MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE MOULOUD MAMMERI, TIZI-OUZOU

FACULTE DES SCIENCES

DEPARTEMENT : MATHEMATIQUES

Mémoire de Master

SPECIALITE: MATHEMATIQUES

Option : PROBABILITES ET STATISTIQUE

Présentée par:

TIGHEDINE DIHIA

Thème

Application de la théorie des valeurs extrêmes en hydrologie

Devant le jury d'examen composé de:

Mme MERABET	Dalila	MCB	U.M.M.T.O	Présidente.
Mme BOUALAM	Karima	MCB	U.M.M.T.O	Rapporteur.
Mme BOUZIANE	Houria	MAA	U.M.M.T.O	Examinatrice.

Soutenue le 03/07/2025

Remerciements

Nous tenons tout d'abord à remercier Dieu le tout puissant et miséricordieux, qui nous a donné la force et la patience d'accomplir ce modeste travail.

En second lieu, je présente mes sincères remerciements à mon encadreur Mme Boualam Karima pour son soutien et son attention exceptionnels durant toute la période du travail et pour son suivi attentif et pertinent qui a mené à l'acheminement de ce travail.

Nos vifs remerciements vont également aux membres du jury, Mme Merabet Dalila et Mme Bouziane Houria pour avoir accepté d'évaluer mon travail.

Enfin, nous tenons également à remercier toutes les personnes qui ont participé de près ou de loin à la réalisation de ce travail.

Table des matières

Introduction générale

L'analyse des phénomènes rares ou extrêmes occupe une place essentielle dans de nombreux domaines scientifiques et techniques, tels que la finance, les télécommunications, la climatologie, et tout particulièrement l'hydrologie (Davison Smith [7] ; Katz, Parlange Naveau [22]) . Des événements comme les crues, les sécheresses, les tempêtes ou les inondations, bien que peu fréquents, peuvent engendrer des conséquences considérables. Situés dans les queues des distributions statistiques, ces phénomènes exigent des outils spécifiques pour leur modélisation et leur prévision.

C'est dans ce cadre que s'inscrit la théorie des valeurs extrêmes (TVE), qui constitue un fondement mathématique rigoureux pour l'étude du comportement asymptotique des maxima (ou minima) observés dans de grands échantillons. Cette théorie fournit des lois limites telles que celles de Fréchet, Gumbel ou Weibull, vers lesquelles convergent les distributions des extrêmes, et propose des outils puissants pour l'estimation de quantiles rares, de périodes de retour et de niveaux de risque élevé.

L'objectif principal de ce mémoire est d'explorer les fondements théoriques de la TVE, d'en étudier les principales distributions asymptotiques et d'en illustrer l'application au domaine hydrologique, notamment pour l'estimation des crues extrêmes. Une attention particulière sera accordée à la loi généralisée de Pareto (GPD) dans le cadre de l'approche Peaks Over Threshold (POT), qui permet une meilleure exploitation des données excédant un certain seuil critique.

Dans un contexte empirique, les données extrêmes peuvent être incomplètes ou censurées, en raison de limitations techniques des instruments de mesure, de périodes d'observation trop courtes, ou de l'incomplétude des séries de données relatives aux événements rares. Le traitement de telles données nécessite le recours à des méthodes spécifiques, issues notamment de l'analyse de survie ou de modèles semi-paramétriques, qui seront également intégrées à notre étude.

Ce mémoire est structuré de la manière suivante :

- Le premier chapitre introduit les fondements théoriques de la théorie des valeurs extrêmes (Fisher Tippett [14] et Gnedenko [15]). Il traite du comportement asymptotique des maxima, de la caractérisation des domaines d'attraction (Fréchet, Gumbel, Weibull) (voir Embrechts et al.[13] et De Haan Ferreira [17]), ainsi que de la modélisation des dépassements de seuil via la loi généralisée de Pareto, en insistant sur le choix du seuil, paramètre crucial pour la qualité des ajustements.
- Le deuxième chapitre est consacré à l'estimation des paramètres des distributions d'extrêmes. Nous y présentons différentes méthodes, allant des approches graphiques aux estimateurs non paramétriques robustes tels que ceux de Hill [18], des moments [8] et de Pickands [26], ainsi que l'estimation par maximum de vraisemblance pour la GPD. L'accent est mis sur l'estimation des quantiles extrêmes, essentielle pour l'analyse du risque.

— Le troisième chapitre propose une application pratique de la TVE à des données hydrologiques réelles. Il détaille le processus d'ajustement des modèles à des séries de débits, incluant la sélection d'observations indépendantes et l'estimation des périodes de retour pour des événements de crue.

À travers cette démarche, ce travail vise à approfondir la compréhension et la quantification du risque associé aux événements hydrologiques extrêmes, dans une perspective de gestion durable des ressources en eau et de prévention des catastrophes naturelles.

Chapitre 1

Théorie des valeurs extrêmes

1.1 Introduction

L'objectif de ce chapitre est de présenter les principaux outils statistiques utilisés pour la modélisation des valeurs extrêmes. La théorie des valeurs extrêmes repose sur des concepts mathématiques rigoureux qui permettent d'étudier le comportement des variables aléatoires dans les cas où les observations sont situées dans les extrêmes de leur distribution. Parmi les principaux outils de cette théorie, on trouve les distributions limites des extrêmes et leurs domaines d'attraction, qui sont essentielles pour comprendre la convergence des lois des extrêmes.
Deux modèles de distribution jouent un rôle central dans cette théorie : la loi Généralisée des Valeurs Extrêmes et la loi des Excédents Généralisée. La loi GEV décrit le comportement asymptotique des maxima d'échantillons, tandis que la loi GPD permet de modéliser les excédents au-dessus d'un seuil donné. Ces deux lois sont au cœur de l'analyse statistique des événements extrêmes, car elles offrent des cadres probabilistes permettant d'estimer des quantités clés telles que les quantiles extrêmes, l'indice des extrêmes, et la période de retour des événements rares.
En outre, ce chapitre met en lumière la notion de domaine d'attraction, qui permet de classifier les distributions de valeurs extrêmes en trois grandes familles : Gumbel, Fréchet et Weibull. Cette classification repose sur le paramètre de forme, appelé indice des valeurs extrêmes, qui décrit le comportement des queues des distributions. L'objectif final de cette approche est de fournir une estimation précise des risques associés à des événements extrêmes en s'appuyant sur des méthodes statistiques robustes et bien fondées.

1.2 Comportement asymptotique des extrêmes

Soit un ensemble de n variables aléatoires indépendantes et identiquement distribuées (i.i.d.), notées $X_1, X_2, ..., X_n$, ayant pour fonction de répartition F. L'échantillon ordonné des n valeurs, noté $X_{1:n}, X_{2:n}, ..., X_{n:n}$, est défini de manière à ce que :

$$X_{1:n} \leq X_{2:n} \leq ... \leq X_{n:n},$$

où $X_{k:n}$ est la k^{ieme} statistique d'ordre.
Parmi les statistiques d'ordre, celles qui sont particulièrement pertinentes pour l'étude des événements extrêmes sont le minimum, $X_{1:n} = \min(X_1, ..., X_n)$, et le maximum, $X_{n:n} = \max(X_1, ..., X_n)$. Il est facile de passer de l'une à l'autre à l'aide de la relation suivante :

$$X_{1:n} = -\max\{-X_1, -X_2, ..., -X_n\},$$

Dans la suite de ce chapitre, nous nous concentrerons principalement sur l'étude du maximum, c'est-à-dire sur la statistique d'ordre $X_{n:n}$.
La fonction de répartition de la statistique d'ordre pour le maximum $X_{n:n}$ est donnée par :

$$F_{X_{n:n}}(x) = P(X_1 \leq x, X_2 \leq x, ..., X_n \leq x) = [F(x)]^n.$$

Bien que cette formule soit correcte, elle présente un intérêt limité, car la loi d'une variable aléatoire X est rarement précisément connue. De plus, même si la loi de X est connue, le calcul de la loi du maximum $X_{n:n}$ reste souvent difficile.

Nous introduisons alors la fonction de survie, notée $\bar{F}$, qui est définie par :

$$\bar{F}(x) = P(X > x) = 1 - F(x).$$

On définit également x_F, le point terminal à droite de la fonction de répartition F, qui peut être soit fini, soit infini.

En analysant le comportement asymptotique du maximum à mesure que $n \to +\infty$, on obtient la distribution limite suivante pour $F_{X_{n:n}}(x)$:

$$\lim_{n\to\infty} F_{X_{n:n}}(x) = \lim_{n\to\infty} [F(x)]^n = \begin{cases} 0 & \text{si } x < x_F, \\ 1 & \text{si } x \geq x_F. \end{cases}$$

Ainsi, la loi du maximum, à mesure que n tend vers l'infini, devient une loi dégénérée, ce qui offre peu d'informations sur le comportement de $X_{n:n}$. Pour remédier à cela, nous appliquons une transformation de normalisation afin de produire une loi asymptotique non dégénérée.
Le théorème fondamental en théorie des valeurs extrêmes suivant, énoncé par Fisher et Tippett (1928) puis généralisé par Gnedenko (1943), établit la loi asymptotique normalisée du maximum d'un échantillon.

Théorème 1 (Théorème de Fisher-Tippett [14] -Gnedenko [15]) *Soit $(X_n)_n$ une suite de variables aléatoires i.i.d. avec fonction de répartition F. S'il existe deux suites normalisantes $(a_n)_{n\geq 1}$, $a_n > 0$, et $(b_n)_{n\geq 1}$, telles que la loi du maximum normalisé converge vers une loi H, alors la fonction de répartition de cette loi est donnée par :*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{X_{n:n} - b_n}{a_n} \leq x\right) = H(x), \quad \forall x \in \mathbb{R}. \tag{1.1}$$

Cette loi H est l'une des trois suivantes :

— *Loi de Gumbel :*

$$\Lambda(x) = \exp\left(-\exp(-x)\right), \quad x \in \mathbb{R},$$

— *Loi de Fréchet :*

$$\phi_\alpha(x) = \begin{cases} 0 & si\ x \leq 0 \\ \exp\left(-(x)^{-\alpha}\right) & si\ x > 0 \end{cases}, \alpha > 0.$$

— *Loi de Weibull :*

$$\Psi_\alpha(x) = \begin{cases} \exp\left(-(-x)^{\alpha}\right) & si\ x \leq 0 \\ 1 & si\ x > 0 \end{cases}, \alpha > 0.$$

Les trois types de distributions extrêmes, Gumbel, Fréchet et Weibull, sont souvent appelées les lois des valeurs extrêmes. Le paramètre α est appelé indice des valeurs extrêmes et caractérise le comportement de la queue de la distribution.
Une démonstration détaillée de ce théorème peut être trouvée dans les travaux de Resnick [28] et Embrechts et al. [13].
Le théorème suivant dû à Von Mises [31] et Jenkinson [20] établit l'unification du comportement du maximum en une seule fonction de répartition, via la loi généralisée des valeurs extrêmes, permet une étude simplifiée du comportement des extrêmes. Cette loi dépend du paramètre de forme γ, appelé indice des valeurs extrêmes.

Définition 1.1 (Représentation de Von Mises-Jenkinson) *Soit $\gamma \in \mathbb{R}$, on appelle **distribution des valeurs extrêmes généralisée standard** toute distribution H_γ définie par :*

$$H_\gamma(x) = \begin{cases} \exp\left(-(1+\gamma x)^{\frac{-1}{\gamma}}\right) & si\ \gamma \neq 0 \quad et \quad 1+\gamma x > 0, \\ \exp\left(-\exp(-x)\right) & si\ \gamma = 0\ , x \in \mathbb{R}. \end{cases}$$

Le paramètre γ caractérise le comportement de la queue de la distribution, et selon sa valeur, on distingue trois domaines d'attraction :

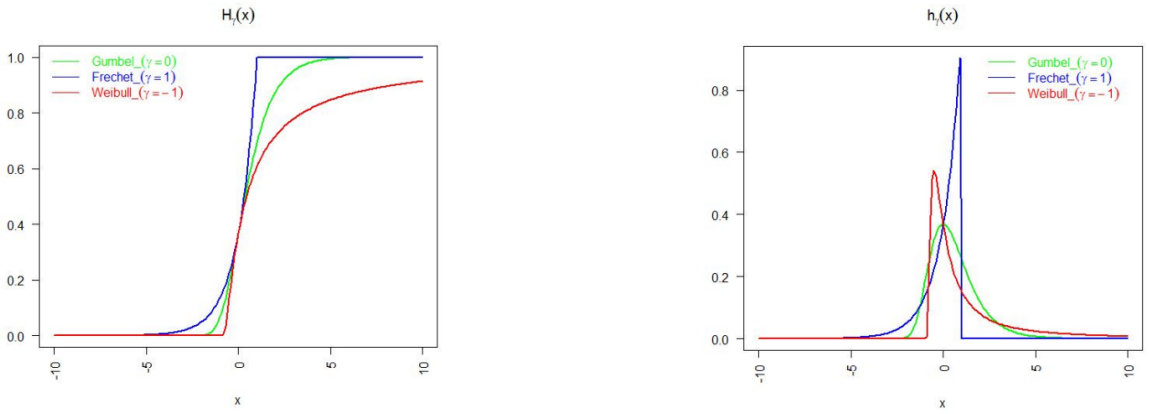
- Si $\gamma > 0$, la fonction de répartition F appartient au domaine d'attraction de Fréchet, correspondant aux lois à queue lourde (comme la loi de Pareto).
- Si $\gamma < 0$, F appartient au domaine d'attraction de Weibull, pour les lois à point terminal fini.
- Si $\gamma = 0$, F appartient au domaine d'attraction de Gumbel, pour les lois à queue légère.

La forme générale de la GEV est :

$$H_{\gamma,\mu,\sigma}(x) = \begin{cases} \exp\left(-\left(1+\gamma\frac{x-\mu}{\sigma}\right)^{\frac{-1}{\gamma}}\right) & \text{si } \gamma \neq 0, \\ \exp\left(-\exp\left(-\frac{x-\mu}{\sigma}\right)\right) & \text{si } \gamma = 0. \end{cases}$$

La densité associée à une variable aléatoire suivant la loi GEV est donnée par :

$$h_{\gamma,\mu,\sigma}(x) = \frac{1}{\sigma}\left(1+\gamma\frac{x-\mu}{\sigma}\right)^{\frac{-1}{\gamma}-1} H_{\gamma,\mu,\sigma}(x), \quad \gamma \neq 0.$$



Fonction de répartition Fonction de densité

FIGURE 1.1 – Représentation graphique de la GEV avec $\mu = 0$, $\sigma = 1$

1.3 Caractérisation des domaines d'attraction

Dans cette section, nous présentons les caractéristiques des trois domaines d'attraction : Fréchet, Weibull et Gumbel. Nous commençons par énoncer quelques définitions des outils nécessaires à leur caractérisation.

1.3.1 Fonction à variation régulière

Définition 1.2 (Fonction à variation régulière, Bingham et al. [2]) *Une fonction f mesurable et positive sur $[a, \infty[$, avec $a > 0$, est dite à variation régulière à l'infini si et seulement si, pour tout $x > 0$, il existe un réel ρ tel que :*

$$\lim_{t\to+\infty} \frac{f(tx)}{f(t)} = x^{\rho}.$$

On note $f \in \mathrm{RV}_{\rho}$, où ρ est appelé l'indice de la fonction à variation régulière.

Définition 1.3 (Fonction à variation lente) *Une fonction L est dite à variation lente, si $\rho = 0$:*

$$\lim_{t\to+\infty} \frac{L(tx)}{L(t)} = 1, \quad \forall x > 0.$$

$L \in RV_0$

Nous énonçons maintenant quelques propriétés fondamentales des fonctions à variation régulière :

Proposition 1 (Représentation de Karamata, Resnick [28]) *Une fonction L est à variation lente à l'infini si et seulement si elle peut être représentée sous la forme suivante :*

$$L(x) = c(x) \exp\left(\int_a^x \frac{\sigma(t)}{t}\, dt\right), \quad \forall x \geq a > 0,$$

où c et σ sont deux fonctions mesurables telles que :

$$\lim_{x\to+\infty} c(x) = c_0 \in (0, \infty) \quad et \quad \lim_{t\to+\infty} \sigma(t) = 0.$$

Le théorème de représentation de Karamata est un outil puissant qui donne une forme explicite pour les fonctions à variation lente et permet de mieux comprendre leur comportement asymptotique.

Proposition 2 (Bingham et al. [2]) *Les propriétés importantes des fonctions à variation lente :*

1. *Si L est une fonction à variation lente à l'infini, alors :*

$$\lim_{x\to+\infty} \frac{\log L(x)}{\log x} = 0.$$

2. *Si L est une fonction à variation lente à l'infini et $\rho > 0$, alors :*

$$\lim_{x\to+\infty} x^{\rho} L(x) = +\infty \quad et \quad \lim_{x\to+\infty} x^{-\rho} L(x) = 0.$$

3. *Si L est une fonction à variation lente à l'infini, alors pour tout $\rho \in \mathbb{R}$, la fonction*

$$L^\rho : x \longmapsto [L(x)]^\rho$$

est également une fonction à variation lente à l'infini.

4. *Si L_1 et L_2 sont des fonctions à variation lente à l'infini, alors*

$$L_1 + L_2 : x \longmapsto L_1(x) + L_2(x) \quad et \quad L_1.L_2 : x \longmapsto L_1(x).L_2(x)$$

sont des fonctions aux variations lentes à l'infini. De plus, si de plus $\lim_{x\to+\infty} L_2(x) = +\infty$, alors la fonction

$$L_1 \circ L_2 : x \longmapsto L_1\left[L_2(x)\right]$$

est aussi à variation lente à l'infini.

Le théorème suivant établit le lien entre une fonction à variation régulière d'indice $\alpha \in \mathbb{R}$ et les fonctions à variation lente.

Théorème 2 (Bingham et al. [2]) *Soit $\alpha \in \mathbb{R}$ et f une fonction mesurable sur $[a, \infty[$, avec $a > 0$, et $f \in \mathrm{RV}_\alpha$. On a :*

$$f(x) = x^\alpha L(x),$$

où L est une fonction à variation lente.

Proposition 3 (Bingham et al., [2]) *Soient α, α_1 et α_2 des réels :*

1. *Si $f \in \mathrm{RV}_\alpha$, alors*

$$\lim_{x\to+\infty} f(x) = \begin{cases} 0 & si\ \alpha < 0, \\ +\infty & si\ \alpha > 0. \end{cases}$$

2. *Si $f \in \mathrm{RV}_\alpha$, alors*

$$\lim_{x\to+\infty} \frac{\log f(x)}{\log x} = \alpha.$$

3. *Si $f \in \mathrm{RV}_\alpha$ et $\rho \in \mathbb{R}$, alors*

$$f^\rho : x \longmapsto [f(x)]^\rho \in \mathrm{RV}_{\alpha\rho}\,.$$

4. *Si $f_1 \in \mathrm{RV}_{\alpha_1}$ et $f_2 \in \mathrm{RV}_{\alpha_2}$, alors*

$$f_1 + f_2 : x \longmapsto f_1(x) + f_2(x) \in \mathrm{RV}_{\max(\alpha_1,\alpha_2)},$$

et si de plus $\lim_{x\to+\infty} f_2(x) = +\infty$, alors

$$f_1 \circ f_2 : x \longmapsto f_1\left[f_2(x)\right] \in \mathrm{RV}_{\alpha_1\alpha_2}\,.$$

1.3.2 Inverse généralisée

Définition 1.4 (Inverse généralisée) *Soit $F^\leftarrow$ l'inverse généralisé ou fonction de quantile de la fonction distribution F définie par :*

$$F^\leftarrow(q) = \inf\{x \in \mathbb{R}, F(x) \geq q\}, \quad 0 < q < 1,$$

Proposition 4 (Propriétés de la fonction inverse généralisée Resnick [28]) *Soient F une fonction croissante et $F^\leftarrow$ son inverse généralisée, alors $F^\leftarrow$ est une fonction continue à gauche. Si, en plus, F est continue à droite, alors on a :*

1. $F(x) \geq q \iff F^{\leftarrow}(q) \leq x$
2. $F(x) < q \iff F^{\leftarrow}(q) > x$
3. $\forall\, q \in]a, b[, \quad F[F^{\leftarrow}(q)] \geq q$ *avec égalité si F est continue.*
4. $\forall\, x \in \mathbb{R}, \quad F^{\leftarrow}[F(x)] \leq x$ *avec égalité si F est strictement croissante.*

Définition 1.5 (Domaine d'attraction) *On dit qu'une variable aléatoire X (ou de sa fonction de répartition F) appartient au domaine d'attraction de la distribution des extrêmes généralisée (ou sa fonction de répartition H) s'il existe des constantes $a_n > 0$ et $b_n \in \mathbb{R}$ telles que (1.1) est vérifiée. On écrit $X \in D(H)$ (ou $F \in D(H)$)*

1.3.3 Domaine d'attraction de Fréchet

Condition nécessaire et suffisante

Théorème 3 (De Haan et Ferreira, [17]) *Une fonction de répartition F appartient au domaine d'attraction de Fréchet $\gamma > 0$ si et seulement si :*

$$1) \quad x_F = \infty,$$
$$2) \quad \lim_{t\to\infty} \frac{\bar{F}(tx)}{\bar{F}(t)} = x^{-1/\gamma}, \quad \gamma > 0.$$

Autrement dit, une fonction de répartition F appartenant au domaine d'attraction de Fréchet s'écrit sous la forme :

$$F(x) = 1 - x^{-1/\gamma} L(x), \quad L \in \mathrm{RV}_0\,.$$

Les constantes de normalisation $(a_n)_n$ et $(b_n)_n$ sont données dans ce cas par :

$$a_n = F^{\leftarrow}\left(1 - \frac{1}{n}\right), \quad b_n = 0.$$

Condition suffisante

Proposition 5 (Condition de Von Mises) *Soit X une variable aléatoire de fonction de répartition F absolument continue et de fonction de densité f vérifiant :*

$$1) \quad x_F = \infty,$$
$$2) \quad \lim_{x\to\infty} \frac{x f(x)}{1 - F(x)} = \gamma, \quad (\gamma > 0).$$

Alors $F \in D(Fréchet)$.

1.3.4 Domaine d'attraction de Weibull

Condition nécessaire et suffisante

Théorème 4 (De Haan et Ferreira, [17]) *Une fonction de répartition F appartient au domaine d'attraction de Weibull si et seulement si :*

$$1) \quad x_F < +\infty,$$
$$2) \quad \forall x \in \mathbb{R}^+, \exists\, \gamma < 0 \quad \textit{tel que} \quad \lim_{t\to 0^+} \frac{1 - F(x_F - tx)}{1 - (x_F - t)} = x^{-1/\gamma}.$$

Les constantes de normalisation $(a_n)_n$ et $(b_n)_n$ sont données par :

$$a_n = x_F - F^{\leftarrow}\left(1 - \frac{1}{n}\right), \quad b_n = x_F.$$

Condition suffisante

Proposition 6 (Condition de Von Mises) *Soit X une variable aléatoire de fonction de répartition F absolument continue et de fonction de densité f vérifiant :*

$$\begin{aligned} &1) \quad x_F < \infty, \\ &2) \quad \lim_{x \to x_F} \frac{(x_F - x) f(x)}{1 - F(x)} = -\gamma, \quad (\gamma < 0). \end{aligned}$$

Alors $F \in D(Weibull)$.

1.3.5 Domaine d'attraction de Gumbel

Condition nécessaire et suffisante

Théorème 5 (De Haan et Ferreira, [17]) *Une fonction de répartition F appartient au domaine d'attraction de Gumbel si et seulement si :*

$$\begin{aligned} &1) \quad \mathbb{E}(X \mid X > c) < \infty \quad pour \quad c < F^{\leftarrow}(1), \\ &2) \quad \lim_{t \to x_F} \frac{1 - F(t + x g(t))}{1 - F(t)} = e^{-x}, \quad avec \quad g(t) = \mathbb{E}(X - t \mid X > t) \end{aligned}$$

où g représente l'espérance conditionnelle de $X - t$ sachant que $X > t$. Cette condition est nécessaire et suffisante pour que la fonction de répartition F appartienne au domaine d'attraction de Gumbel.

Les constantes de normalisation $(a_n)_n$ et $(b_n)_n$ qui permettent de définir la loi limite de F dans ce domaine sont données par :

$$a_n = F^{\leftarrow}\left(1 - \frac{1}{n}\right), \quad \text{et} \quad b_n = \mathbb{E}(X - a_n \mid X > a_n).$$

Ici, $F^{\leftarrow}(q)$ représente l'inverse généralisé de la fonction de répartition F, et a_n est une suite de normalisation qui dépend de F. La constante b_n est l'espérance conditionnelle de $X - a_n$ sachant que $X > a_n$, ce qui correspond à la moyenne ajustée de la variable aléatoire X après avoir normalisé par a_n.

Condition suffisante

Proposition 7 (Condition de Von Mises) *Soit X une variable aléatoire de fonction de répartition F absolument continue et de fonction de densité f. Si $\dfrac{f(x)}{1 - F(x)} > 0$ et que la fonction $\dfrac{1 - F(x)}{f(x)}$ est dérivable au voisinage de x_F, avec la propriété suivante :*

$$\lim_{x \to x_F} \frac{d}{dx}\left(\frac{1 - F(x)}{f(x)}\right) = 0,$$

alors $F \in D(Gumbel)$.
Dans ce cas, les constantes de normalisation $(a_n)_n$ et $(b_n)_n$ peuvent être choisies comme suit :

$$a_n = F^{\leftarrow}\left(1 - \frac{1}{n}\right), \quad et \quad b_n = F^{\leftarrow}\left(1 - \frac{1}{ne}\right) - F^{\leftarrow}\left(1 - \frac{1}{n}\right).$$

Voici un classement de quelques lois par domaine d'attraction dans le tableau 1.1.

Fréchet ($\gamma > 0$)	Gumbel ($\gamma = 0$)	Weibull ($\gamma < 0$)
Pareto	Normale	Uniforme
Student	Exponentielle	Beta
Burr	Log-normale	
Chi-deux	Gamma	
Fréchet	Weibull	
Log-gamma	Gumbel	
Log-logistique	Logistique	
Cauchy		

TABLE 1.1 – Quelques lois de probabilité et leurs domaines d'attraction.

1.4 Loi généralisée de Pareto

La loi des valeurs extrêmes généralisée est utilisée pour modéliser le comportement des maxima d'un échantillon sur une période donnée. Elle est couramment appliquée dans des contextes où l'on cherche à étudier les valeurs extrêmes au sein de l'ensemble de la distribution. En revanche, la méthode des dépassements de seuil, en anglais *Peaks-Over-Threshold* (POT), proposée par Pickands en 1975, repose sur la distribution de Pareto généralisée. Cette méthode permet d'analyser directement les valeurs qui dépassent un seuil prédéfini, ce qui la rend particulièrement utile dans des applications centrées sur les événements extrêmes, tels que les inondations ou les pertes financières.

Distribution des excès

Définition 1.6 *On appelle excès de la variable aléatoire X au-delà d'un seuil u (avec $u < x_F$) la variable aléatoire Y, qui prend ses valeurs dans l'intervalle $]0, x_F - u[$, et est définie par :*

$$Y = X - u \mid X > u.$$

Définition 1.7 (Distribution des excès) *On appelle distribution des excès de la variable aléatoire X par rapport à un seuil $u < x_F$ la loi de probabilité de la variable aléatoire Y, représentant l'excès de X au-delà du seuil u, donnée par sa fonction de répartition F_u :*

$$\forall\, y \in \mathbb{R}, \quad F_u(y) = P(X - u \leq y \mid X > u) = \begin{cases} 0 & si\ y \leq 0, \\ 1 - \dfrac{1 - F(u+y)}{1 - F(u)} & si\ 0 < y < x_F - u, \\ 1 & si\ y \geq x_F - u. \end{cases}$$

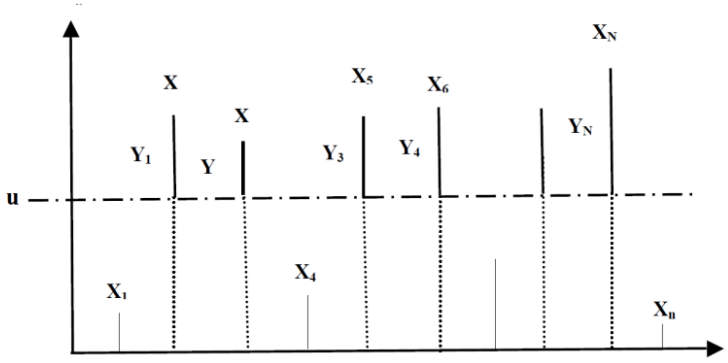


FIGURE 1.2 – Illustration de la méthode des excès.

Exemple(La loi exponentielle)
Soit X une variable aléatoire distribuée selon une loi exponentielle de fonction de répartition F définie par $F(x) = (1 - e^{-x})\mathbb{I}_{\{x>0\}}(x)$. Par calcul direct, on trouve

$$\begin{aligned} P(X > u + y | X > u) &= \frac{P(X > u + y \text{ et } X > u)}{P(X > u)} \\ &= \frac{P(X > u + y)}{P(X > u)} \\ &= \frac{1 - F(u + y)}{1 - F(u)} \\ &= \frac{1 - (1 - e^{-(u+y)})}{1 - (1 - e^{-u})} \\ &= \frac{e^{-(u+y)}}{e^{-u}} \\ &= e^{-y}, \quad y > 0 \end{aligned}$$

Définition 1.8 *Une distribution $\mathcal{G}_{\gamma,\sigma}$ est dite de Pareto généralisée avec paramètres γ et σ si elle est définie comme suit :*

$$\mathcal{G}_{\gamma,\sigma}(y) = \begin{cases} 1 - \left(1 + \gamma \frac{y}{\sigma}\right)^{\frac{-1}{\gamma}} & si\ \gamma \neq 0, \\ 1 - \exp\left(-\frac{y}{\sigma}\right) & si\ \gamma = 0. \end{cases}$$

Elle est définie pour $y \in \mathbb{R}^+$ sous la condition $1 + \gamma \frac{y}{\sigma} > 0$.

Cette loi dépend de deux paramètres :

(a) $\sigma > 0$: paramètre d'échelle.

(b) $\gamma \in \mathbb{R}$: paramètre de forme.

Remarques

— Si $\gamma = 0$, la loi devient une loi exponentielle de paramètre $\frac{1}{\sigma}$:

$$\mathcal{G}_{0,\sigma}(y) = 1 - \exp\left(-\frac{y}{\sigma}\right), \quad y \geq 0.$$

— Si $\gamma = -1$, la loi devient une loi uniforme sur l'intervalle $[0, \sigma]$.

— La forme générale de la GPD est :

$$\mathcal{G}_{\gamma,\mu,\sigma}(y) = \begin{cases} 1 - \left(1 + \gamma \frac{y-\mu}{\sigma}\right)^{\frac{-1}{\gamma}} & \text{si } \gamma \neq 0, \\ 1 - \exp\left(\frac{\mu - y}{\sigma}\right) & \text{si } \gamma = 0, \end{cases}$$

définie pour $y > \mu$, sous la condition $1 + \gamma \frac{y-\mu}{\sigma} > 0$, où $\mu \in \mathbb{R}$ est le paramètre de position et $\sigma > 0$ est le paramètre d'échelle.

— Densité de la GPD :

$$\forall x > 0, \quad g_\gamma(x) = \begin{cases} (1+\gamma x)^{-\frac{1}{\gamma}-1} \mathbb{I}_{1+\gamma x>0}(x) & \text{si } \gamma \neq 0, \\ e^{-x} & \text{si } \gamma = 0. \end{cases}$$

— Moments de la GPD : pour $k \in \mathbb{N}^*$, on a

$$\mathbb{E}(Y^k) = \frac{\sigma^k \Gamma\left(\frac{1}{\gamma} - k\right)}{\gamma^{k+1} \Gamma\left(\frac{1}{\gamma} + 1\right)} k!,$$

avec la condition $\gamma < \dfrac{1}{k}$. En particulier, on obtient :

$$\mathbb{E}(Y) = \frac{\sigma}{1-\gamma}, \quad \text{pour} \quad \gamma < 1,$$

et

$$\mathbb{V}(Y) = \frac{\sigma^2}{(1-\gamma)^2(1-2\gamma)}, \quad \text{pour} \quad \gamma < \frac{1}{2}.$$

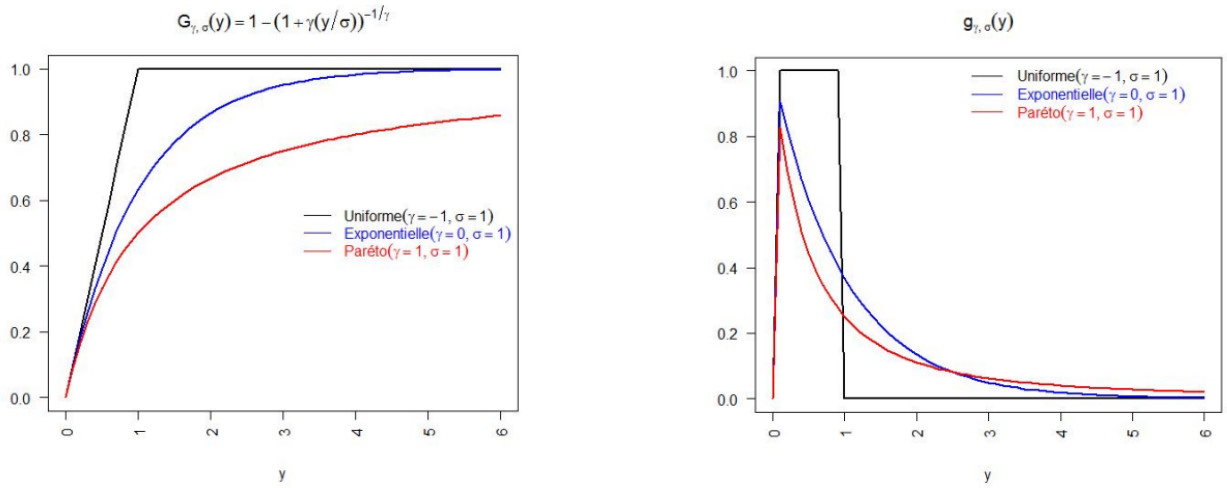
Le théorème suivant établit la relation entre le comportement asymptotique de la distribution des excès et la loi de Pareto généralisée.

Théorème 6 (Théorème de Balkema et Haan, Pickands) *Soit F la fonction de répartition d'une variable aléatoire. Celle-ci appartient au domaine d'attraction de H_γ si et seulement s'il existe $\sigma > 0$ et $\gamma \in \mathbb{R}$ tels que la loi des excès F_u peut être approximée uniformément par une loi de Pareto généralisée $\mathcal{G}_{\gamma,\sigma}$. Plus précisément :*

$$\lim_{u \to x_F} \sup_{x \in]0, x_F - u[} |F_u(x) - \mathcal{G}_{\gamma,\sigma}(x)| = 0,$$

où $\mathcal{G}_{\gamma,\sigma}$ est la fonction de répartition de la loi de Pareto généralisée.

Par conséquent, la distribution limite des dépassements de seuils est la loi exponentielle qui est également la loi GPD de paramètre $\gamma = 0$ et $\sigma = 1$.



Fonction de répartition Fonction de densité

FIGURE 1.3 – Représentation graphique de la loi Pareto généralisée.

La distribution de Pareto généralisée est couramment utilisée pour modéliser les observations excédant un seuil u. L'un des principaux avantages de cette approche par rapport aux distributions des valeurs extrêmes réside dans le fait qu'elle permet de prendre en compte plusieurs observations qui dépassent le seuil, contrairement à la modélisation des seules valeurs maximales ou minimales sur une période donnée, comme c'est le cas dans l'approche Block Maxima. En effet, la méthode POT repose sur l'analyse des excès au-dessus du seuil, ce qui permet de modéliser plus efficacement les événements extrêmes sur la base de multiples observations.
Par ailleurs, la méthode POT présente l'avantage de nécessiter moins de paramètres à estimer que l'approche Block Maxima et est souvent préférable lorsqu'on dispose d'un nombre limité de données historiques. De plus, la manière dont la série des extrêmes est construite dans cette méthode permet d'obtenir des estimateurs plus efficaces que ceux de l'approche Block Maxima. Toutefois, l'une des difficultés majeures de la méthode POT réside dans le choix du seuil u, un problème similaire à celui de la sélection du paramètre k_n dans l'estimation des valeurs extrêmes généralisées.

1.4.1 Sélection du seuil

La détermination du seuil optimal constitue une étape cruciale de l'approche POT, car elle a une influence directe sur la qualité du modèle. En effet, la convergence des excès vers une loi de Pareto généralisée repose sur le choix d'un seuil adéquat, qui ne doit être ni trop bas ni trop élevé. Le seuil doit être suffisamment élevé pour exploiter les propriétés asymptotiques de la GPD, car un seuil trop bas pourrait entraîner des estimateurs biaisés. À l'inverse, un seuil trop élevé risquerait de rendre les estimations moins fiables, en augmentant les écarts-types des estimateurs.
Le choix du seuil repose souvent sur l'examen du graphique dit "Mean Excess Plot", qui permet d'identifier le seuil optimal en fonction du comportement des excès au-delà de ce seuil.

Définition 1.9 *Soit $X_1, X_2, \ldots, X_n$ un échantillon aléatoire provenant d'une distribution F, et u un seuil donné. Le "Mean Excess Plot" est le graphique des points $(u, e(u))$, où $e(u)$ est la moyenne des excès au-dessus du seuil u, définie par :*

$$e(u) = \mathbb{E}(X - u \mid X > u).$$

En pratique, la fonction moyenne des excès $e(u)$ est estimée par $\hat{e}_n(u)$, donnée par :

$$\hat{e}_n(u) = \frac{\sum_{i=1}^{n} x_i \mathbb{I}_{\{x_i>u\}}}{\sum_{i=1}^{n} \mathbb{I}_{\{x_i>u\}}} - u, \quad u < x_F,$$

où x_i représente les réalisations de la i-ème variable aléatoire X_i suivant la loi F, et $\mathbb{I}_{\{x_i>u\}}$ est la fonction indicatrice définie par :

$$\mathbb{I}_{\{x_i>u\}} = \begin{cases} 1 & \text{si } x_i > u, \\ 0 & \text{sinon.} \end{cases}$$

Proposition 8 *Si $X_1, \ldots, X_{N_u}$ suivent une loi $GPD_{\gamma,\sigma}$, alors, pour $\gamma < 1$, on a :*

$$\mathbb{E}(X - u \mid X > u) = \frac{\gamma}{1-\gamma} u + \frac{\sigma}{1-\gamma}, \quad \textit{avec} \quad \gamma u + \sigma > 0.$$

Il convient de noter que $\mathbb{E}(X - u \mid X > u)$ est une fonction linéaire de u.
Dans le cadre du choix du seuil, le graphique de la fonction moyenne des excès, ou "Mean Excess Plot", joue un rôle fondamental dans la détermination du seuil optimal. Plus précisément, le seuil est choisi comme étant la première valeur positive de u pour laquelle la courbe de la fonction moyenne des excès empirique $\hat{e}_n(u)$ devient approximativement linéaire.

Définition 1.10 *Le "Mean Excess Plot" (ME-plot) est défini par l'ensemble des points suivants :*

$$\left\{(u, \hat{e}_n(u)),\ x_{1:n} < u < x_{n:n}\right\},$$

où $x_{n:n}$ et $x_{1:n}$ désignent respectivement le maximum et le minimum de l'échantillon.

Lorsqu'on trace ce graphique, on détermine le seuil comme étant la première valeur positive de u pour laquelle la courbe de $\hat{e}_n(u)$ s'approxime à une droite.

Chapitre 2

Estimation des paramètres de la loi GEV et loi GPD

2.1 Introduction

L'estimation de l'indice des extrêmes et des paramètres de la loi de Pareto généralisée constitue un élément fondamental dans l'étude des phénomènes extrêmes. L'indice des extrêmes, qui décrit le comportement de la queue d'une distribution, permet de comprendre comment les événements extrêmes sont distribués. Son estimation offre une quantification de la probabilité d'occurrence de phénomènes rares d'une certaine intensité. Cet indice, noté γ, est essentiel pour élaborer des modèles prévisionnels des événements extrêmes, notamment dans des domaines tels que l'hydrologie, la climatologie et la finance.

Après l'estimation de ces paramètres, il devient possible de déterminer les quantiles extrêmes, c'est-à-dire les valeurs seuils au-delà desquelles des événements rares peuvent se produire avec une probabilité donnée. Ces quantiles sont cruciaux dans les analyses de risques, car ils permettent de prédire des événements extrêmes associés à des périodes de retour spécifiques. Dans ce cadre, l'estimation des quantiles extrêmes joue un rôle clé dans l'anticipation des impacts de tels événements et dans l'orientation des stratégies de gestion des risques, que ce soit pour les catastrophes naturelles, les fluctuations des marchés financiers ou la protection des infrastructures essentielles.

2.2 Méthode graphique

Dans cette section, nous étudions l'estimateur de l'indice extrême γ dans le cadre d'une loi de Pareto de distribution F, telle que :

$$1 - F(x) = x^{-\frac{1}{\gamma}} L_F(x), \quad x > 0,\ \gamma > 0$$

où L_F est une fonction à variation lente à l'infini.
Dans ce contexte, nous pouvons introduire la fonction de queue $U(x)$ définie par :

$$U(x) = \inf \left\{ y : F(y) \geq 1 - \frac{1}{x} \right\}.$$

On a alors l'expression suivante pour U :

$$U(x) = x^{\gamma} L_U(x),$$

où L_U est également une fonction à variation lente à l'infini.
En appliquant le logarithme à cette relation, nous obtenons :

$$\log U(x) = \gamma \log x + \log L_U(x) = \gamma \log x \left(1 + \frac{\log L_U(x)}{\gamma \log x}\right).$$

En utilisant les propriétés des fonctions à variation lente, on obtient que :

$$\frac{\log L_U(x)}{\log x} \longrightarrow 0 \quad \text{lorsque} \quad x \longrightarrow \infty.$$

Cela implique que :

$$\log U(x) \sim \gamma \log x \quad \text{lorsque} \quad x \longrightarrow \infty.$$

En remplaçant la fonction de queue U par sa version empirique $\hat{U}_n$, et en remarquant que $\hat{U}_n\left(\frac{n+1}{i}\right) = X_{n-i+1:n}$, nous obtenons l'équivalence suivante :

$$\log X_{n-i+1:n} \sim \gamma \log\left(\frac{n+1}{i}\right) \quad \text{lorsque} \quad \frac{n+1}{i} \longrightarrow \infty.$$

Définition 2.1 (Pareto Quantile Plot) *Soit $\{X_{1,n}, \ldots, X_{n,n}\}$ la statistique d'ordre associée à notre échantillon, le* Pareto quantile plot *est le graphique suivant :*

$$\left\{\left(\log\left(\frac{n+1}{i}\right), \log X_{n-i+1:n}\right) : i = 1, \ldots, n\right\}.$$

Le *Pareto quantile plot* est une représentation graphique très utile pour vérifier si les données suivent une loi du domaine de Fréchet. Dans ce cas, le graphique devrait approximativement être une droite avec une pente égale à γ pour les petites valeurs de i, c'est-à-dire pour les points extrêmes.

Remarque(Exponential Quantile Plot)
Le *exponential quantile plot* est une variation de la représentation précédente, mais appliquée au domaine de Gumbel ($\gamma = 0$). Il consiste à remplacer $\log x$ par x sur l'axe des ordonnées. Dans ce cas, la pente asymptotique dans le *exponential quantile plot* est égale au paramètre σ.

Définition 2.2 (Quantile Plot Généralisé) *Une approche permettant de contourner le choix a priori du domaine d'attraction a été proposée par Beirlant et al. (1996). Elle consiste à utiliser un* quantile plot généralisé*, défini par le graphique :*

$$\left\{\left(\log\left(\frac{n+1}{j}\right), \log UH_{j,n}\right) : j = 1, \ldots, n\right\}.$$

où $UH_{j,n}$ est défini par :

$$UH_{j:n} = X_{n-j+1:n}\left(j^{-1}\sum_{i=1}^{j} \log X_{n-i+1:n} - \log X_{n-j+1:n}\right).$$

En fonction de la courbure de ce graphique, il est possible de déduire dans quel domaine d'attraction les données se situent. Si les points extrêmes forment une droite de pente positive, on est alors dans le domaine de Fréchet. Si la courbe est plutôt constante, cela indique que les données appartiennent au domaine de Gumbel. Enfin, si la décroissance est linéaire, les données appartiennent au domaine de Weibull.

2.3 Estimation des paramètres de la loi GEV

2.3.1 Estimateur de Hill

L'estimateur de Hill a été introduit en 1975 [18] pour estimer non paramétriquement les lois appartenant au domaine d'attraction de Fréchet, c'est à dire pour les valeurs de $\gamma > 0$.

Définition 2.3 *Considérons une suite de variables aléatoires $X_1, X_2, \ldots, X_n$ de fonction de répartition $F \in D(H_\gamma)$ avec $\gamma > 0$. Soit k_n une suite d'entiers telle que $1 < k_n < n$. L'estimateur de Hill est défini par :*

$$\hat{\gamma}_n^H = \frac{1}{k_n} \sum_{i=1}^{k_n} \log(X_{n-i+1:n}) - \log(X_{n-k_n:n}).$$

Propriétés de l'estimateur de Hill

Proposition 9 *Soit $F \in D(H_\gamma)$ avec $\gamma > 0$. Si $k_n \to \infty$ et $\frac{k_n}{n} \to 0$ (k_n est dite intermédiaire) quand $n \to \infty$, alors on a :*

1. *Convergence en probabilité : Mason [25]*

$$\hat{\gamma}_n^H \xrightarrow{P} \gamma.$$

2. *Convergence presque sûre : Deheuvels et al. [10] Si de plus, $\frac{k_n}{\log\log n} \to \infty$ quand $n \to \infty$, alors*

$$\hat{\gamma}_n^H \xrightarrow{P.S} \gamma.$$

La normalité asymptotique de l'estimateur de Hill a fait l'objet de nombreux travaux. En effet, une hypothèse sur la fonction à variations lentes L est nécessaire, condition portant sur la vitesse de convergence du rapport des fonctions à variations lentes vers 1.

Définition 2.4 (Condition de variation du second ordre) *Il existe un paramètre $\rho < 0$ et une fonction $b(t) \to 0$ lorsque $t \to \infty$ tels que pour tout $\lambda > 1$, on a :*

$$\lim_{t\to\infty} \log\left(\frac{L(\lambda t)}{L(t)}\right) = b(t)\frac{\lambda^\rho - 1}{\rho}.$$

Si cette condition est vérifiée, on dit que la fonction L satisfait la condition du second ordre.

Remarque Le paramètre $\rho < 0$ contrôle la vitesse de convergence du ratio $L(\lambda t)/L(t)$ vers 1. Plus ρ est proche de 0, plus la convergence sera lente, rendant l'estimation de γ plus difficile.

Théorème 7 (Normalité asymptotique de l'estimateur de Hill) *(Beirlant et al [1]) Soit $(k_n)_{n\geq 1}$ une suite d'entiers telle que $1 < k_n \leq n$, $k_n \to \infty$ et $k_n/n \to 0$ quand $n \to \infty$. Supposons que la condition ci-dessus est satisfaite par la fonction auxiliaire de F avec $\sqrt{k_n} b(n/k_n) \to 0$ lorsque $n \to \infty$, alors :*

$$\sqrt{k_n}\left(\hat{\gamma}_n^H - \gamma\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \gamma^2).$$

Remarques

1. L'estimateur de Hill est biaisé et le résultat sur la normalité asymptotique de l'estimateur de Hill permet de construire un intervalle de confiance pour l'estimation de γ. Pour un niveau de confiance de $(1-\alpha)$, l'intervalle est donné par :
$$\gamma \in \left]\hat{\gamma}_n^H - t_{1-\alpha/2}\frac{\hat{\gamma}_n^H}{\sqrt{k_n}}, \hat{\gamma}_n^H + t_{1-\alpha/2}\frac{\hat{\gamma}_n^H}{\sqrt{k_n}}\right[$$
où $t_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale centrée et réduite.
2. Cet estimateur est influencé par le choix de la suite k_n :
• Si k_n est trop grand : l'approximation par une loi de Pareto sera mauvaise, le biais de $\hat{\gamma}_n^H$ est important.

• Si k_n est très petit : on aura peu d'observations pour l'estimation de γ, la variance de $\hat{\gamma}_n^H$ est importante.

Le bon choix de k_n est donc celui de meilleur compromis biais/variance de telle sorte : $k_n \to \infty$ et $k_n/n \to 0$ (assez grand mais pas trop grand).

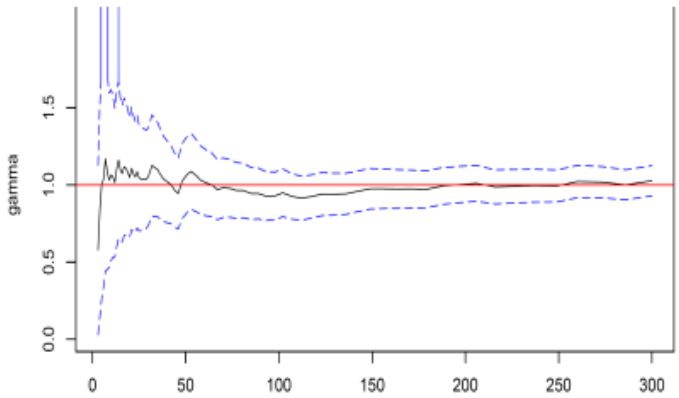


FIGURE 2.1 – L'estimateur de Hill avec un intervalle de confiance à 95% de γ basé sur 100 échantillons de taille 3000 pour la loi de Pareto standard.

2.3.2 Estimateur des moments

L'estimateur des moments, proposé par Dekkers et al. en 1989 ([8]), est une extension de l'estimateur de Hill, applicable pour tout $\gamma \in \mathbb{R}$.

Définition 2.5 *Considérons $X_1, X_2, \ldots, X_n$ une suite de variables aléatoires iid de fonction de répartition $F \in D(H_\gamma)$ avec $\gamma \in \mathbb{R}$. Soit k_n une suite d'entiers telle que $1 < k_n < n$. L'estimateur des moments est défini par :*
$$\hat{\gamma}_n^D = M^{(1)} + 1 - \frac{1}{2}\left(1 - \frac{(M^{(1)})^2}{M^{(2)}}\right)^{-1},$$
avec
$$M^{(a)} = \frac{1}{k_n}\sum_{i=1}^{k_n}(\log X_{n-i+1:n} - \log X_{n-k_n:n})^a, \quad a \in \{1,2\}.$$
On remarque que $M^{(1)}$ correspond à l'estimateur de Hill.

Propriétés de l'estimateur $\hat{\gamma}_n^D$ (Dekkers et al. [8])

Proposition 10 *Considérons $F \in D(H_\gamma)$, $\gamma \in \mathbb{R}$, avec $k_n \to \infty$ et $\dfrac{k_n}{n} \to 0$ lorsque $n \to \infty$, alors :*

1. *Convergence en probabilité :*
$$\hat{\gamma}_n^D \xrightarrow{P} \gamma.$$
2. *Convergence presque sûre : Si de plus, $k_n/(\log n)^\delta \to \infty$ quand $n \to \infty$ pour $\delta > 0$, alors :*
$$\hat{\gamma}_n^D \xrightarrow{P.S} \gamma \quad quand \quad n \to \infty.$$
3. *Normalité asymptotique : Sous certaines conditions sur la loi F :*
$$\sqrt{k_n}\left(\hat{\gamma}_n^D - \gamma\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_D^2) \quad quand \quad n \to \infty,$$
où
$$\sigma_D^2 = \begin{cases} 1+\gamma^2 & si\ \gamma \geq 0, \\ (1-\gamma^2)(1-2\gamma)\left(4 - 8\dfrac{1-2\gamma}{1-3\gamma} + \dfrac{(5-11\gamma)(1-2\gamma)}{(1-3\gamma)(1-4\gamma)}\right) & si\ \gamma < 0. \end{cases}$$

2.3.3 Estimateur de Pickands

L'estimateur de Pickands, introduit par James Pickands en 1975 [26], est valable quel que soit le domaine d'attraction de la distribution F.

Définition 2.6 *Considérons $X_1, X_2, \ldots, X_n$ une suite de variables aléatoires iid de fonction de répartition $F \in D(H_\gamma)$, et k_n une suite intermédiaire. L'estimateur de Pickands est défini par :*
$$\hat{\gamma}_n^P = \frac{1}{\log 2} \log\left(\frac{X_{n-k_n+1:n} - X_{n-2k_n+1:n}}{X_{n-2k_n+1:n} - X_{n-4k_n+1:n}}\right).$$

Propriétés de l'estimateur de Pickands

Proposition 11 *Supposons que $F \in D(H_\gamma)$, $\gamma \in \mathbb{R}$, et que k_n soit une suite intermédiaire.*

1. *Convergence en probabilité : Pickands [26]*
$$\hat{\gamma}_n^P \xrightarrow{P} \gamma \quad quand \quad n \to \infty.$$
2. *Convergence presque sûre : Dekkers et de Haan [9] Si de plus, $\lim\limits_{n\to\infty} \dfrac{k_n}{\log\log n} = \infty$, alors :*
$$\hat{\gamma}_n^P \xrightarrow{P.S} \gamma \quad quand \quad n \to \infty.$$
3. *Normalité asymptotique : Sous des conditions supplémentaires sur la suite k_n et la distribution F, Dekkers et de Haan [9] on a :*
$$\sqrt{k_n}\left(\hat{\gamma}_n^P - \gamma\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\gamma^2\left(2^{2\gamma+1}+1\right)}{4(\log 2)^2\left(2^\gamma - 1\right)^2}\right) \quad quand \quad n \to \infty.$$

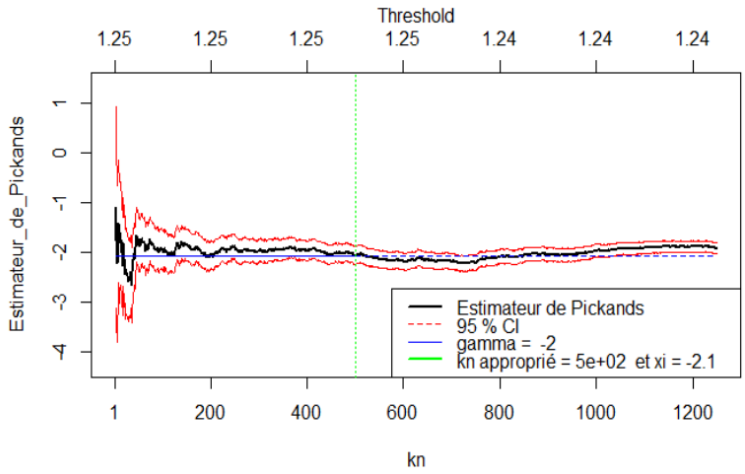


FIGURE 2.2 – Représentation graphique de l'estimateur de Pickands.

2.3.4 Autres estimateurs

Estimateur à noyau

Csörgo et al. (Voir [6]) ont proposé les estimateurs à noyau $\hat{\gamma}_n^N$. Cette classe d'estimateurs ne peut être utilisée que pour $\gamma > 0$, et est définie par :

$$\hat{\gamma}_n^N = \frac{\sum_{i=1}^m \frac{i}{m} K\left(\frac{i}{m}\right) \left(\log X_{(n-i+1):n} - \log X_{(n-i):n}\right)}{\sum_{i=1}^m \frac{i}{m} K\left(\frac{i}{m}\right)}.$$

où K est un noyau d'intégrale égale à 1, c'est-à-dire que la fonction $K(t)$ doit satisfaire la condition suivante :

$$0 \leq K(t) < \infty, \quad \forall t, \quad \text{et} \quad \int_{-\infty}^{+\infty} K(t)\, dt = 1.$$

Suivant le choix du noyau K, différents estimateurs peuvent en résulter, le plus connu étant l'estimateur de Hill, correspondant au cas particulier $K(x) = I_{(0,1]}(x)$, on a aussi le noyau d'Epanechnikov défini par :

$$K(t) = \frac{3}{2}(1 - t^2), \text{ si } t \in [0, 1]$$

La consistance et la normalité asymptotique de cet estimateur ont été établies par Groeneboom et Lopuhaa (Voir [16]).

Estimateur de Zipf

Schultze et Steinebach [29] ont proposé d'estimer $\gamma > 0$ par la méthode des moindres carrés classiques. Leur estimateur, connu sous le nom de *estimateur de Zipf*, est défini par :

$$\hat{\gamma}_n^Z = \frac{\sum_{i=1}^m \log \frac{m+1}{i} \log X_{(n-i+1):n} - \frac{1}{m} \sum_{i=1}^m \log \frac{m+1}{i} \sum_{i=1}^m \log X_{(n-i+1):n}}{\sum_{i=1}^m \log^2 \frac{m+1}{i} - \frac{1}{m} \left(\sum_{i=1}^m \log \frac{m+1}{i}\right)^2}.$$

Cet estimateur a été proposé dans le but d'améliorer le biais asymptotique des estimateurs de Hill et de Pickands. Cependant, la variance de cet estimateur est deux fois supérieure à celle de l'estimateur de Hill.

Théorème 8 (Normalité asymptotique de l'estimateur de Zipf) *Soient $F \in DA(H_\gamma)$, avec $\gamma > 0$, vérifiant la variation régulière de second ordre, et k une suite intermédiaire telle que $\lim_{n\to\infty} \sqrt{k}\, b\left(\frac{n}{k}\right) = \lambda$, où $\lambda \in \mathbb{R}$, alors :*

$$\sqrt{k}\left(\hat{\gamma}_n^Z - \gamma\right) \xrightarrow{\mathcal{L}} \mathcal{N}(\mu, \sigma^2),$$

où

$$\sigma^2 = \begin{cases} 2(1+\gamma^2+\gamma) & si\ \gamma \geq 0, \\ \dfrac{2(1-\gamma)(1+2\gamma+\gamma^2-2\gamma^3)}{(1-2\gamma)(1-\gamma)} & si\ \gamma < 0, \end{cases}$$

et

$$\mu = \frac{\lambda}{(1-\tilde{\rho})^2}.$$

Analyse comparative des estimateurs
L'estimateur de Hill est simple et adapté aux queues lourdes ($\gamma > 0$), mais très sensible au choix du seuil et inutilisable pour les queues légères. L'estimateur de Moment est plus robuste, fonctionne pour $\gamma > -0.5$, et offre une meilleure stabilité que Hill. L'estimateur de Pickands, basé sur des quantiles extrêmes, est simple mais souvent instable, surtout avec peu de données. L'estimateur à noyau lisse les données extrêmes via une fonction noyau, ce qui réduit la variance et améliore la robustesse, au prix d'une complexité de mise en œuvre. Enfin, l'estimateur de Zipf, basé sur une représentation log-log des rangs, est rapide et visuel, utile en exploration mais moins précis. Le choix dépend du type de queue, de la taille de l'échantillon et du niveau de précision recherché.(voir [5], [3], [27])

2.4 Estimation des paramètres de la loi GPD

Dans cette section, nous nous intéressons à l'estimation des paramètres de la loi de Pareto généralisée.

2.4.1 Méthode du maximum de vraisemblance

Considérons un échantillon $Y_1, Y_2, \ldots, Y_{k_n}$, i.i.d., issu d'une loi de GPD $G_{\gamma,\sigma}$. La fonction de vraisemblance est alors définie par :

$$\mathcal{L}(\gamma, \sigma) = \prod_{i=1}^{k_n} g_{\gamma,\sigma}(y_i).$$

Nous cherchons à maximiser cette fonction de vraisemblance, ce qui revient à maximiser la fonction de log-vraisemblance suivante :

$$\log \mathcal{L}(\gamma, \sigma; y_1, y_2, \ldots, y_{k_n}) = \sum_{i=1}^{k_n} \log\left(g_{\gamma,\sigma}(y_i)\right).$$

En remplaçant par l'expression de la densité de la GPD, on obtient :

$$= \sum_{i=1}^{k_n} \log\left(\frac{1}{\sigma}\left(1+\frac{\gamma}{\sigma}y_i\right)^{-\frac{1}{\gamma}-1}\right)$$

$$= -k_n \log\sigma - \left(\frac{1}{\gamma}+1\right)\sum_{i=1}^{k_n} \log\left(1+\frac{\gamma}{\sigma}y_i\right),$$

avec $1 + \frac{\gamma}{\sigma} y_i > 0$ pour tout $i = 1, \ldots, k_n$.
Si $\gamma = 0$, la fonction de log-vraisemblance, correspond à une loi exponentielle donnée par :

$$\log \mathcal{L}(\sigma, 0) = -k_n \log \sigma - \frac{1}{\sigma} \sum_{i=1}^{k_n} y_i.$$

Pour faciliter les calculs, on peut effectuer une reparamétrisation des paramètres (γ, σ) en introduisant $\tau = \frac{\gamma}{\sigma}$. Ainsi, la fonction de log-vraisemblance devient :

$$\log \mathcal{L}(\tau, \gamma) = -k_n \log \gamma + k_n \log \tau - \left(\frac{1}{\gamma} + 1\right) \sum_{i=1}^{k_n} \log\left(1 + \tau y_i\right).$$

Les estimateurs $\hat{\tau}_n^{ML}$ et $\hat{\gamma}_n^{ML}$ peuvent être obtenus en résolvant l'équation suivante :

$$\frac{1}{\hat{\tau}_n^{ML}} - \left(\frac{1}{\hat{\gamma}_n^{ML}} + 1\right) \frac{1}{k_n} \sum_{i=1}^{k_n} \frac{y_i}{1 + \hat{\tau}_n^{ML} y_i} = 0.$$

Cela conduit à l'estimateur de γ :

$$\hat{\gamma}_n^{ML} = \frac{1}{k_n} \sum_{i=1}^{k_n} \log\left(1 + \hat{\tau}_n^{ML} y_i\right).$$

Propriétés de l'estimateur $\hat{\gamma}_n^{ML}$ Sous quelques conditions (voir [27]), on a les convergences suivantes :

1. **Convergence en probabilité :**

$$\hat{\gamma}_n^{ML} \xrightarrow{P} \gamma \quad \text{lorsque} \quad n \to \infty.$$

2. **Convergence presque sûre :**

$$\hat{\gamma}_n^{ML} \xrightarrow{P.S} \gamma \quad \text{lorsque} \quad n \to \infty.$$

3. **Normalité asymptotique :**

$$\sqrt{n}\left(\hat{\gamma}_n^{ML} - \gamma, \frac{\hat{\sigma}_n}{\sigma} - 1\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, M^{-1}),$$

où

$$M^{-1} = (1 + \gamma) \begin{pmatrix} 1 + \gamma & -1 \\ -1 & 2 \end{pmatrix}.$$

2.4.2 Méthode des moments

La méthode des moments, introduite par Hosking et Wallis en 1987 [19], permet d'estimer les paramètres de la loi GPD. Pour que l'espérance et la variance de la variable aléatoire Y de loi GPD $G_{\gamma,\sigma}$ existent, il faut que $\gamma < 1/2$. Dans ce cas, on a :

$$\mathbb{E}(Y) = \frac{\sigma}{1 - \gamma} \quad \text{et} \quad \mathbb{V}(Y) = \frac{\sigma^2}{(1 - \gamma)^2 (1 - 2\gamma)}.$$

Les paramètres γ et σ peuvent alors être exprimés en fonction de l'espérance et de la variance de Y :

$$\gamma = \frac{1}{2}\left(1 - \frac{\mathbb{E}(Y)^2}{\mathbb{V}(Y)}\right) \quad \text{et} \quad \sigma = \frac{\mathbb{E}(Y)}{2}\left(1 + \frac{\mathbb{E}(Y)^2}{\mathbb{V}(Y)}\right).$$

En remplaçant $\mathbb{E}(Y)$ et $\mathbb{V}(Y)$ par leurs estimateurs empiriques :

$$\bar{Y} := \frac{1}{k_n}\sum_{i=1}^{k_n} Y_i \quad \text{et} \quad S^2 := \frac{1}{k_n - 1}\sum_{i=1}^{k_n}(Y_i - \bar{Y})^2,$$

on obtient les estimateurs des moments de γ et σ :

$$\hat{\gamma}_n^{MOM} = \frac{1}{2}\left(1 - \frac{\bar{Y}^2}{S^2}\right) \quad \text{et} \quad \hat{\sigma}_n^{MOM} = \frac{\bar{Y}}{2}\left(1 + \frac{\bar{Y}^2}{S^2}\right).$$

Propriétés de l'estimateur $\hat{\gamma}_n^{MOM}$ Sous quelques conditions (voir [27]), on a les convergences suivantes :

1. **Convergence en probabilité :**

$$\hat{\gamma}_n^{MOM} \xrightarrow{P} \gamma \quad \text{lorsque} \quad n \to \infty.$$

2. **Convergence presque sûre :**

$$\hat{\gamma}_n^{MOM} \xrightarrow{P.S} \gamma \quad \text{lorsque} \quad n \to \infty.$$

2.4.3 Méthode des moments pondérés

Hosking et Wallis (Voir [19]) ont proposé en 1987 une méthode basée sur les moments pondérés pour estimer les paramètres de la GPD. Soit $Y_1, \ldots, Y_n$ un échantillon de n variables aléatoires i.i.d. de fonction de répartition $\mathcal{G}_{\gamma,\sigma}$. Les statistiques d'ordre associées sont données par $Y_{1,k_n} \leq \cdots \leq Y_{k_n,k_n}$.
Le moment pondéré d'ordre s de la variable Y de loi GPD $\mathcal{G}_{\gamma,\sigma}$ est défini par :

$$\mu_s = M(1,0,s) = \mathbb{E}\left[Y(1 - \mathcal{G}_{\gamma,\sigma}(Y))^s\right] = \frac{\sigma}{(1+s)(1+s-\gamma)} \quad \text{avec} \quad \gamma < 1.$$

Les estimateurs des moments pondérés peuvent être obtenus en résolvant les équations suivantes :

$$\mu_0 = \frac{\sigma}{1-\gamma} \quad \text{et} \quad \mu_1 = \frac{\sigma}{2(2-\gamma)}.$$

Les paramètres γ et σ peuvent alors être exprimés en fonction de μ_0 et μ_1 :

$$\gamma = \frac{\mu_0 - 4\mu_1}{\mu_0 - 2\mu_1} \quad \text{et} \quad \sigma = \frac{2\mu_0\mu_1}{\mu_0 - 2\mu_1}.$$

En remplaçant les moments μ_s, pour $s \in \{0,1\}$, par leurs estimateurs empiriques :

$$\hat{\mu}_s := \frac{1}{k_n}\sum_{i=1}^{k_n}\left(1 - \frac{i}{1+k_n}\right)^r Y_{i:k_n},$$

on obtient alors les estimateurs pondérés de γ et σ :

$$\hat{\gamma}_n^{MOP} = 2 - \frac{\hat{\mu}_0}{\hat{\mu}_0 - 2\hat{\mu}_1} \quad \text{et} \quad \hat{\sigma}_n^{MOP} = \frac{2\hat{\mu}_0\hat{\mu}_1}{\hat{\mu}_0 - 2\hat{\mu}_1}.$$

Les propriétés de l'estimateur $\hat{\gamma}_n^{MOP}$ Sous quelques conditions (voir [27]), on a les convergences suivantes :

1. **Convergence en probabilité :**

$$\hat{\gamma}_n^{MOP} \xrightarrow{P} \gamma \quad \text{lorsque} \quad n \to \infty.$$

2. **Convergence presque sûre :**

$$\hat{\gamma}_n^{MOP} \xrightarrow{P.S} \gamma \quad \text{lorsque} \quad n \to \infty.$$

2.5 Estimation des quantiles extrêmes

Rappelons que le quantile d'ordre p de la fonction de répartition F est défini par :

$$q(p) = q_p = F^{\leftarrow}(p) = \inf\left\{x \in \mathbb{R} : \bar{F}(x) \leq p\right\}, \quad \text{avec } p \in]0, 1]$$

où $F^{\leftarrow}$ désigne l'inverse généralisé de la fonction de répartition F, et $\bar{F}(x) = 1 - F(x)$ est la fonction de survie de X.

Définition 2.7 (Fonction empirique)
Soit $X_1, X_2, \ldots, X_n$ un échantillon de variables aléatoires indépendantes et identiquement distribuées (i.i.d.) de fonction de répartition F (continue). Les statistiques d'ordre associées sont notées $X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$. La fonction de répartition empirique F_n est définie par :

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leq x\}} = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_{i:n} \leq x\}},$$

avec la forme suivante selon la valeur de x :

$$F_n(x) = \begin{cases} 0 & si\ x < X_{1:n}, \\ \frac{i-1}{n} & si\ X_{i-1:n} \leq x < X_{i:n}, \\ 1 & si\ x \geq X_{n:n}. \end{cases}$$

On peut estimer le quantile q_p en inversant la fonction de répartition empirique :

$$\hat{q}(p) = F_n^{\leftarrow}(1-p) = \inf\left\{x : F_n(x) \geq 1 - p\right\}.$$

Définition 2.8 (Quantile extrême)
Le quantile extrême d'ordre $1 - p$ de la fonction de répartition F est défini par :

$$q(1-p) = q_{1-p} = \inf\left\{x \in \mathbb{R} : \bar{F}(x) \leq 1 - p\right\} = F^{\leftarrow}(1-p),$$

avec :

$$1 - p \to 0 \quad \textit{lorsque} \quad n \to \infty.$$

Les quantiles extrêmes sont particulièrement importants lorsqu'il s'agit d'estimer des événements rares, c'est-à-dire les quantiles associés aux valeurs extrêmes d'une distribution. Ces quantiles sont utilisés pour estimer les niveaux de retour, c'est-à-dire les valeurs des quantiles au-delà desquelles un événement rare (tel qu'une inondation, un tremblement de terre, etc.) est susceptible de se produire avec une certaine probabilité.

La méthode *POT* (Peak Over Threshold) s'appuie sur le théorème de Balkema-de Haan-Pickands pour estimer q_{1-p}. L'estimateur de ce quantile est obtenu en inversant la fonction de répartition de la loi de Pareto généralisée et en estimant les paramètres de cette loi à l'aide des observations supérieures au seuil u.

$$\begin{aligned} F_u(x) &= \frac{F(u+x) - F(u)}{1 - F(u)} \quad \text{si } x \geq 0, \\ &\Leftrightarrow F_u(x-u) = 1 - \frac{1 - F(x)}{1 - F(u)} \quad \text{si } x \geq u, \\ &\Leftrightarrow \bar{F}(x) = \bar{F}(u) \cdot \bar{F}_u(x-u) \quad \text{si } x \geq u. \end{aligned}$$

Sachant que :

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

où N_u est le nombre d'observations supérieures au seuil u, et d'après le théorème de Balkema-Pickands-de Haan, on peut approximer la distribution par :

$$F_u(x) \approx \mathcal{G}_{\gamma,\sigma}(x),$$

et de plus :

$$\bar{F}_u(x) \approx 1 - \mathcal{G}_{\gamma,\sigma}(x),$$

où $\mathcal{G}_{\gamma,\sigma}$ est la fonction de répartition de la loi de Pareto généralisée.
L'estimateur de la fonction de survie $\bar{F}$ peut alors être écrit comme :

$$\widehat{\bar{F}}(x) = \frac{N_u}{n}\left(1 + \widehat{\gamma}\frac{x-u}{\widehat{\sigma}}\right)^{-\frac{1}{\widehat{\gamma}}}, \quad \forall \gamma \neq 0.$$

Une fois les paramètres γ et σ estimés, on peut obtenir une estimation du quantile extrême q_{1-p} en inversant la fonction de répartition de la loi de Pareto généralisée. L'estimateur du quantile extrême est donné par :

$$\widehat{q}_{1-p}^{\text{GPD}} = u + \frac{\widehat{\sigma}}{\widehat{\gamma}}\left[\left(\frac{n}{N_u}(1-p)\right)^{-\widehat{\gamma}} - 1\right], \ \gamma \neq 0.$$

Cet estimateur permet d'obtenir une estimation du quantile extrême pour une probabilité $1-p$ très faible (typiquement associée à un événement rare), ce qui est utile pour l'estimation des niveaux de retour dans les phénomènes extrêmes.

Chapitre 3

Application sur des données en hydrologie

3.1 Introduction

L'étude des événements extrêmes, comme les crues, les tempêtes ou les vagues de chaleur, est essentielle pour évaluer les risques liés à des phénomènes rares mais potentiellement dévastateurs pour les infrastructures, l'environnement et les populations. Une question centrale est d'estimer la probabilité qu'un événement dépasse un certain seuil d'intensité sur une période donnée, souvent exprimée par la période de retour, c'est-à-dire l'intervalle moyen entre deux événements similaires.
L'estimation des quantiles extrêmes, liés à ces périodes, permet de quantifier l'intensité attendue des événements rares. Elle joue un rôle clé dans la gestion des risques, la planification des infrastructures et l'élaboration de politiques publiques.
Une méthode largement utilisée repose sur la loi des valeurs extrêmes généralisées, qui permet d'extrapoler le comportement des événements rares à partir de données historiques. Des outils comme les quantile plots et les ajustements de modèles facilitent cette estimation, améliorant ainsi la compréhension et la prévision des risques extrêmes.

3.2 Ajustement du modèle

L'étude des débits extrêmes est fondamentale en hydrologie pour comprendre les phénomènes rares comme les crues. Deux types de débits sont particulièrement étudiés :
- Les débits d'écoulement (souvent influencés par les activités humaines),
- Les débits de rivière (fortement dépendants de la dynamique naturelle).
La théorie des valeurs extrêmes est un outil statistique souvent appliqué à ces phénomènes, pour évaluer le risque des inondations.
Dans tous les cas,la première chose à faire est une distinction entre les domaines $\gamma > 0$, $\gamma = 0$ *et* $\gamma < 0$. Signalons cependant que le domaine $\gamma < 0$ est très rare en hydrologie puisque cela revient à dire que la distribution a une limite finie. Les données de précipitations montrent le plus souvent une distribution avec un indice nul ou positif strict (Buishand, 1989 ; Harremoes et Mikkelsen, 1995). Par conséquent, les débits de rivières et les écoulements n'apparaissent pas avec des limites supérieures, sauf s'il y a eu des influences humaines qui les limitent ou des inondations qui réduisent les débits de pointes.
Exemple 1
Nous simulons un échantillon de $n = 50$ crues maximales annuelles selon une loi de Fréchet, caractérisée par un paramètre de forme $\gamma = 0,25$. Cette valeur correspond à une queue lourde, typique des phénomènes extrêmes en hydrologie.
Nous utilisons le graphique quantile de Pareto pour vérifier si les données simulées suivent un

comportement du domaine de Fréchet. Ce graphique trace :

$$x_j = \log\left(\frac{n+1}{j}\right), \quad y_j = \log(Q_{(n-j+1)})$$

où $Q_{(n-j+1)}$ est la j-ième plus grande valeur dans l'échantillon qui représente X_{n-j+1}

— Si les points s'alignent sur une droite, cela indique un bon ajustement au modèle de Fréchet.

— La pente de cette droite est une estimation du paramètre γ.

Le code R

```r

install.packages("evd")
library(evd)

# Simulation de crues maximales annuelles (par exemple en m³/s)
set.seed(123)
n <- 50  # 50 années
gamma_true <- 0.25  # queue assez lourde
Q_max <- rfrechet(n, shape = 1 / gamma_true)

# Ordre décroissant
Q_sorted <- sort(Q_max)
log_Q <- log(rev(Q_sorted))  # log(Q_(n-j+1))
log_rank <- log((n + 1) / (1:n))  # log((n+1)/j)

# Graphe
plot(log_rank, log_Q, type = "p", col = "black",
     xlab = "log((n+1)/j)", ylab = "log(Q_{(n-j+1)})",
     main = "Pareto Quantile Plot - Crues maximales annuelles")
abline(lm(log_Q ~ log_rank), col = "blue", lwd = 2)

# Estimation de gamma par régression linéaire
coef(lm(log_Q ~ log_rank))  # pente  gamma


```

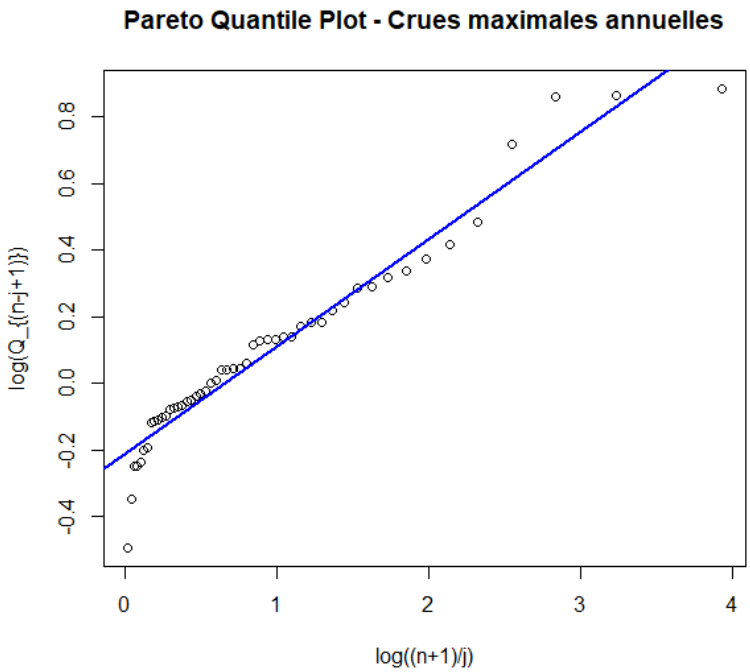


FIGURE 3.1 – Graphique quantile de Pareto pour des crues simulées selon une loi de Fréchet ($\gamma = 0{,}25$)

Le graphique montre que la queue de la distribution des crues maximales annuelles simulés, tendent à s'aligner le long d'une droite, ce qui valide l'hypothèse d'une loi à queue lourde. La pente de la droite de régression est proche de 0,25, ce qui est cohérent avec la valeur utilisée pour la simulation.
Le résultat indique une pente estimée de 0.3234, soit $\gamma \approx 0.32$. Ce résultat confirme que le modèle de Fréchet est adapté à la modélisation de crues maximales annuelles.

(Intercept)	log_rank
-0.2145101	0.3234028

Exemple 2(Cas de loi de Gumbel)
Toutefois, dans certains cas, notamment pour des crues modérées ou des séries de données moins extrêmes, un ajustement par la loi de Gumbel (cas $\gamma = 0$) peut être plus approprié. Cette loi, à queue exponentielle, est également un cas particulier du modèle GEV et fournit une alternative efficace lorsque la queue des données est ni trop lourde ni trop bornée.
Pour vérifier si les données suivent une loi de Gumbel, on peut utiliser un Gumbel quantile plot, où l'on trace les valeurs observées $X_{(j)}$ triées par ordre croissant en fonction de la transformation suivante :

$$x_j = -\log\left(-\log\left(\frac{j}{n+1}\right)\right)$$

Si les données suivent effectivement une loi de Gumbel, les points du graphique devraient s'aligner approximativement sur une droite.
Nous avons simulé un échantillon de 100 valeurs extrêmes à partir d'une loi de Gumbel.
Le code R

```r

install.packages("evd")
library(evd)

# Simulation d'un échantillon suivant une loi de Gumbel
set.seed(123)
n <- 100
mu <- 10       # paramètre de localisation
sigma <- 2     # paramètre d'échelle
u <- runif(n) # génération uniforme
data_gumbel <- -log(-log(u)) * sigma + mu

# Transformation pour le Gumbel plot
j <- 1:n
x <- -log(-log(j / (n + 1)))        # axe x transformé
y <- sort(data_gumbel)              # ordre croissant des observations

# Gumbel Quantile Plot
plot(x, y, main = "Gumbel Quantile Plot",
     xlab = "-log(-log(j / (n+1)))", ylab = "X_(j)",
     pch = 19, col = "darkgreen")
abline(lm(y ~ x), col = "blue", lwd = 2)

```

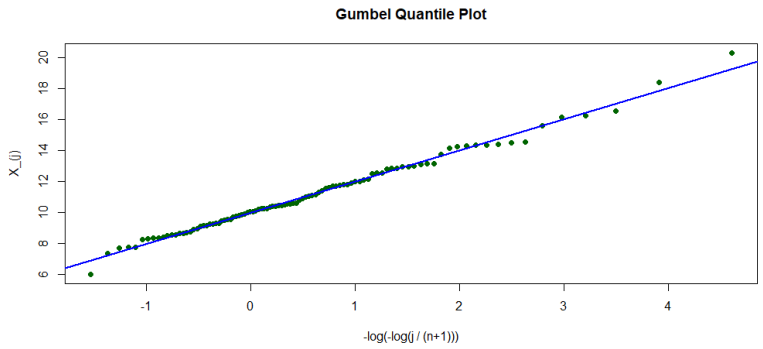


FIGURE 3.2 – Graphique quantile de Pareto pour des crues simulées selon une loi de Gumbel

3.2.1 Données brutes

Pour déterminer le débit d'une rivière, il faut d'abord mesurer les hauteurs d'eau, puis les convertir en débit à l'aide d'une courbe de tarage. Une courbe de tarage est une fonction empirique ou modélisée qui établit la relation entre la hauteur de l'eau mesurée à une station hydrométrique et le débit correspondant. Elle permet ainsi de transformer des mesures simples de hauteur en estimations du débit, sans avoir à effectuer des mesures complexes à chaque instant.
Cependant, dans le cas des rivières, il est difficile de mesurer directement le débit, qui correspond à un volume d'eau écoulé par unité de temps. On peut toutefois mesurer les vitesses d'écoulement à différents points de la section de la rivière, puis les intégrer pour obtenir le débit

global. Ces mesures sont effectuées pendant de courtes périodes, et les vitesses sont supposées constantes sur toute la section.
Le débit ainsi calculé est considéré comme instantané, car les mesures sont faites sur de courts intervalles de temps. En regroupant ces données selon les hauteurs d'eau, on construit alors une courbe de tarage fiable qui permet, à partir des hauteurs observées, d'estimer en continu les débits de la rivière.

3.2.2 Sélection des observations indépendantes

Dans cet étude, il est important de considérer des événements extrêmes de manière indépendante pour appliquer correctement les théories statistiques, notamment la théorie des valeurs extrêmes. Les pics d'inondation consécutifs seront considérés comme indépendants si deux conditions sont remplies : d'une part, l'intervalle de temps entre ces deux pics doit dépasser un temps critique, et d'autre part, le débit entre ces deux événements doit descendre en dessous d'un niveau proche du débit de base. Ce dernier est interprété comme un débit « normal », dans le sens où il ne correspond ni à une période d'inondation ni à une période de sécheresse. De nombreux critères de sélection des événements extrêmes ont été proposés dans la littérature (cf. par exemple USWRC, 1976 [30] ; Lang et al., 1999 [24] ; Claps et Laio, 2003 [4]). Nous allons détailler ci-dessous le critère que nous avons choisi pour cette étude et l'illustrer dans la figure 1.3.

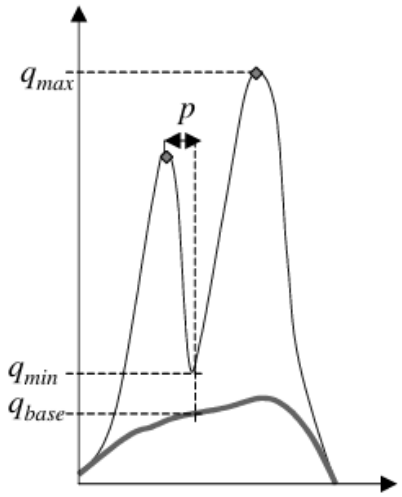


FIGURE 3.3 – Paramètres utilisés dans le critère de sélection des valeurs POT indépendantes.

Plus précisément, deux pics consécutifs seront considérés comme indépendants lorsque les deux conditions suivantes sont satisfaites :

$$p > k, \quad \frac{q_{\min}}{q_{\max}} < f, \quad q_{\max} > q_{\lim}.$$

Interprétation des conditions :
1. **Le critère** • $\frac{q_{\mathbf{min}}}{q_{\mathbf{max}}} < f$: est un critère pour évaluer la variabilité hydrologique.
$q_{\min}$:Crue minimale : Le débit ou le niveau d'eau le plus bas mesuré sur une période donnée.
$q_{\max}$: Crue maximale : Le débit ou le niveau d'eau le plus haut mesuré sur la même période.
$q_{\lim}$:Crue limite représente un seuil ou une valeur de débit (ou de niveau) que l'on considère comme critique pour un système hydrologique donné.

f : fraction de pic.
Si $\frac{\text{Crue minimale}}{\text{Crue maximale}}$ est très petit (inférieur à f), cela signifie qu'il existe une grande variabilité entre les crues minimales et maximales, avec des périodes où les débits sont extrêmement faibles par rapport aux crues importantes. Cela reflète un régime hydrologique très variable, souvent typique des cours d'eau saisonniers ou soumis à des conditions climatiques extrêmes (alternance de fortes pluies et de sécheresses).
Si f est très bas, cela peut être un indicateur de pénurie hydrique en période de basses eaux, même si des crues importantes surviennent.
- Un rapport inférieur à f peut nécessiter une gestion spécifique pour éviter les effets négatifs (comme les inondations en crue maximale ou les pénuries en crue minimale).

2.**Le critère $q_{\max} > q_{\lim}$**

- Ce seuil est souvent déterminé en fonction :
- De la capacité des infrastructures hydrauliques (digues, barrages, canaux).
- De la topographie et de l'occupation des sols (zones habitées ou agricoles).
- Des risques acceptables pour les populations et les activités économiques.
Lorsque la crue maximale dépasse la crue limite, cela signifie que l'événement de crue est plus intense que prévu ou tolérable, ce qui peut entraîner des conséquences graves sur les plans humain, matériel ou environnemental.
3.**Le critère L'inégalité** $p > k$, où :
• p représente le temps entre deux pics (par exemple, entre deux crues maximales ou deux événements de forte intensité hydrologique), k est le facteur de récession (un paramètre décrivant la vitesse à laquelle un cours d'eau revient à son débit de base après un événement de crue). Il indique la capacité du bassin versant ou du cours d'eau à évacuer l'eau excédentaire après une crue.
• Plus k est élevé, plus le débit revient lentement à la normale (reflétant une récession lente, comme dans les bassins à faible pente ou avec de grands réservoirs naturels).
• Si k est faible, l'eau s'évacue rapidement et le bassin retourne vite à des conditions de base.
le critère $p > k$ signifie que :
• Le débit du cours d'eau revient à son niveau de base avant que la crue suivante ne se produise.
• Cela traduit un système hydrologique résilient où les événements hydrologiques sont bien espacés.
• Les débits de crue successifs ne se cumulent pas, limitant les risques d'inondation prolongée ou sévère.

Exemple illustratif
- Crue minimale / Crue maximale : Supposons Crue minimale = $50m^3/s$ et Crue maximale= $500m^3/s$ avec $f = 0.3$.

$$\frac{\text{Crue minimale}}{\text{Crue maximale}} = \frac{50}{500} = 0.1 < 0.3$$

- Crue maximale > Crue limite : Si Crue limite = $400\,\text{m}^3/\text{s}$, alors Crue maximale = $500\,\text{m}^3/\text{s} > 400\,\text{m}^3/\text{s}$.
- Temps entre deux pics $p > k$: Si $p = 15\,\text{jours}$ (temps entre deux crues) et $k = 10\,\text{jours}$ (temps de récession), alors $p > k$, indiquant que le bassin peut se stabiliser entre deux crues.

Exemple
Dans cet exemple, une série temporelle simulée de débits journaliers a été générée pour représenter un comportement hydrologique réaliste sur une période de 2 ans (du 1er janvier 2020 au 31 décembre 2021).

Ces données simulées servent à tester la méthode de détection de crues indépendantes présentée ci-dessus en l'absence de mesures réelles.

Le code R

```r
#  Charger les packages
library(tidyverse)

#  1. Simuler des données de débit journalier sur 2 ans (ex. 2020-2021)
set.seed(123)  # Pour reproductibilité
n_days <- 365 * 2
dates <- seq.Date(from = as.Date("2020-01-01"), by = "day", length.out = n_days)

# Générer des débits avec une composante saisonnière + bruit
debits <- 10 +
  5 * sin(2 * pi * 1:n_days / 365) +  # saisonnalité
  rgamma(n_days, shape = 2, scale = 2)  # bruit + extrêmes

# Ajouter quelques crues artificielles (pics forts)
crue_days <- sample(100:n_days, 10)
debits[crue_days] <- debits[crue_days] + runif(10, 30, 60)

# Créer le data frame simulé
data_sim <- data.frame(Date = dates, Debit = debits)

#  2. Paramètres pour la détection des crues
qlim <- quantile(data_sim$Debit, 0.95) e
f <- 0.3
k <- 3

#  3. Détection des pics
find_peaks <- function(x) {
  which(diff(sign(diff(x))) == -2) + 1
}
pic_indices <- find_peaks(data_sim$Debit)
pics <- data_sim[pic_indices, ]

# 4. Filtrage des pics indépendants
selected <- data.frame()

for (i in 2:nrow(pics)) {
  qmax <- pics$Debit[i]
  date_max <- pics$Date[i]
  date_prev <- pics$Date[i - 1]
  p_days <- as.numeric(date_max - date_prev)

  i1 <- pic_indices[i - 1]
  i2 <- pic_indices[i]
  qmin <- min(data_sim$Debit[i1:i2], na.rm = TRUE)

  if (qmax > qlim && (qmin / qmax) < f && p_days > k) {
```

```
      selected <- rbind(selected, pics[i, ])
    }
  }
  5. Résultat
  print(selected)
```

Résultat de la sélection des données indépendantes :

Index	Date	Débit (m^3/s)
172	2020-06-20	60.23
310	2020-11-05	64.09
355	2020-12-20	53.15
519	2021-06-02	54.88
561	2021-07-14	66.52

Ce tableau montre 5 événements de crue significatifs (les pics de crues jugés indépendants et extrêmes), selon tes critères hydrologiques.

— Le débit dépasse le 95% de la série, noté q_{lim} ;

— Le minimum entre deux pics successifs est suffisamment bas, c'est-à-dire que :

$$\frac{q_{\text{min}}}{q_{\text{max}}} < 0.3,$$

ce qui garantit une bonne séparation hydraulique entre les crues ;

— Les pics sont séparés par au moins $k = 3$ jours.

3.3 Estimation de Période de Retour

Définition 3.1 (Période de retour) *La période de retour T est définie comme l'intervalle de temps moyen entre deux événements dont l'intensité dépasse un certain seuil donné.*

Par exemple, une crue centennale, qui se produit en moyenne une fois tous les 100 ans, a une période de retour de 100 ans, ce qui implique qu'il y a une probabilité de 1 % chaque année qu'une crue de cette ampleur ou plus se produise.

Définition 3.2 (Niveau de retour) *Le niveau de retour p_t est le seuil au-delà duquel la variable aléatoire X dépasse une certaine valeur avec une probabilité $p = \frac{1}{T}$, ce qui correspond à la probabilité d'occurrence d'un événement plus intense que ce seuil sur une période de temps donnée.*

La relation entre la période de retour T, le niveau de retour p_t, et la probabilité qu'un événement dépasse un seuil fixé peut être formulée de la manière suivante.
Si t événements sont observés, on s'attend à ce qu'en moyenne un seul dépassement du seuil p_t se produise. Cela implique que la probabilité qu'un événement dépasse p_t est égale à $\frac{1}{t}$, ce qui relie directement la période de retour à la probabilité d'occurrence d'un événement extrême. Cette

relation peut être exprimée à travers la fonction de répartition $F(p_t)$ de la variable aléatoire X :

$$\begin{aligned}\mathbb{E}\left(\sum_{i=1}^{t}\mathbb{I}_{\{X_i>p_t\}}\right) &= 1 \Leftrightarrow \mathbb{P}\left(X_i > p_t\right) = \frac{1}{t} \text{ avec } i = 1, ..., t \\ &\Leftrightarrow 1 - F(p_t) = \frac{1}{t}\end{aligned}$$

Estimation de la période de retour

L'estimation d'un niveau de retour d'ordre t consiste à déterminer un quantile extrême d'ordre $q_p = 1 - \frac{1}{t}$. Cela peut être réalisé à l'aide de deux approches principales : l'approche GPD et l'approche GEV. Bien que reposant sur des principes statistiques différents, ces deux méthodes permettent d'estimer la période de retour en fonction des données disponibles.

Estimation avec l'approche GPD

L'approche GPD se base sur l'idée que la distribution des excès au-delà d'un seuil u suit une loi GPD paramétrée par γ (pente) et σ (échelle). Pour un seuil $u = X_{n-k+1:n}$, le k-ème plus grand élément de l'échantillon, la période de retour T est estimée par la relation suivante :

$$T = \frac{n}{k+1} \cdot \frac{1}{1 - G_{\gamma,\sigma}(x_p)},$$

où $G_{\gamma,\sigma}(x_p)$ est la fonction de répartition de la GPD, et x_p est le quantile correspondant à la période de retour recherchée.

L'estimation de la période de retour peut aussi se faire graphiquement à l'aide des « quantile plots ». Ces graphiques montrent souvent une relation linéaire au-delà du seuil $X_{n-k+1:n}$, avec une pente γ, ce qui simplifie l'estimation. L'équation de cette droite est :

$$y = \log X_{n-k+1:n} + \gamma\left(x - \log\left(\frac{n}{k+1}\right)\right),$$

où $y = \log(X)$ et $x = \log\left(\frac{n}{k+1}\right)$ sont les coordonnées logarithmiques des points. À partir de cette droite, on obtient les relations suivantes pour T en fonction du débit X :

- Si $\gamma > 0$ (pente positive) :

$$\log(X) = \log X_{n-k+1:n} + \hat{\gamma}\left(\log(T) - \log\left(\frac{n}{k+1}\right)\right),$$

- Si $\gamma = 0$ (pente nulle, correspond à une distribution de type Gumbel) :

$$X = X_{n-k+1:n} + \hat{\sigma}\left(\log(T) - \log\left(\frac{n}{k+1}\right)\right).$$

Les paramètres $\hat{\gamma}$ et $\hat{\sigma}$ sont des estimateurs calculés à partir des données.

Exemple

Nous simulons $n = 1000$ des données qui représentent des débits journaliers sur une durée fictive de 1000 jours, simulés à partir d'une loi Gamma avec $\xi = 5$ et $\beta = 10$ une distribution fréquemment utilisée en hydrologie pour modéliser les précipitations ou les débits.

Le code R

```r
  # install.packages("evd")
library(evd)

# 1. Simulation de données (ex. : débits journaliers sur 1000 jours)
set.seed(123)
debits <- rgamma(1000, shape = 5, scale = 10)  # distribution gamma

# 2. Choix d'un seuil élevé (ex : 95e percentile)
u <- quantile(debits, 0.95)  # seuil

# 3. Ajustement du modèle GPD aux excès
fit_gpd <- fpot(debits, threshold = u)

# 4. Affichage des paramètres estimés
params <- fit_gpd$estimate
xi <- params["shape"]      # paramètre de forme ()
beta <- params["scale"]    # paramètre d'échelle ()
cat("Paramètres estimés :\nShape () =", round(xi, 3), "\nScale () =", round(beta, 3), "\n")

# 5. Proportion de dépassements
n <- length(debits)
nu <- sum(debits > u)
p_exc <- nu / n    # proportion d'excès

# 6. Estimation des quantiles de retour pour différentes périodes (en jours)
T_return <- c(2, 5, 10, 20, 50, 100, 200)  # périodes de retour en jours
q_T <- numeric(length(T_return))

for (i in seq_along(T_return)) {
  T <- T_return[i]
  if (xi != 0) {
    q_T[i] <- u + (beta / xi) * ((T * p_exc)^xi - 1)
  } else {
    q_T[i] <- u + beta * log(T * p_exc)
  }
}

# 7. Résultats dans un tableau
res <- data.frame(
  Periode_de_retour = T_return,
  Quantile = round(q_T, 2)
)
print(res)

# 8. Tracer la courbe
plot(T_return, q_T, type = "b", col = "blue", pch = 19, log = "x",
     xlab = "Période de retour (jours)",
     ylab = "Débit estimé (m³/s)",
     main = "Quantiles de retour estimés via modèle GPD")
```

```
50 grid()
```

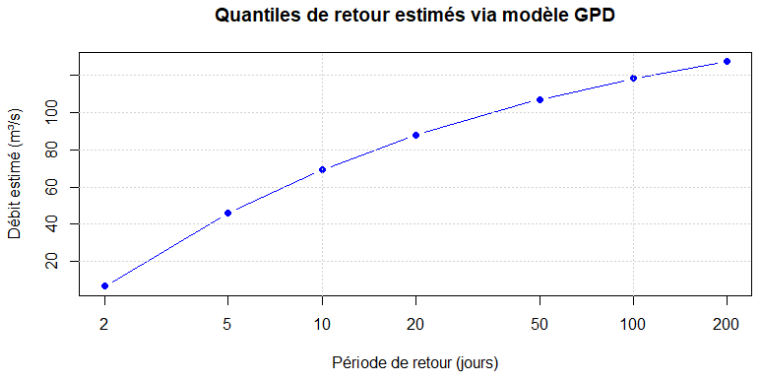


FIGURE 3.4 – Estimation de période de retour avec modèle GPD

	Periode	Quantile
1	2	6.69
2	5	46.13
3	10	69.27
4	20	87.88
5	50	107.02
6	100	118.25
7	200	127.28

La courbe des quantiles de retour (figure du code) montre une croissance asymptotique, caractéristique de la distribution GPD. Cette simulation met en évidence la pertinence de la loi GPD pour modéliser les valeurs extrêmes de débit. L'estimation des quantiles de retour permet d'évaluer la gravité potentielle des événements rares. Par exemple, un débit supérieur à 100 m^3/s n'est attendu qu'une fois tous les 50 à 100 jours.

Estimation avec l'approche GEV

L'approche GEV modélise les maxima des échantillons de données en utilisant la loi des valeurs extrêmes généralisées. La période de retour est directement exprimée par :

$$T = \frac{1}{1 - H_\gamma(x_p)},$$

où $H_\gamma(x_p)$ est la fonction de répartition de la loi GEV, et x_p est le quantile associé à la période de retour.
Une fois les paramètres de la distribution GEV estimés (notamment γ), on peut utiliser la formule précédente pour estimer T.

Exemple
Nous simulons un échantillon de $n = 50$ crues maximales annuelles selon une loi de Fréchet, caractérisée par un paramètre de forme $\gamma = 0.2$. On ajuste le modèle GEV aux données afin d'estimer les quantiles de retour associés à différentes périodes de retour (2 à 200 ans).

Le code R

```r
install.packages("evd")  # pour la GEV
library(evd)

# Exemple de données : maxima annuels simulés ou observés
# Ici, on simule des débits maximaux annuels
set.seed(123)
n <- 50
gamma_true <- 0.2
Q_max <- rfrechet(n, shape = 1 / gamma_true)

# Ajustement de la loi GEV aux maxima annuels
fit <- fgev(Q_max)

# Affichage des paramètres estimés : location (mu), scale (sigma), shape (xi)
print(fit$estimate)

# Estimation du quantile pour différentes périodes de retour
T_return <- c(2, 5, 10, 20, 50, 100, 200)  # périodes de retour en années
p <- 1 - 1 / T_return  # probabilités associées

# Quantiles de retour (inverse de la CDF de la GEV)
quantiles <- qgev(p, loc = fit$estimate["loc"],
                     scale = fit$estimate["scale"],
                     shape = fit$estimate["shape"])

# Affichage
data.frame(Periode = T_return, Quantile = round(quantiles, 2))

# Tracé du graphique quantile vs période de retour
plot(T_return, quantiles, type = "b", log = "x",
     xlab = "Période de retour (années)",
     ylab = "Débit estimé (m³/s)",
     main = "Courbe des quantiles de retour (modèle GEV)",
     col = "blue", pch = 19)


```

La courbe montre que les débits associés aux crues deviennent progressivement plus élevés à mesure que la période de retour augmente. Cela traduit la présence d'une queue lourde.
Le modèle GEV ajusté fournit une estimation des quantiles associés à différentes périodes de retour. Ces quantiles représentent les débits théoriques attendus avec une probabilité d'excès annuelle de $1/T$, où T est la période de retour en années.

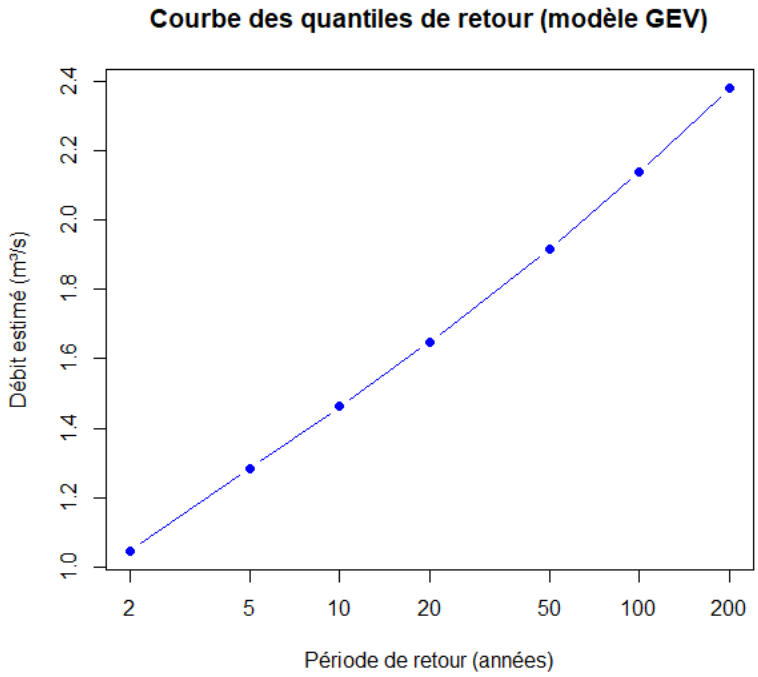


FIGURE 3.5 – Quantiles de retour estimés à partir du modèle GEV appliqué aux crues maximales

	Periode	Quantile
1	2	1.05
2	5	1.29
3	10	1.46
4	20	1.65
5	50	1.92
6	100	2.14
7	200	2.38

3.4 Traitement de données incomplètes

3.4.1 Notions de base pour l'analyse de la survie

L'analyse de la survie est une branche des statistiques qui s'intéresse à l'étude du temps qu'il faut pour qu'un événement particulier se produise (notamment un événement extrême), comme un décès, une crue. Plusieurs concepts de base sont utilisés pour comprendre et quantifier ces événements.

1. **Date d'origine** : Elle correspond à la date à laquelle a débuté l'observation. Pour un phénomène hydrologique, cela peut être la date de début d'une période de mesure des débits dans une rivière, la date à laquelle un enregistrement commence sur un bassin versant, ou encore la date à partir de laquelle un événement extrême est observé.

2. **Date des dernières nouvelles** : Est la dernière date à laquelle des informations pertinentes sur l'événement de crue ont été collectées. Par exemple, il pourrait s'agir de la dernière mesure de débit avant la fin d'une étude, de la dernière observation d'une crue ou de la date à laquelle l'étude prend fin.

3. **Date de point** : C'est la date au-delà de laquelle on arrêtera l'étude et on ne tiendra plus compte des informations sur les sujets. La date de point dans l'analyse des crues correspond à la date limite après laquelle les informations sur les événements de crue ne sont plus prises en compte. Par exemple, après cette date, il n'y a plus de nouvelles observations ou l'étude prend fin, et aucune nouvelle donnée n'est incluse dans l'analyse.

3.4.2 Fonction d'intérêt

Dans l'analyse hydrologique, des fonctions similaires à celles utilisées en analyse de la survie sont employées pour décrire les comportements des crues et leur probabilité d'occurrence (voir Klein et Moeschberger[23])

Définition 3.3 (Fonction de survie) *Dans ce contexte, la fonction de survie $S(x)$ représente la probabilité qu'une crue d'intensité supérieure à x ne se produise pas avant un certain temps ou une certaine période. Plus précisément, si x représente un niveau de crue, $S(x)$ serait la probabilité qu'un niveau de crue supérieur à x ne se produise pas au-delà d'un temps t.*

$$S(x) = \mathbb{P}(X > x) = 1 - F(x),$$

où $F(x)$ est la fonction de répartition qui représente la probabilité qu'un débit de crue dépasse une certaine valeur x.

La fonction $S(x)$ est utilisée pour estimer les chances de survie d'une rivière ou d'un bassin versant sans atteindre des niveaux critiques (crues importantes) pendant un intervalle de temps donné.

Définition 3.4 (Taux instantané de défaillance) *Le taux instantané de défaillance, ou taux de crue, représente la probabilité conditionnelle qu'une crue dépasse un certain niveau x dans un très petit intervalle, étant donné qu'une crue n'a pas encore dépassé ce seuil jusqu'à ce moment-là. Il est défini par :*

$$h(x) := \lim_{dx \to 0} \frac{\mathbb{P}(x < X < x + dx \mid X > x)}{dx}$$

Ce taux est essentiel pour comprendre l'intensité des crues sur une période donnée. Par exemple, un taux de crue élevé pour un niveau donné x indique que la probabilité d'atteindre ce seuil de crue augmente rapidement.
La relation entre le taux instantané de crue et la fonction de survie $S(x)$ est la suivante :

$$h(x) = \frac{f(x)}{S(x)} = -\frac{S'(x)}{S(x)} = -\frac{d}{dx}\log(S(x)).$$

Cette relation montre comment le taux instantané de crue est lié à la variation de la fonction de survie, et donc à la probabilité cumulative d'événements extrêmes.

Définition 3.5 (Taux de hasard cumulé) *Le taux de crue cumulé, noté $\Phi(x)$, est une mesure de l'accumulation du risque de crue au-delà d'un seuil donné. C'est l'intégrale du taux instantané de crue sur un intervalle de temps, et elle est définie par :*

$$\Phi(x) = \int_0^x h(u)du$$

La fonction $\Phi(x)$ permet de quantifier l'accumulation du risque de crue jusqu'à un certain niveau x. Elle est reliée à la fonction de survie $S(x)$ par la formule suivante :

$$\Phi(x) = -\log(S(x)).$$

Cette relation montre que le taux cumulé de crue est étroitement lié à la probabilité de survie d'un bassin versant ou d'une rivière sans dépasser un seuil de crue critique.

3.4.3 Données censurées

Dans l'analyse des crues et des données hydrologiques extrêmes, il est fréquent de rencontrer des données incomplètes. Cette incomplétude peut résulter de périodes sans mesure, de données censurées ou tronquées, notamment dans les séries historiques où seules les grandes crues sont enregistrées.

Définition 3.6 *La variable de censure C est définie par la non-observation de l'événement étudié. Si l'on observe C et non X, et que l'on sait que $X > C$ (respectivement $X < C$, $C_1 < X < C_2$), on parle de censure à droite (respectivement censure à gauche, censure par intervalle).*

Si l'événement se produit, X est "réalisé". S'il ne se produit pas, c'est C qui est "réalisé".

Trois notions principales sont utilisées pour le traitement statistique de ces cas :

- *La censure* :
  - Censure à droite : on sait qu'un événement a dépassé un certain seuil, sans connaître sa valeur exacte.
  - Censure à gauche : la valeur observée est inférieure à un seuil connu.
  - Censure par intervalle : la valeur réelle se situe dans un intervalle.

  En hydrologie, cela correspond à des situations où seules les crues supérieures à un certain niveau sont enregistrées, ou lorsque les extrêmes sont rapportés comme des intervalles.
- *La troncature* : les observations situées en dehors d'un intervalle donné sont absentes de l'échantillon. En pratique, cela peut se produire lorsque seules les années comportant une crue exceptionnelle sont retenues, les autres étant ignorées.
- *L'estimateur de Kaplan-Meier* : cet estimateur est utilisé pour estimer la fonction de survie en présence de données censurées. Il permet d'intégrer l'information partielle fournie par les observations incomplètes dans l'estimation des probabilités de dépassement.

Ces outils permettent de valoriser au mieux l'information disponible dans les séries hydrologiques, même incomplètes, afin d'estimer correctement les quantiles extrêmes et les périodes de retour associées.
Dans les études hydrologiques, les séries de crues ou de débits extrêmes sont fréquemment incomplètes, en particulier lorsqu'il s'agit de données anciennes ou issues de campagnes de mesure discontinues. Ce type de données peut inclure des observations censurées, c'est-à-dire des cas où on ne connaît pas la valeur exacte d'un débit extrême, mais seulement le fait qu'il dépasse (ou non) un certain seuil. Pour exploiter statistiquement ces données incomplètes, on utilise la fonction de survie, notée $S(x) = P(X > x)$, qui représente la probabilité qu'un débit dépasse une valeur donnée x.
Lorsque les observations sont censurées, l'estimation empirique classique de la fonction de survie est biaisée. L'estimateur de Kaplan-Meier permet alors d'obtenir une estimation non paramétrique de $S(x)$, tout en tenant compte des données censurées. Il est défini par :

Définition 3.7 *L'estimateur de Kaplan-Meier [21] de la fonction de survie S en présence de censure, pour $x < T_{n:n}$, $i = 1, \ldots, n$ est donné par la formule suivante :*

$$\widehat{S}(x) = \prod_{T_{i:n} \leq x} \left(1 - \frac{\delta_{[i,n]}}{n_i}\right) = \prod_{T_{i:n} \leq x} \left(1 - \frac{\delta_{[i,n]}}{n - i + 1}\right)$$

où $T_{1:n} \leq \cdots \leq T_{n:n}$ sont les statistiques d'ordre associées à $T_1, \ldots, T_n$, et $\delta_{[i,n]}$ est l'indicateur de censure associé à $T_{i:n}$.

Remarque
Il existe d'autres formes pour l'estimateur de Kaplan-Meier :

$$\widehat{S}(x) = \prod_{T_{i:n} \leq x} \left(\frac{n - i}{n - i + 1}\right)^{\delta_{[i,n]}} = \prod_{i=1}^{n} \left(1 - \frac{\delta_{[i,n]}}{n - i + 1}\right)^{\mathbb{I}_{\{T_{i:n} \leq x\}}}, \quad i = 1, \ldots, n.$$

L'estimateur de Kaplan-Meier présente plusieurs propriétés importantes : il est non paramétrique, cohérent, et converge uniformément vers la vraie fonction de survie lorsque la taille de l'échantillon augmente (voir Droesbeke et Saporta [12]). En hydrologie, il permet ainsi de tirer parti d'informations partielles dans l'estimation des quantiles extrêmes et des périodes de retour, en intégrant de façon rigoureuse les données censurées souvent présentes dans les séries historiques.

Exemple :
Ce script illustre l'utilisation de l'estimateur de Kaplan-Meier pour estimer la fonction de survie données de crues (valeurs en m^3/s) avec censure, 'temps' représente les hauteurs ou débits enregistrés.

Le code R

```r
install.packages("survival")
library(survival)

# 'censure' = 1 si valeur observée, 0 si censurée (on connaît juste un seuil dépassé)

temps <- c(120, 135, 140, 150, 160, 170, 145, 155, 175, 180)
censure <- c(1, 1, 0, 1, 0, 1, 1, 1, 0, 1)  # les 3e, 5e et 9e observations sont censurées

# Création de l'objet de type Surv pour données censurées
# Surv(time, event) : 'event = 1' si observé, 'event = 0' si censuré
obj_surv <- Surv(temps, censure)

# Estimation de la fonction de survie par l'estimateur de Kaplan-Meier
km_fit <- survfit(obj_surv ~ 1)

# Affichage des résultats
summary(km_fit)

# Tracé de la fonction de survie estimée
plot(km_fit, conf.int = TRUE,
     main = "Estimateur de Kaplan-Meier pour des crues censurées",
     xlab = "Débit (m³/s)",
     ylab = "Probabilité de dépassement S(x) = P(X > x)",
```

```
24        col = "blue", lwd = 2)
25  grid()
26
```

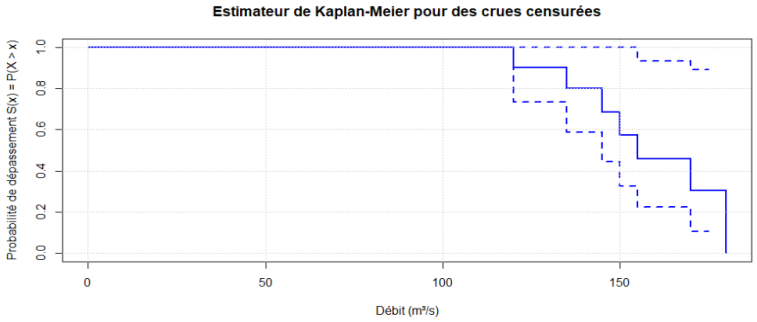


FIGURE 3.6 – Estimation de la fonction de survie par l'estimateur de Kaplan-Meier

Débit (m^3/s)	n.risk	n.event	$\hat{S}(x)$	Erreur standard	IC 95% Inf.	IC 95% Sup.
120	10	1	0.900	0.0949	0.732	1.000
135	9	1	0.800	0.1265	0.587	1.000
145	7	1	0.686	0.1515	0.445	1.000
150	6	1	0.571	0.1638	0.326	1.000
155	5	1	0.457	0.1662	0.224	0.932
170	3	1	0.305	0.1666	0.104	0.890
180	1	1	0.000	—	—	—

TABLE 3.1 – Estimation de la fonction de survie par l'estimateur de Kaplan-Meier

avec,

- **Débit (m^3/s)** : valeur du débit observé ou censuré.
- **n.risk** : nombre d'observations encore "à risque", c'est-à-dire non encore observées ni censurées, juste avant cette valeur.
- **n.event** : nombre de crues effectivement observées à cette valeur (événements non censurés).
- $\hat{S}(x)$: estimation de la probabilité que le débit dépasse x (fonction de survie).
- **Erreur standard** : erreur associée à l'estimateur $\hat{S}(x)$.
- **IC 95 %** : intervalle de confiance à 95 % pour la fonction de survie.

Ce tableau présente les résultats de l'estimation de la fonction de survie $\hat{S}(x)$, qui représente la probabilité qu'une crue dépasse un certain débit x. Par exemple, on estime que la probabilité qu'une crue dépasse 120 m^3/s est de 90 %, celle de dépasser 150 m^3/s est de 57 %, et celle de dépasser 170 m^3/s est de 30 %. Enfin, la fonction de survie atteint 0 à 180 m^3/s, ce qui signifie qu'aucune crue plus intense n'a été observée dans l'échantillon.

- Les intervalles de confiance deviennent plus larges à mesure que le nombre d'observations diminue, notamment dans les queues de distribution.
- À 180 m^3/s, la fonction de survie est nulle, car il s'agit de la crue maximale observée. Par conséquent, lerreur standard et les bornes de l'intervalle de confiance ne sont pas définies.

— L'estimateur de Kaplan-Meier est particulièrement utile dans ce contexte, car il permet d'incorporer rigoureusement l'information partielle contenue dans les données censurées pour estimer la distribution des crues.

3.5 Application sur des données réelles

On utilise un jeu de données s'intitule `V312401001_QIXnJ(n=1_non-glissant)(1).csv`. Il contient les débits maximaux annuels extraits du site (www.hydro.eaufrance.fr), observés à une station hydrométrique "Le Gier à Givors"(code station :V312 4010 01 et le libellé : Le Gier à Givors — station au confluent du Gier et du Rhône) sur une période donnée(du 1er octobre 2024 au 1er décembre2024). Ces données résultent d'un traitement non-glissant, ce qui signifie que seul le débit maximal par année civile est retenu.
Le fichier comporte les colonnes suivantes :

— **Date..TU.** : la date associée au débit maximal enregistré pour une année donnée ;

— **Date.de.la.mesure.du.min.max..TU.** : date précise de la mesure du maximum ou minimum ;

— **Valeur..en.m..s.** : valeur du débit maximal annuel, en mètres cubes par seconde (m^3/s), qui constitue la variable principale analysée ;

— **Statut**, **Qualification**, **Méthode**, **Continuité** : variables complémentaires sur la qualité des mesures, leur origine ou leur validité.

Le fichier contient un total de 62 jours de données de débit journalier maximal, après nettoyage et suppression des valeurs manquantes. La variable principale `Valeur..en.m..s.` a été renommée `Debit` dans le cadre du traitement sous R. Toutes les observations ont été converties au format numérique et les éventuelles valeurs manquantes ont été éliminées.

L'objectif de cette partie est de modéliser statistiquement les crues maximales annuelles à l'aide de la loi des valeurs extrêmes généralisée. Ce type de loi est particulièrement adapté pour représenter des phénomènes rares et extrêmes comme les inondations.

```r
# Installer les packages
install.packages("tidyverse")
install.packages("evd")

# Charger les packages
library(tidyverse)
library(evd)

# Lire le fichier
data <- read.csv("V312401001_QIXnJ(n=1_non-glissant) (1).csv", stringsAsFactors = FALSE)

# Extraire les débits maximaux annuels
debit_max <- data$Valeur..en.m..s.
debit_max <- na.omit(debit_max)

# Analyse graphique Pareto
Q_sorted <- sort(debit_max, decreasing = TRUE)
n <- length(Q_sorted)
```

```r
log_Q <- log(Q_sorted)
log_rank <- log((n + 1) / (1:n))

plot(log_rank, log_Q, type = "p", col = "black",
     xlab = "log((n+1)/j)", ylab = "log(Q_{(j)})",
     main = "Pareto Quantile Plot - Crues maximales annuelles")
abline(lm(log_Q ~ log_rank), col = "blue", lwd = 2)
# Ajustement GEV
fit <- fgev(debit_max)
summary(fit)

# Valeurs de retour
T <- c(10, 50, 100)
p <- 1 - 1/T
mu <- fit$estimate["loc"]
sigma <- fit$estimate["scale"]
xi <- fit$estimate["shape"]

q_return <- evd::qgev(p, loc = mu, scale = sigma, shape = xi)

data.frame(Période = T, Débit_Q = round(q_return, 2))
```

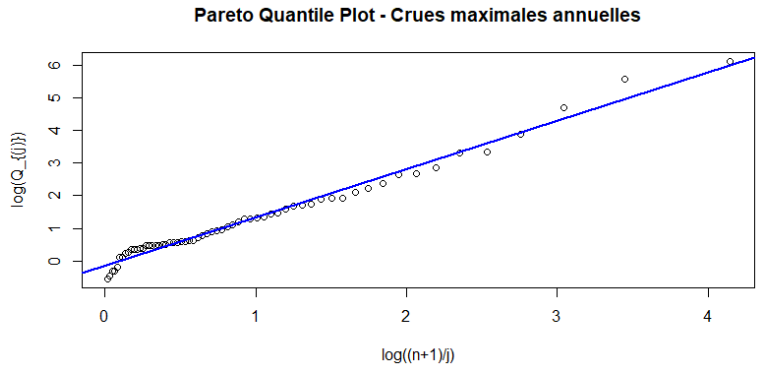


FIGURE 3.7 – Graphique quantile de Pareto en échelle log-log pour les crues maximales annuelles

Periode	Débit_Q
10	18.36
50	105.62
100	222.23

Le graphique ci-dessus montre que la distribution des crues maximales annuelles suit une tendance quasi-linéaire en échelle log-log, indiquant un comportement en queue lourde. Ce résultat appuie le choix d'un modèle GEV avec un paramètre de forme positif (loi de Fréchet) pour modéliser les quantiles de retour extrêmes.
Une fois la loi GEV ajustée avec fgev(), on peut vérifier visuellement si l'ajustement est bon ou non grâce au graphique :

```r
# QQ-plot
qqplot(qgev(ppoints(length(debit_max)), loc = fit$estimate["loc"],
            scale = fit$estimate["scale"],
            shape = fit$estimate["shape"]),
       sort(debit_max),
       xlab = "Quantiles théoriques (GEV)",
       ylab = "Quantiles observés",
       main = "QQ-plot : Ajustement GEV")
abline(0, 1, col = "blue", lwd = 2)
```

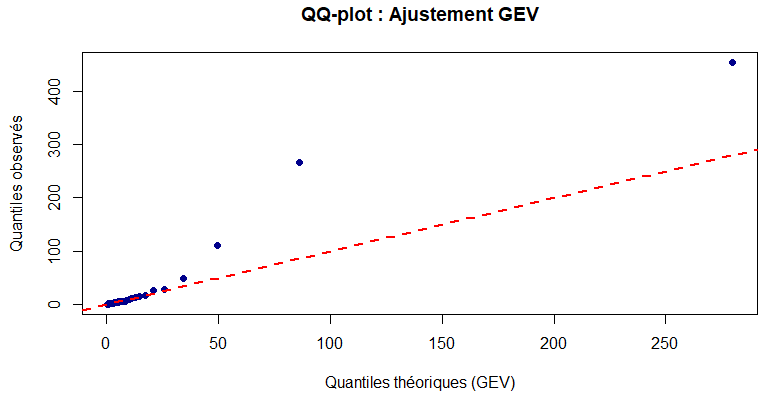


FIGURE 3.8 – QQ-plot des quantiles observés vs. théoriques — Ajustement GEV

Le QQ-plot montre que le modèle GEV ajusté reproduit bien les quantiles empiriques, la majorité des points se situant à proximité de la diagonale. Cela confirme la qualité de l'ajustement.

Estimation du quantile de retour avec modèle GPD

```r
#1
install.packages("evd")
library(evd)

# 2. Charger les données
data <- read.csv("V312401001_QIXnJ(n=1_non-glissant) (1).csv", stringsAsFactors = FALSE)

# 3. Vérifier les colonnes
names(data)

# 4. Extraire le vecteur de débits journaliers ou maxima selon la colonne
debits <- data$Valeur..en.m..s.
debits <- na.omit(debits)  # Supprimer les valeurs manquantes

# 5. Choisir un seuil (exemple : 90e percentile)
u <- quantile(debits, 0.90)

# 6. Extraire les excès par rapport au seuil
exces <- debits[debits > u] - u
```

```r

# 7. Ajuster la loi GPD sur les excès
fit_gpd <- fpot(exces, threshold = 0)  # On centre déjà les excès à 0
summary(fit_gpd)

# 8. Estimer les quantiles de retour
n <- length(debits)
nu <- length(exces)
pu <- nu / n

# Récupérer les paramètres estimés
xi <- fit_gpd$estimate["shape"]
sigma <- fit_gpd$estimate["scale"]

# Définir les périodes de retour
T_return <- c(2, 5, 10, 20, 50, 100)

# Calcul des quantiles de retour
quantiles <- u + (sigma / xi) * ((T_return * pu)^xi - 1)

# Affichage des résultats
resultats <- data.frame(Période = T_return, Quantile_m3s = round(quantiles, 2))
print(resultats)

# 9. La courbe des quantiles de retour
plot(T_return, quantiles, type = "b", log = "x",
     xlab = "Période de retour (années)",
     ylab = "Débit estimé (m³/s)",
     main = "Courbe de quantiles (loi GPD - Poste V312401001)",
     col = "darkgreen", pch = 19)
grid()
```

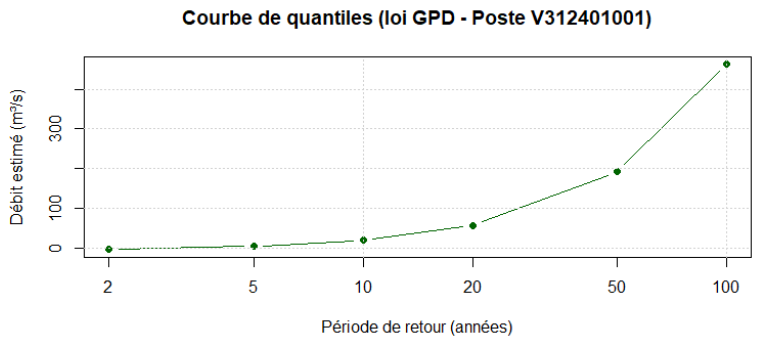


FIGURE 3.9 – Courbe des quantiles de retour estimés selon la loi GPD

Période	quantile_m3s
2	-2.66
5	5.25
10	21.02
20	58.06
50	192.94
100	461.77

Le graphique présente la courbe des quantiles de retour estimés à partir de la loi GPD, ajustée aux excès au-delà du 90% (nous avons choisis ce seuil car on a un petit échantillon). Cette courbe permet d'associer à chaque période de retour une valeur estimée du débit extrême. On observe une croissance rapide du débit estimé avec la période de retour, ce qui traduit l'augmentation de la gravité des crues rares. Ces résultats sont essentiels pour les études hydrologiques liées à la sécurité des ouvrages (barrages, digues) et à la gestion des risques d'inondation.

Les résultats montrent aussi une croissance rapide des débits estimés avec l'augmentation de la période de retour, ce qui est attendu dans les processus extrêmes.
Le débit estimé pour une période de retour de 100 ans atteint environ 462 m^3/s, ce qui reflète un événement très rare et potentiellement dangereux.
Estimation du quantile de retour avec modèle GEV

```r
#install.packages("evd")
library(evd)

# Lire les données depuis le fichier CSV
data <- read.csv("V312401001_QIXnJ(n=1_non-glissant) (1).csv", stringsAsFactors = FALSE)

# Vérifier les noms de colonnes (optionnel mais utile)
names(data)

# Extraire la colonne des débits maximaux annuels
Q_max <- data$Valeur..en.m..s.
Q_max <- na.omit(Q_max)  # Supprimer les NA si présents

# Ajuster une loi GEV
fit <- fgev(Q_max)

# Afficher les paramètres estimés : loc, scale, shape
print(fit$estimate)

# Définir les périodes de retour souhaitées (en années)
T_return <- c(2, 5, 10, 20, 50, 100, 200)
p <- 1 - 1 / T_return

# Calculer les quantiles de retour
quantiles <- evd::qgev(p,
                       loc = fit$estimate["loc"],
                       scale = fit$estimate["scale"],
                       shape = fit$estimate["shape"])
```

```r

# Afficher les résultats
resultats <- data.frame(Période_de_retour = T_return,
                        Débit_estimé_m3s = round(quantiles, 2))
print(resultats)

# Tracer la courbe des quantiles
plot(T_return, quantiles, type = "b", log = "x",
     xlab = "Période de retour (années)",
     ylab = "Débit estimé (m³/s)",
     main = "Courbe de quantiles de retour - Poste V312401001",
     col = "darkblue", pch = 19)
grid()
```

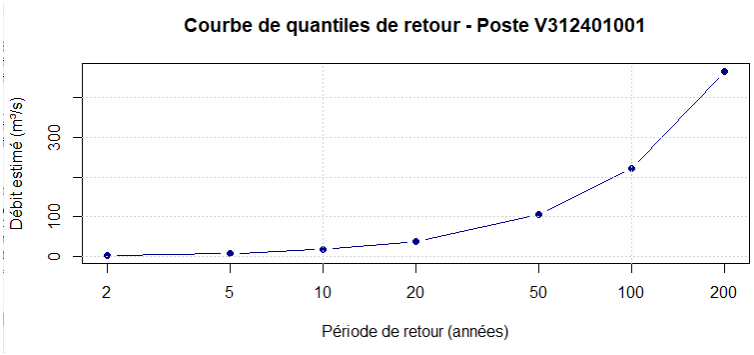


FIGURE 3.10 – Courbe des quantiles de retour estimés selon la loi GEV

Période	quantile_m3s
2	2.71
5	8.40
10	18.36
20	39.25
50	105.62
100	222.23
200	466.63

Le graphique présente la courbe des quantiles de retour estimés à partir de la loi GEV ajusté aux débits maximaux annuels observés. La courbe nous montre que le débit estimé augmente rapidement avec la période de retour. Cela traduit un comportement de queue lourde, ce qui est typique des phénomènes extrêmes modélisés avec une loi de Fréchet. Le modèle GEV permet d'estimer des débits extrêmes croissants avec la période de retour.

Ces résultats représentent les valeurs de débit susceptibles d'être atteintes ou dépassées avec une probabilité annuelle donnée.Par exemple, un débit estimé à 222.23 m^3/s pour une période de retour de 100 ans signifie qu'un tel débit a 1% de probabilité d'être dépassé au cours d'une année donnée. La progression rapide des débits avec l'augmentation de la période de retour reflète la nature extrême des crues rares (Le graphique des quantiles en fonction de la période de retour (en échelle logarithmique) confirme la tendance croissante des valeurs extrêmes modélisées par la loi GEV).

Comparaison des deux approches

L'ajustement GEV est plus stable et réaliste pour les périodes courtes à moyennes, car il considère l'ensemble des maxima annuels. L'ajustement GPD est plus sensible mais plus conservateur pour les crues rares (périodes longues), ce qui peut être utile dans une perspective de gestion des risques extrêmes. Dans la pratique, les deux méthodes sont complémentaires.

Estimation de la période de retour selon plusieurs estimateur de γ

```r

install.packages("evir")
library(evir)

# 2. Lire les données

data <- read.csv("V312401001_QIXnJ(n=1_non-glissant) (1).csv", stringsAsFactors = FALSE)

# Vérifier les noms des colonnes
names(data)

# Adapter ici le nom de la colonne de débits
debits <- na.omit(data$Valeur..en.m..s.)
n <- length(debits)

# Trier les données en ordre décroissant
debits_sorted <- sort(debits, decreasing = TRUE)

# Choix automatique d'un k raisonnable
k <- min(50, floor(n / 5))

# 3. Estimateur de Hill
hill_obj <- hill(debits_sorted)
gamma_hill <- hill_obj$y[k]

# 4. Estimateur de Pickands
pickands_estimator <- function(data, k) {
  x_sorted <- sort(data, decreasing = TRUE)
  if (4 * k > length(x_sorted)) stop("Pas assez de données pour Pickands")
  X_k  <- x_sorted[k]
  X_2k <- x_sorted[2 * k]
  X_4k <- x_sorted[4 * k]
  gamma <- (1 / log(2)) * log((X_k - X_2k) / (X_2k - X_4k))
  return(gamma)
}
gamma_pick <- pickands_estimator(debits, k)

# 5. Estimateur des Moments
moment_estimator <- function(data, k) {
  x_sorted <- sort(data, decreasing = TRUE)
  u <- x_sorted[k]
  excesses <- x_sorted[1:k] - u
  mean_excess <- mean(excesses)
```

```r
  var_excess <- var(excesses)
  gamma <- mean_excess / (mean_excess - var_excess)
  return(gamma)
}
gamma_moment <- moment_estimator(debits, k)

# 6. Quantiles de retour avec les estimateurs
# Seuil u (ex. 95e percentile)
u <- quantile(debits, 0.9)
exces <- debits[debits > u] - u
nu <- length(exces)
pu <- nu / n

# Si trop peu d'excès, baisser le seuil
if (nu < 10) {
  u <- quantile(debits, 0.90)
  exces <- debits[debits > u] - u
  nu <- length(exces)
  pu <- nu / n
}
sigma_hat <- sd(exces)

# Périodes de retour
T_vals <- c(2, 5, 10, 20, 50, 100)

# Fonction de calcul des quantiles
quantile_retour <- function(gamma, sigma, T, u, pu) {
  if (abs(gamma) < 1e-5) {
    return(u + sigma * log(T * pu))  # approximation pour gamma  0
  } else {
    return(u + (sigma / gamma) * ((T * pu)^gamma - 1))
  }
}
# Calculs
QT_hill   <- quantile_retour(gamma_hill, sigma_hat, T_vals, u, pu)
QT_pick   <- quantile_retour(gamma_pick, sigma_hat, T_vals, u, pu)
QT_moment <- quantile_retour(gamma_moment, sigma_hat, T_vals, u, pu)

# Résultats comparés
quantiles_comparatif <- data.frame(
  `Période (ans)` = T_vals,
  Hill = round(QT_hill, 2),
  Pickands = round(QT_pick, 2),
  Moments = round(QT_moment, 2)
)

print(quantiles_comparatif)
```

Période..ans.	Hill	Pickands	Moments
2	-123.43	-102.25	-229.31
5	-57.87	-52.49	-77.34
10	38.31	38.68	37.25
20	209.97	237.18	151.52
50	659.21	896.77	302.10
100	1318.18	2105.26	415.63

Ce tableau donne les valeurs de débit estimé (en m^3/s) pour différentes périodes de retour (en années), obtenues à partir d'un modèle de type Peaks Over Threshold (POT) avec une loi de Pareto généralisée (GPD).
Les valeurs pour les périodes courtes (T=2 et T=5) ne sont pas interprétables ni fiables.
À partir de 10 ans, les valeurs sont positives, réalistes et croissantes. Les trois méthodes donnent des valeurs cohérentes en tendance. Les trois méthodes confirment une forte croissance des quantiles avec la période de retour, compatible avec un comportement de type Fréchet (queue lourde) et les différences de valeurs indiquent la sensibilité des estimateurs à la sélection du seuil ou de k.
Les estimations finales des quantiles de retour ont été basées principalement sur l'estimateur des Moments, réputé plus robuste en présence de données bruitées. L'estimateur de Pickands a été utilisé pour comparaison, tandis que l'estimateur de Hill a été écarté pour les faibles périodes (T = 2, 5 ans) en raison d'une extrapolation négative non interprétable.

```

plot(T_vals, QT_hill, type = "b", pch = 19, col = "blue", log = "x",
     xlab = "Période de retour (années)", ylab = "Débit estimé (m³/s)",
     main = "Quantiles de retour - Comparaison des estimateurs",
     ylim = range(c(QT_hill, QT_pick, QT_moment), na.rm = TRUE))

lines(T_vals, QT_pick, type = "b", pch = 17, col = "red")
lines(T_vals, QT_moment, type = "b", pch = 15, col = "darkgreen")

legend("topleft", legend = c("Hill", "Pickands", "Moments"),
       col = c("blue", "red", "darkgreen"),
       pch = c(19, 17, 15), lty = 1)

grid()
```

L'estimation de l'indice des extrêmes à l'aide de plusieurs méthodes permet une analyse robuste du risque de crue. Les quantiles obtenus peuvent être utilisés pour dimensionner des ouvrages de protection ou pour définir des scénarios de crues extrêmes dans les documents de gestion du risque.

Estimation de la Période de Retour en présence de censure

```

install.packages("survival")
library(survival)
library(dplyr)
```

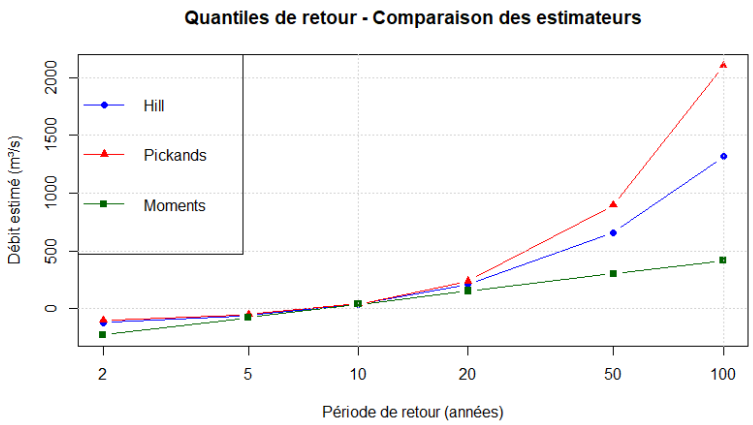


FIGURE 3.11 – Comparaison graphique des estimateurs de quantiles de retour

```r
library(tidyr)

# 2. Lire les données
data <- read.csv("V312401001_QIXnJ(n=1_non-glissant) (1).csv", stringsAsFactors = FALSE)

# 3. Préparation des données
data <- data %>%
  rename(Date = Date..TU., Debit = Valeur..en.m..s.) %>%
  mutate(Debit = as.numeric(Debit)) %>%
  drop_na()

# 4. Créer les variables pour Kaplan-Meier
temps <- data$Debit

censure <- rep(1, length(temps))  # 1 = observé ; 0 = censuré

# 5. Créer un objet Surv pour Kaplan-Meier
obj_surv <- Surv(temps, censure)

# 6. Estimer la fonction de survie
km_fit <- survfit(obj_surv ~ 1)

# 7. Résumé des résultats
summary(km_fit)

# 8. Tracer la courbe de survie
plot(km_fit, conf.int = TRUE,
     main = "Estimateur de Kaplan-Meier - Données de crues",
     xlab = "Débit (m³/s)",
     ylab = "S(x) = P(X > x)",  # fonction de survie
     col = "blue", lwd = 2)
grid()
```

Ce graphique présente la fonction de survie estimé $S(x) = P(X > x)$ des débits journaliers observés, à l'aide de l'estimateur de Kaplan-Meier.

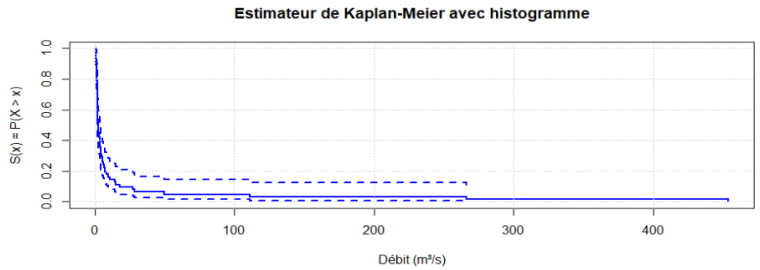


FIGURE 3.12 – Fonction de survie estimée par l'estimateur de Kaplan-Meier

On constate une forte décroissance de la courbe pour les faibles valeurs de débit (inférieures à 50 m^3/s), ce qui indique que la majorité des observations correspondent à des débits faibles.
La fonction de survie devient ensuite plus stable entre 50 et 200 m^3/s, ce qui reflète la rareté des crues modérées. Pour les débits supérieurs à 200 m^3/s, la courbe tend vers zéro, ce qui suggère que les crues extrêmes sont très peu fréquentes.
Cette représentation met en évidence une distribution très asymétrique des débits, caractérisée par une forte densité de petites valeurs et une queue droite lourde. Elle confirme l'intérêt de recourir à des méthodes adaptées à l'analyse des extrêmes telles que la théorie des valeurs extrêmes (GEV, GPD) ou l'analyse de survie.
Dans notre cas nos fichier de données ne contient pas de données incomplétes. Nous allons introduire artificiellement de la censure le fichier de données de débits, pour faire une estimation Kaplan-Meier avec données partiellement censurées. donc on va dire que tous *les dbits* $< 20m/s$ sont censurés, les autres sont observés. Pour cela on rajoute cette étape juste après la lecture et nettoyage des données :

```r
# Débits (temps)
temps <- data$Debit

# Statut de censure : 0 = censuré (valeurs faibles), 1 = observé
censure <- ifelse(temps < 20, 0, 1)

# Ajout dans les données
data$censure <- censure
```

Aprés avoir appliquer l'estimateur de Kaplan-Meier avec données censurées, on a :

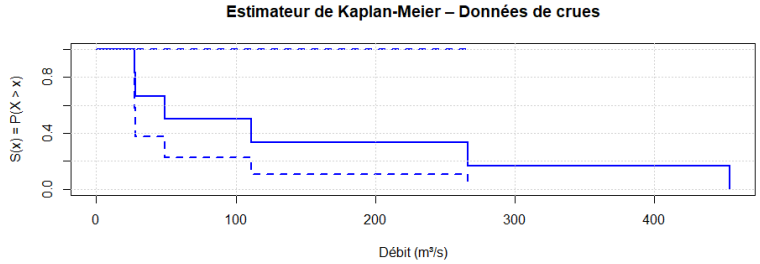


FIGURE 3.13 – Fonction de survie estimée par l'estimateur de Kaplan-Meier avec données censurées

La courbe montre une décroissance ce qui montre que les fortes valeurs de débit sont rares. On observe une décroissance rapide de la fonction de survie pour les faibles valeurs de débit (inférieures à environ 40 m^3/s), ce qui indique que la majorité des observations concernent des écoulements faibles à modérés. Entre 40 et 100 m^3/s, la courbe présente une pente plus douce, traduisant une fréquence plus faible mais non négligeable des crues modérées. Au-delà de 100 m^3/s, la courbe se stabilise avec des valeurs proches de zéro, ce qui suggère que les crues fortes ou extrêmes sont rares dans l'échantillon analysé. La présence de données censurées se traduit par des paliers horizontaux dans la courbe, qui marquent des incertitudes sur certains débits élevés, tout en permettant une estimation prudente de la probabilité de dépassement.

Conclusion

Ce mémoire a porté sur l'application de la théorie des valeurs extrêmes à l'analyse des phénomènes rares en hydrologie. Crues, débits extrêmes ou précipitations intenses, bien que peu fréquents, peuvent avoir des impacts majeurs sur les territoires et nécessitent une modélisation statistique adaptée.
Nous avons présenté les fondements de la TVE, en exposant les trois lois limites associées aux maxima (Fréchet, Gumbel, Weibull) et les deux approches principales : la loi généralisée des valeurs extrêmes pour les maxima de blocs, et la loi généralisée de Pareto dans le cadre de la méthode Peaks Over Threshold. Le choix du seuil pour la GPD s'est révélé déterminant, tout comme l'ajustement des modèles aux queues de distribution observées.
L'application à des données hydrologiques réelles, à partir d'une station de mesure, a permis d'estimer des niveaux de retour pour différentes périodes, utiles pour la prévention des inondations ou le dimensionnement d'ouvrages. La comparaison entre les modèles GEV et GPD a montré leur complémentarité selon le format des données disponibles.
Enfin, la question des données incomplètes a été abordée à travers l'utilisation de l'estimateur de Kaplan-Meier, qui permet une estimation fiable de la fonction de survie malgré la censure.
Ainsi, la TVE s'impose comme un cadre rigoureux et un outil pratique pour analyser les événements extrêmes et appuyer la gestion des risques en hydrologie.

Bibliographie

[1] Beirlant, J., Goegebeur, Y., Teugels, J., Segers, J., De Waal, D., Ferro, C. (2004). John Wiley Sons Ltd. Chichester, UK.

[2] Bingham, N. H., Goldie, C. M., Teugels, J. L. (1989). Regular variation (Vol. 27). Cambridge university press.

[3] Cairns, J. A. (1987). Evaluating changes in league structure : the reorganization of the Scottish Football League. Applied Economics, 19(2), 259-275.

[4] Claps, P., Laio, F. (2003). Can continuous streamflow data support flood frequency analysis. An alternative to the partial duration series approach. Water Resources Research, 39(8).

[5] Coles, S., Bawa, J., Trenner, L., Dorazio, P. (2001). An introduction to statistical modeling of extreme values (Vol. 208, p. 208). London : Springer.

[6] Csörgoő, S., Mason, D. M. (1985, November). Central limit theorems for sums of extreme values. In Mathematical Proceedings of the Cambridge Philosophical Society (Vol. 98, No. 3, pp. 547-558). Cambridge University Press.

[7] 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.

[8] Dekkers, A. L., Einmahl, J. H., De Haan, L. (1989). A moment estimator for the index of an extreme-value distribution. The Annals of Statistics, 1833-1855.

[9] 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.

[10] Deheuvels, P., Haeusler, E., Mason, D. M. (1988, September). Almost sure convergence of the Hill estimator. In Mathematical Proceedings of the Cambridge Philosophical Society (Vol. 104, No. 2, pp. 371-381). Cambridge University Press.

[11] Deme, E. H. (2013). Quelques contributions à la Théorie univariée des Valeurs Extrêmes et Estimation des mesures de risque actuariel pour des pertes à queues lourdes. Université Gaston Berger.

[12] Droesbeke, J. J., Saporta, G. (2011). Approches non paramétriques en régression. Editions Technip.

[13] Embrechts, P., Klüppelberg, C., Mikosch, T., Emberchts, P., Klüppelberg, C., Mikosch, T. (1997). Risk theory. Modelling Extremal Events : for Insurance and Finance, 21-57.

[14] Fisher, R. A., Tippett, L. H. C. (1928, April). Limiting forms of the frequency distribution of the largest or smallest member of a sample. In Mathematical proceedings of the Cambridge philosophical society (Vol. 24, No. 2, pp. 180-190). Cambridge University Press.

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

[16] Groeneboom, P., Lopuhaä, H. P., De Wolf, P. P. (2003). Kernel-type estimators for the extreme value index. The Annals of Statistics, 31(6), 1956-1995.

[17] Haan, L., Ferreira, A. (2006). Extreme value theory : an introduction (Vol. 3). New York : springer.

[18] Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. The annals of statistics, 1163-1174.

[19] Hosking, J. R., Wallis, J. R. (1987). Parameter and quantile estimation for the generalized Pareto distribution. Technometrics, 29(3), 339-349.

[20] 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.

[21] Kaplan, E. L., Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American statistical association, 53(282), 457-481.

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

[23] Klein, J.P., Moeschberger M.L. (1998). Survival Analysis - Techniques for Censored and Truncated Data. Statistics for Biology and Health. Springer.

[24] Lang, M., Ouarda, T.B.M.J., Bobée, B. (1999). Towards operational guidelines for over-threshold modeling. J. Hydrol. 225, 103–117.

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

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

[27] Reiss, R. D., Thomas, M. (2007). Statistical analysis of extreme values : with applications to insurance, finance, hydrology and other fields. Basel : Birkhäuser Basel.

[28] Resnick, S. I. (2008). Extreme values, regular variation, and point processes (Vol. 4). Springer Science Business Media.

[29] Schultze, J., Steinebach, J. (1996). On least squares estimates of an exponential tail coefficient. Statistics Risk Modeling, 14(4), 353-372.

[30] USWRC. (1976). Guidelines for determining flood flow frequency. United States Water Resources Council, Bull. vol. 17, Hydrol. Comm. Washington, DC, 73p.

[31] Von Mises, R. (1936). Wahrscheinlichkeit Statistik und Wahrheit : Einführung in die neue Wahrscheinlichkeitslehre und ihre Anwendung.