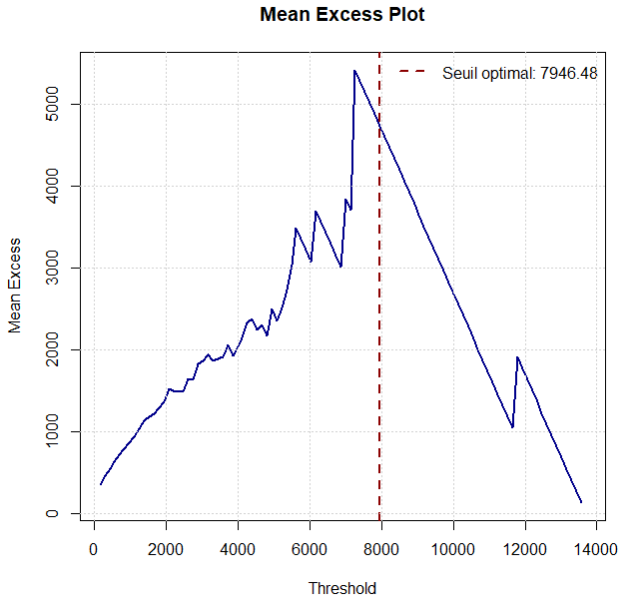

FIGURE 5.5 – Mean Excess Plot d'identification du Seuil Optimal à partir duquel les données suivent la Distribution de Pareto Généralisée

3.1 Présentation des résultats

Dans cette section nous avons calculer les estimateurs de nos données à savoir $\hat{\xi}^H$, $\hat{\xi}^P$, $\hat{\xi}^{DEdH}$ ainsi que les biais et MSE en fonction des valeurs de $k = 10, 20, 30$ et $\xi = 0.5, 1, 1.5$ avec $\%Tronc = 80$. Nous pouvons voir les résultats dans le tableau 5.8.

ξ						0.5				
$\%Tronc$						80%				
k			10			20			30	
		$\hat{\xi}^{H}$	$\hat{\xi}^{P}$	$\hat{\xi}^{DEdH}$	$\hat{\xi}^{H}$	$\hat{\xi}^{P}$	$\hat{\xi}^{DEdH}$	$\hat{\xi}^{H}$	$\hat{\xi}^{P}$	$\hat{\xi}^{DEdH}$
730		0.325	-0.169	0.378	0.376	-0.095	0.302	0.424	0.052	0.272
	Biais	-0.17	-0.67	-0.12	-0.12	-0.60	-0.20	-0.08	-0.45	-0.23
	MSE	0.03	0.45	0.01	0.02	0.35	0.04	0.01	0.20	0.05
ξ						1				
730		0.325	-0.169	0.378	0.376	-0.095	0.302	0.424	0.052	0.272
	Biais	-0.67	-1.17	-0.62	-0.62	-1.10	-0.70	-0.58	-0.95	-0.73
	MSE	0.46	1.37	0.39	0.39	1.20	0.49	0.33	0.90	0.53
ξ						1.5				
730		0.325	-0.169	0.378	0.376	-0.095	0.302	0.424	0.052	0.272
	Biais	-1.17	-1.67	-1.12	-1.12	-1.60	-1.20	-1.08	-1.45	-1.23
	MSE	1.38	2.79	1.26	1.26	2.55	1.43	1.16	2.10	1.51

TABLE 5.8 – Résultats des calculs des estimateurs (Hill, Pickands et DEdH), des biais et MSE en fonction d'IVE $\xi = (0.5, 1, 1.5)$, pourcentages de troncature $\%Tronc$ et de valeurs de k

3.2 Interprétation des résultats

Les résultats du tableau montrent que pour l'estimation de l'indice de valeur extrême sous un scénario de troncature à gauche de 80%, la méthode de Hill ($\hat{\xi}^{H}$) et la méthode DEdH ($\hat{\xi}^{DEdH}$) fournissent des estimations plus proches de la vraie valeur comparativement à la méthode de Pickands ($\hat{\xi}^{P}$). Toutefois, toutes les méthodes présentent un certain biais et une erreur quadratique moyenne significative, suggérant la nécessité de prudence dans l'interprétation des résultats, surtout avec des tailles d'échantillons et des valeurs de k variables. La méthode de Pickands semble être moins performante dans ce contexte spécifique, montrant une plus grande sensibilité à la troncature appliquée.

Conclusion et Perspectives

L'estimation de l'indice des valeurs extrêmes (EVI) sous troncature est un domaine d'étude essentiel dans l'analyse des données extrêmes, notamment dans les situations où les échantillons sont tronqués, limitant ainsi l'accès aux observations complètes. Les estimateurs classiques tels que ceux de Hill, Pickands et DEdH ont montré leur efficacité dans diverses situations, mais doivent être adaptés pour tenir compte de la troncature. En particulier, la troncature à gauche (ou à droite) modifie la nature des observations disponibles, nécessitant des ajustements spécifiques des méthodes d'estimation pour obtenir des résultats fiables. Les simulations réalisées sur des données tronquées ont permis de démontrer que ces estimateurs, bien que biaisés dans certains cas, peuvent fournir des estimations correctes de l'EVI lorsqu'ils sont correctement calibrés.

Les études empiriques montrent que l'ajustement des paramètres, tels que la proportion de troncature et le choix optimal du nombre d'observations extrêmes (k), jouent un rôle crucial dans l'efficacité des estimations. L'utilisation de méthodes comme l'analyse de biais et de la MSE (Mean Squared Error) nous a aidé à mieux comprendre les performances des différents estimateurs dans ces situations tronquées.

Une voie d'amélioration consiste à développer des estimateurs qui intègrent mieux les effets de la troncature, réduisant ainsi le biais introduit par la perte de données. Cela pourrait inclure des méthodes basées sur des approches bayésiennes ou des techniques de pondération des données.

Bien que la distribution de Pareto soit couramment utilisée pour étudier les valeurs extrêmes, il serait pertinent d'appliquer les méthodes d'estimation sous troncature à d'autres distributions comme celles de Weibull ou Fréchet, afin de généraliser les résultats et les rendre applicables à un éventail plus large de données réelles.

L'une des perspectives majeures est d'appliquer ces méthodes à des données issues de domaines variés comme la finance, la météorologie etc. En particulier, l'application à des phénomènes rares mais critiques,comme les risques financiers extrêmes ou les événements climatiques, offrirait des résultats très concrets et utiles pour la prise de décision.

Annexe : Démonstrations

Démonstration des calculs de la Variance et du MSE

Variance

La variance d'un estimateur $\hat{\xi}$ est définie par :

$$\mathrm{Var}(\hat{\xi}) = \mathbb{E}\left[(\hat{\xi} - \mathbb{E}[\hat{\xi}])^2\right] \tag{2}$$

La démonstration est la suivante :

$$\begin{aligned}
\mathrm{Var}(\hat{\xi}) &= \mathbb{E}\left[(\hat{\xi} - \mathbb{E}[\hat{\xi}])^2\right] \\
&= \mathbb{E}\left[\hat{\xi}^2 - 2\hat{\xi}\mathbb{E}[\hat{\xi}] + (\mathbb{E}[\hat{\xi}])^2\right] \\
&= \mathbb{E}[\hat{\xi}^2] - 2\mathbb{E}[\hat{\xi}]\mathbb{E}[\hat{\xi}] + (\mathbb{E}[\hat{\xi}])^2 \\
&= \mathbb{E}[\hat{\xi}^2] - (\mathbb{E}[\hat{\xi}])^2
\end{aligned}$$

où $\mathbb{E}[\hat{\xi}^2]$ est l'espérance du carré de l'estimateur et $(\mathbb{E}[\hat{\xi}])^2$ est le carré de l'espérance de l'estimateur.

Formule du MSE

La formule du Mean Squared Error (MSE) d'un estimateur $\hat{\xi}$ est :

$$\mathrm{MSE}(\hat{\xi}) = \mathbb{E}\left[(\hat{\xi} - \xi)^2\right] = \mathrm{Var}(\hat{\xi}) + (\mathrm{Bias}(\hat{\xi}))^2 \tag{3}$$

La démonstration est la suivante :

$$\begin{aligned}
\mathrm{MSE}(\hat{\xi}) &= \mathbb{E}\left[(\hat{\xi} - \xi)^2\right] \\
&= \mathbb{E}\left[(\hat{\xi} - \mathbb{E}[\hat{\xi}] + \mathbb{E}[\hat{\xi}] - \xi)^2\right] \\
&= \mathbb{E}\left[(\hat{\xi} - \mathbb{E}[\hat{\xi}])^2 + 2(\hat{\xi} - \mathbb{E}[\hat{\xi}])(\mathbb{E}[\hat{\xi}] - \xi) + (\mathbb{E}[\hat{\xi}] - \xi)^2\right] \\
&= \mathbb{E}\left[(\hat{\xi} - \mathbb{E}[\hat{\xi}])^2\right] + 2\mathbb{E}\left[(\hat{\xi} - \mathbb{E}[\hat{\xi}])(\mathbb{E}[\hat{\xi}] - \xi)\right] + (\mathbb{E}[\hat{\xi}] - \xi)^2 \\
&= \mathrm{Var}(\hat{\xi}) + (\mathrm{Bias}(\hat{\xi}))^2
\end{aligned}$$

où $\mathrm{Bias}(\hat{\xi}) = \mathbb{E}[\hat{\xi}] - \xi$.

Bibliographie

[1] Haldane, J. B. S., & Jayakar, S. D. (1963). The distribution of extremal and nearly extremal values in samples from a normal distribution. Biometrika, 50(1/2), 89-94.

[2] Katz, R. W., Parlange, M. B., & Naveau, P. (2002). Statistics of extremes in hydrology. Advances in water resources, 25(8-12), 1287-1304.

[3] Cheng, S., et Peng, L. (2001). Cheng, S., et Peng, L. (2001). Intervalles de confiance pour l'indice de queue. Bernoulli, 751-760.

[4] Ferreira*, A., de Haan*, L., & Peng, L. (2003). Sur l'optimisation de l'estimation des quantiles élevés d'une distribution de probabilité. Statistiques, 37(5), 401-434

[5] Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une serie aleatoire. Annals of mathematics, 44(3), 423-453.

[6] de Haan, L., & Ferreira, A. (2006). Extreme Value Theory : An Introduction. Springer Series in Operations Research and Financial Engineering.

[7] Embrechts, P., Klüppelberg, C., & Mikosch, T. (1997). Modelling Extremal Events for Insurance and Finance. Springer.

[8] Resnick, S. I. (1987). Extreme Values, Regular Variation, and Point Processes. Springer-Verlag.

[9] Alvarado, E., D. Sand Berg, and S. Picford (1998). Modelling large forest fires as extreme events. Northwest Sci. 72, 66–75.

[10] Peng, L. (1998). Asymptotically unbiased estimators for the extreme-value index. Statistics and Probability Letters 38 (2), 107–115.

[11] Hill, B. (1975). A simple general approach to inference about the tail of a distribution. Annals of Statistics 3 (5), 1163–1174.

[12] McNeil, A. J. (1998). Calculating quantile risk measures for financial return series using extreme value theory. ETH Zurich.

[13] Leadbetter, M. R. (1983). Extremes and local dependence of stationnary sequences. Z. Wahr. verw. Gebiete (65), 291–306.

[14] Deheuvels, P., Haeusler, E., & Mason, D. M. (1988, septembre). Convergence presque certaine de l'estimateur de Hill. Dans Mathematical Proceedings of the Cambridge Philosophical Society (vol. 104, n° 2, pp. 371-381). Presses de l'Université de Cambridge.

[15] Pickands III, J. (1975). Statistical inference using extreme order statistics. the Annals of Statistics, 119-131.

[16] Lo, G. (1986). Sur quelques estimateurs de lâindex dâune loi de Pareto : Estimation de Deheuvels Csorgo, Mason, de De Haan-Resnick et lois limites pour des sommes de valeurs extrêmes pour une variable dans le domaine de Gumbel. Thèse de Doctorat. Universit´ Paris VI. France. Ph. D. thesis.

[17] Reiss, R. D., Thomas, M. et Reiss, R. D. (1997). Analyse statistique des valeurs extrêmes (Vol. 2). Bâle : Birkhäuser.

[18] Reiss, R.-D., & Thomas, M. (2007). Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology, and Other Fields (3rd ed.). Birkhäuser.

[19] Coles, S. (2001). An Introduction to Statistical Modeling of Extreme Values. Springer.

[20] R. Fisher and L. Tippett (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. Proceedings of the Cambridge Philosophical Society, 24, 180-190.

[21] Beirlant, J., Goegebeur, Y., Segers, J., & Teugels, J. (2004). Statistics of Extremes : Theory and Applications. John Wiley & Sons.

[22] Von Mises, R. (1936). La distribution de la plus grande des valeurs. Revue de l'Institut International de Statistique, 4(3), 5-18.

[23] De Haan, L. D., & Peng, L. (1998). Comparison of tail index estimators. Statistica Neerlandica, 52(1), 60-70.

[24] Jenkinson, A. F. (1955). The Frequency Distribution of the Annual Maximum (or Minimum) Values of Meteorological Elements. Quarterly Journal of the Royal Meteorological Society, 81(348), 158-171.

[25] Borchani, A. (2010). Statistiques des valeurs extrêmes dans le cas de lois discrètes. Centre de Recherche de l'ESSEC ISSN.

[26] Dekkers, A., J. H. J. Einmalh, and L. De Hann (1989). A moment estimator for the index of an extreme-value distribution. Annals of Statistics 17 (4), 1833–1855.

[27] Klüppelberg, C. (2004). Risk management with extreme value theory. Extreme Values in Finance, Telecommunication and the Environment, 101-168.

[28] Lagakos, S. W., Barraj, L. M., & Gruttola, V. D. (1988). Nonparametric analysis of truncated survival data, with application to AIDS. Biometrika, 75(3), 515-523.

[29] Vitolo, C. (2017). hddtools : Hydrological Data Discovery Tools. R package version 0.7.5. Disponibilité

[30] Smith, R. L. (1987). Estimating tails of probability distributions. Annals of Statistics 15, 1174–1207.

[31] Danielsson, J., D. W. J., and C. G. de Vries (1996). The method of moments ratio estimator for the tail shape parameter. Communication in Statistics, Theory and Methods 4 (25), 711–720.

[32] Bhatt, S., Weiss, D. J., Cameron, E., Bisanzio, D., Mappin, B., Dalrymple, U., ... & Gething, P. W. (2015). The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015. Nature, 526(7572), 207-211.

[33] Reiss, R. and M. Thomas (2001). Statistical analysis of extreme values. Birkhanser Verlag.

[34] Davison, A. C., & Smith, R. L. (1990). Models for exceedances over high thresholds. Journal of the Royal Statistical Society Series B : Statistical Methodology, 52(3), 393-425.

[35] Klein, J. P., & Moeschberger, M. L. (2003). Survival analysis : techniques for censored and truncated data (Vol. 2, pp. 3-5). New York : Springer

[36] Klein JP et Moeschberger, ML (1997). Survival analysis : techniques for censored and truncated data. Springer-Verlag, New York.

[37] Bingham, N. H., C. M. Goldie, and J. L. Teugels (1987). Regular Variation, Volume 27. Encyclopedia of Mathematics and its Applications, Cambridge University Press.

[38] Tsourti, Z., & Panaretos, J. (2003). Extreme Value Index Estimators and Smoothing Alternatives : A Critical Review. STOCHASTIC MUSINGS : PERSPECTIVES FROM THE PIONEERS OF THE LATE 20TH CENTURY, J. Panaretos, ed., Laurence Erlbaum, Publisher, USA, 141-160.

[39] El-Adlouni, S., B. Bobée, and T. B. Ouarda (2007). Caracterisation des distributions à queues lourdes pour l'analyse des crues. Technical Report no r-929, INRS-ETE, Université du Québec.

[40] Hosking, J. R. M. and J. R. Wallis (1987, August). Parameter and quantile estimation for the generalized pareto distribution. Technometrics 29 (3), 339–1349.

[41] Beirlant, J., & Goegebeur, Y. (2004). Local polynomial maximum likelihood estimation for Pareto-type distributions. Journal of Multivariate Analysis, 89(1), 97-118.

[42] Dekkers, A., J. H. J. Einmalh, and L. De Hann (1989). A moment estimator for the index of an extreme-value distribution. Annals of Statistics 17 (4), 1833–1855.

[43] Deheuvels, P., Häeusler, E, and Mason, D. M. (1988). Almost sure convergence of the hill estimator. Math. Proc. Cambridge Philos. Soc, 104(02) ;371 381.

[44] Beirlant, J., & Goegebeur, Y. (2003). Regression with response distributions of Pareto-type. Computational statistics & data analysis, 42(4), 595-619.

[45] Mason, D. M. (1982). Laws of large numbers for sums of extreme values. The Annals of Probability, 754-764.

[46] Davis, R., & Resnick, S. (1984). Tail estimates motivated by extreme value theory. The Annals of Statistics, 1467-1487.

[47] Haeusler, E., & Teugels, J. L. (1985). On asymptotic normality of Hill's estimator for the exponent of regular variation. The Annals of Statistics, 743-756.

[48] Goldie, C. M., & Smith, R. L. (1987). Slow variation with remainder : Theory and applications. The Quarterly Journal of Mathematics, 38(1), 45-71.

[49] Bertail, P. (2002). Evaluation des risques d'exposition à un contaminant alimentaire : quelques outils statistiques. INSEE.

[50] Sousa, B. C. (2002). A Contribution to the Estimation of the Tail index of Heavy-tailed Distributions. University of Michigan.

[51] Vandewalle, B., Beirlant, J., Christmann, A., & Hubert, M. (2007). A robust estimator for the tail index of Pareto-type distributions. Computational Statistics & Data Analysis, 51(12), 6252-6268.

[52] Dekkers, A. L., & De Haan, L. (1989). On the estimation of the extreme-value index and large quantile estimation. The annals of statistics, 1795-1832.

[53] Drees, H. (1995). Refined Pickands estimators of the extreme value index. The Annals of Statistics, 2059-2080.