République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

*Université Mohamed Khider, Biskra*

Faculté des Sciences Exactes

Département de Mathématiques


Mémoire présenté pour obtenir le diplôme de

Master en "**Mathématiques Appliquées**"

Option : **Statistique**

Par : BOUTA Oumeima

**Titre :**

# Modélisation des Valeurs Extrêmes

**Devant le Jury :**

| | | | | |
|---|---|---|---|---|
| Mr. | MERAGHNI Djamel | Pr | U. Biskra | Président |
| Mme. | BENAMEUR Sana | M.C.B | U. Biskra | Rapporteur |
| Mme. | SOLTANE Louiza | M.C.B | U. Biskra | Examinatrice |

**Soutenu Publiquement le 02/06/2025**

# *Dédicace*

Je dédie ce modeste travail

À mon grand amour, ma mère qui a sacrifié sa vie pour notre bonheur et notre réussite

À mon père à qui je témoigne de l'affection et du respect.

À ma défunte sœur : Djanatt el frdwousse

À mes frères

À toute la famille Bouta

À tous mes amis : Souhila, Selsabil, Imen.

# Remerciements

Je rends tout d'abord grâce à Allah, le Tout-Puissant et le Miséricordieux, qui m'a accordé la force, la patience et la persévérance nécessaires pour mener à bien ce travail.

Au début de ce mémoire, je souhaite exprimer ma profonde gratitude à toutes les personnes qui, de près ou de loin, ont contribué à la réalisation de ce travail.

Je tiens à remercier tout particulièrement mon encadreur, Mme **Benameur Sana**, pour son accompagnement bienveillant, ses conseils éclairés et son soutien indéfectible tout au long de cette recherche. Son encadrement rigoureux et sa disponibilité ont grandement contribué à l'avancement et à la qualité de ce travail.

Je remercie également les membres du jury, **Mr. Meraghni Djamel** et **Mme. Soltane Louiza** pour le temps qu'ils ont consacré à l'évaluation de ce mémoire, ainsi que pour leurs remarques pertinentes et constructives, qui ont permis d'enrichir ce travail.

Je n'oublie pas de remercier profondément ma famille, mes parents, pour leur soutien et leurs encouragements inébranlables tout au long de mes études. Ils ont toujours été mon pilier et ma source de force.

Enfin, je n'oublie pas mes collègues et amis, qui m'ont soutenue moralement et encouragée dans les moments difficiles. Merci à vous tous pour votre aide précieuse et votre bienveillance.

BOUTA Oumeima

# Table des matières

**Dédicace** i

**Remerciements** ii

**Table des matières** iii

**Table des figures** vi

**Liste des tableaux** vii

**Introduction** 1

**1 Distributions GEV** 3

1.1 Définitions et notions de base . . . . . 3

1.1.1 Statistique d'ordre . . . . . 3

1.1.2 Distributions d'une statistique d'ordre . . . . . 4

1.1.3 Loi jointe d'un couple de statistique d'ordre . . . . . 6

1.2 Approche des maxima par blocs . . . . . 9

1.3 Convergence du maximum renormalisé . . . . . 9

1.4 Distributions des valeurs extrêmes généralisée . . . . . 11

1.5 Fonctions à variation régulière . . . . . 13

1.6 Domaines d'attraction . . . . . 14

1.6.1 Domaine d'attraction de Fréchet . . . . . 15
1.6.2 Domaine d'attraction de Weibull . . . . . 16
1.6.3 Domaine d'attraction de Gumbel . . . . . 17
1.7 Estimation des paramètres . . . . . 18
1.7.1 Estimation paramétrique . . . . . 18
1.7.2 Estimation semi-paramétrique . . . . . 23
1.8 Estimation des quantiles extrêmes . . . . . 25

**2 Distrubutions GPD** **28**
2.1 Approche POT . . . . . 28
2.1.1 Modélisation des excès . . . . . 29
2.1.2 Distribution des excès . . . . . 30
2.2 Distribution de Pareto Généralisée . . . . . 31
2.3 Sélection du seuil . . . . . 34
2.4 Estimation des paramètres de la GPD . . . . . 36
2.4.1 Méthode du maximum de vraisemblance . . . . . 36
2.4.2 Méthode des moments de probabilités pondérés . . . . . 38
2.5 Estimation de la queue de la distribution . . . . . 38
2.6 Estimation des quantiles extrêmes . . . . . 39

**3 Application sous R** **40**
3.1 Description des données . . . . . 40
3.1.1 Ajustement par une loi normale . . . . . 42
3.2 Modélisation via une distribution *GEV* . . . . . 43
3.2.1 Série des maxima annuels . . . . . 43
3.2.2 Estimation des paramètres . . . . . 44

3.2.3 Validation du modèle ajusté . . . . . 45
3.2.4 Estimation des quantiles extrêmes . . . . . 47
3.3 Modélisation via une $GPD$ . . . . . 47
3.3.1 Sélection du seuil . . . . . 47
3.3.2 Série des excès . . . . . 49
3.3.3 Estimation des paramètres . . . . . 50
3.3.4 Validation du modèle ajusté . . . . . 51
3.3.5 Estimation des quantiles extrêmes . . . . . 52

**Conclusion** **53**

**Bibliographie** **54**

**Annexe : Abréviations et Notations** **57**

**Résumé** **59**

# Table des figures

1.1 Densités et Distributions de Lois des Valeurs Extrêmes . . . . . 13

2.1 Schéma illustratif de l'approche POT . . . . . 29

2.2 Densité et fonction de répartition de de lois de Pareto Généralisées . . . . . 32

3.1 Série des précipitations journalières en Angleterre couvrant la période du 1914 au 1962 . . . . . 42

3.2 Histogramme de la série des précipitations journalières . . . . . 42

3.3 Série des maxima annuels des précipitations journalières . . . . . 43

3.4 Graphiques de diagnostic de l'ajustement avec la distribution $GEV$ . . . . 46

3.5 Résultats graphiques de la sélection du seuil appliquée aux précipitations journalières en Angleterre (MRL-plot) . . . . . 48

3.6 Résultats graphiques de la sélection du seuil appliquée aux précipitations journalières en Angleterre (TC-plot) paramètre de forme (à gauche) et d'échelle (à droite) pour la $GPD$ . . . . . 48

3.7 Série des excès au-delà du seuil $u = 30\ mm$ (ligne horizontale), extraite à partir des précipitations journalières observées en Angleterre . . . . . 49

3.8 Graphiques de diagnostic de l'ajustement avec la $GPD$ . . . . . 51

# Liste des tableaux

1.1 Domaine d'attractions de quelques lois usuelles . . . . . 18

3.1 Caractéristiques statistiques des données . . . . . 41

3.2 Estimateurs des paramètres de la distribution GEV selon les méthodes MLE et PWM, avec intervalles de confiance à 95 . . . . . 44

3.3 Niveau de retour pour différentes période de retour et intervalles de confiance (GEV) . . . . . 47

3.4 Caractéristiques statistiques de la série des excès . . . . . 50

3.5 Estimateurs des paramètres de la GPD selon les méthodes MLE et PWM, avec intervalles de confiance à 95 . . . . . 51

3.6 Niveau de retour pour différentes période de retour et intervalles de confiance (GPD) . . . . . 52

# Introduction

Les événements extrêmes, bien que rares, jouent un rôle primordial dans de nombreux domaines tels que la météorologie, l'hydrologie, la finance, l'environnement, etc. Leur étude est essentielle pour évaluer et gérer les risques associés à ces événements. La modélisation des valeurs extrêmes constitue une branche fondamentale de la statistique, permettant d'estimer la probabilité d'occurrence et l'intensité de ces phénomènes.

L'intérêt pour la modélisation des extrêmes remonte au début du $XXe$ siècle, avec les travaux pionniers de Fisher et Tippett (1928) [14] ainsi que de Gnedenko (1943) [15], qui ont introduit et formalisé les lois limites pour les distributions des maxima. Ces travaux ont conduit au développement de deux distributions clés pour modéliser les valeurs extrêmes : la distribution des valeurs extrêmes généralisée ($GEV$) et la distribution de Pareto généralisée ($GPD$). La distribution $GEV$ s'applique à l'analyse des maxima observés sur des blocs de données, tandis que la distribution $GPD$ est utilisée pour modéliser les excès au-delà d'un seuil élevé.

L'objectif de ce travail est d'explorer ces deux distributions fondamentales, d'étudier leurs cadres théoriques et les méthodes d'estimation associées, puis de les appliquer à une série de données réelles.

Ce mémoire s'articule en trois chapitres, comme suit :

**Chapitre 1 (Distributions GEV) :** Ce chapitre présente le cadre théorique de la distribution $GEV$, en commençant par une introduction aux statistiques d'ordre. Il détaille

les caractéristiques des domaines d'attraction et les principales méthodes d'estimation de ses paramètres.

**Chapitre 2 (Distributions GPD) :** Ce chapitre est dédié à la distribution de Pareto généralisée, dans le cadre de l'approche $POT$. Il présente les méthodes graphiques de sélection du seuil, les propriétés de la $GPD$, les méthodes d'estimation des paramètres, ainsi que les quantiles extrêmes.

**Chapitre 3 (Application sous R) :** Ce chapitre présente une application pratique sur des données réelles, à savoir la série des précipitation journalières observées en Angleterre, en utilisant le logiciel statistique R.

Enfin, on achève ce travail par une brève conclusion.

# Chapitre 1

# Distributions GEV

Ce chapitre est consacré aux distributions de valeurs extrêmes généralisées ($GEV$, pour Generalized Extreme Value), des outils statistiques essentiels pour modéliser le comportement des événements extrêmes. La distribution $GEV$ regroupe, dans une même formulation unifiée, les lois de Gumbel, Fréchet et de Weibull. On y présente sa définition, ses principales propriétés, ainsi que les méthodes d'estimation de ses paramètres. Pour commencer, on introduit les notions de base et les propriétés des statistiques d'ordre, sur lesquelles repose les lois des valeurs extrêmes. Pour plus d'informations, on peut consulter les livres de Arnold et al (1992) [1], David et Nagaraja (2003) [7] et Embrechts et al (1997) [13].

## 1.1 Définitions et notions de base

### 1.1.1 Statistique d'ordre

**Définition 1.1.1 (Statistique d'ordre)**

*Soient $X_1, X_2, \cdots, X_n$, $n$ variables aléatoires (va's) indépendantes et identiquement distribuées $(i.i.d)$ d'une densité $f$ et d'une fonction de répartition commune $F$. Organisons ces va's en les classant par ordre croissant, on obtient les statistiques d'ordre associées à*

*l'échantillon* $(X_1, X_2, \cdots, X_n)$ *et que l'on note par*

$$X_{1,n} \leq X_{2,n} \leq \cdots \leq X_{n-1,n} \leq X_{n,n}.$$

**Remarque 1.1.1**

– *La première statistique d'ordre extrême* $X_{1,n}$ *correspond au minimum de l'échantillon, tandis que* $X_{n,n}$ *représente le maximum. En d'autres termes,*

$$X_{1,n} = \min(X_1, X_2, \cdots, X_n) \text{ et } X_{n,n} = \max(X_1, X_2, \cdots, X_n).$$

– *Pour* $1 \leq k \leq n$*, la va* $X_{k,n}$ *est appelée la* $k^{ième}$ *statistique d'ordre.*

### 1.1.2 Distributions d'une statistique d'ordre

Soient $X_{1,n}, X_{2,n}, \cdots, X_{n,n}$ les statistiques d'ordre associées à l'échantillon $(X_1, X_2, \cdots, X_n)$ constitué de n va's, ayant une fonction de répartition commune $F$ et une densité $f$.

**Proposition 1.1.1 (Distribution du minimum)**

*La fonction de répartition et la densité de la statistique d'ordre du minimum* $X_{1,n}$ *sont données respectivement par :*

$$F_{X_{1,n}}(x) = 1 - [1 - F(x)]^n, \ x \in \mathbb{R}, \tag{1.1}$$

*et*

$$f_{X_{1,n}}(x) = n[1 - F(x)]^{n-1} f(x), \ x \in \mathbb{R}. \tag{1.2}$$

En effet, pour $x \in \mathbb{R}$

$$F_{X_{1,n}}(x) = \mathbb{P}(X_{1,n} \leq x) = 1 - \mathbb{P}(X_{1,n} > x) = 1 - \mathbb{P}\left\{\bigcap_{i=1}^{n} X_i > x\right\} = 1 - \prod_{i=1}^{n} \mathbb{P}(X_i > x)$$
$$= 1 - \prod_{i=1}^{n} [1 - \mathbb{P}(X_i \leq x)] = 1 - [1 - F(x)]^n.$$

*Et on déduit la densité :*

$$f_{X_{1,n}}(x) = \frac{\partial F_{X_{1,n}}(x)}{\partial x} = n\left[1 - F(x)\right]^{n-1} f(x).$$

**Proposition 1.1.2 (Distribution du maximum)**

*La fonction de répartition et la densité de la statistique d'ordre du maximum $X_{n,n}$ sont données respectivement par :*

$$F_{X_{n,n}}(x) = \left[F(x)\right]^n, \ x \in \mathbb{R}, \tag{1.3}$$

*et*

$$f_{X_{n,n}}(x) = n\left[F(x)\right]^{n-1} \ f(x), \ x \in \mathbb{R}. \tag{1.4}$$

En effet,

$$F_{X_{n,n}}(x) = \mathbb{P}\left(X_{n,n} \leq x\right) = \mathbb{P}\left\{\bigcap_{i=1}^{n} X_i \leq x\right\} = \prod_{i=1}^{n} \mathbb{P}\left(X_i \leq x\right) = \left[F(x)\right]^n, \ x \in \mathbb{R}.$$

Et on déduit la densité :

$$f_{X_{n,n}}(x) = \frac{\partial F_{X_{n,n}}(x)}{\partial x} = n\left[F(x)\right]^{n-1} f(x), \ x \in \mathbb{R}.$$

**Proposition 1.1.3 (Distribution de la $k^{ième}$ statistique d'ordre $X_{k,n}$)**

*La fonction de répartition de la $k^{ième}$ statistique d'ordre $X_{k,n}$ $(1 \leq k \leq n)$, est donnée par :*

$$F_{X_{k,n}}(x) = \sum_{r=k}^{n} C_n^r \left[F(x)\right]^r \left[1 - F(x)\right]^{n-r}, \ x \in \mathbb{R}, \tag{1.5}$$

*et la fonction de densité de $X_{k,n}$ est la suivante*

$$f_{X_{k,n}}(x) = \frac{n!}{(k-1)!\,(n-k)!} f(x) \left[F(x)\right]^{k-1} \left[1 - F(x)\right]^{n-k}. \tag{1.6}$$

En effet, on a

$$\begin{aligned} F_{X_{k,n}}(x) &= \mathbb{P}(X_{k,n} \leq x) \\ &= \mathbb{P}(\text{Il y a un nombre supérieur ou égale à } k \text{ de } X_i \text{ inférieurs ou égale } x) \\ &= \mathbb{P}\left(\sum_{r=k}^{n} \mathbb{1}_{(X_i \leq x)} \geq k\right) \\ &= \sum_{r=k}^{n} C_n^r \left[F(x)\right]^r \left[1 - F(x)\right]^{n-r}, \ x \in \mathbb{R}, \end{aligned}$$

où $\mathbb{1}_A$ dénote la fonction indicatrice de l'ensemble $A$ et $C_n^r = \dfrac{n!}{r!\,(n-r)!}$ est la combinaison linéaire de $r$ éléments parmi $n$ éléments (sans remis).

Pour $1 \leq i \leq n$, les va's $\mathbb{1}_{(X_i \leq x)}$ étant $(i.i.d)$ de loi de Bernoulli de paramètres $F(x)$, ce qui implique que la loi $\sum_{i=1}^{n} \mathbb{1}_{(X_i \leq x)}$ est une loi Binomiale de paramètres $n$ et $F(x)$.
Pour la preuve de la densité $f_{k,n}$, voir Arnold et al (1992), page 10 [1].

### 1.1.3 Loi jointe d'un couple de statistique d'ordre

**Proposition 1.1.4 (Densité jointe d'un couple de statistique d'ordre)**

*Soit $(X_1, X_2, \cdots, X_n)$ un échantillon de va's de fonction de répartition commune $F$ continue et de fonction de densité $f$, la densité jointe du couple $(X_{j,n}, X_{k,n})$ avec $1 \leq i \leq j \leq n$ est donnée par :*

$$\begin{aligned} f_{(X_{j,n}, X_{k,n})}(x,y) = &\frac{n!}{(j-1)!\,(k-j-1)!\,(n-j)!} f(x)\, f(y) \left[F(x)\right]^{j-1} \\ &\times \left[F(y) - F(x)\right]^{k-j-1} \left[1 - F(y)\right]^{n-k}, \ \textit{avec} \ -\infty < x < y < +\infty. \end{aligned} \tag{1.7}$$

Afin d'obtenir la densité jointe d'un couple de statistique d'ordre, on visualise d'abord l'événement $(x \leq X_{i,n}, \leq x + \delta x, y \leq X_{j,n} \leq y + \delta y)$ comme suit (voir Shahbaz et al *[29])*

$$\underset{-\infty}{\overset{j-1}{\text{———}}}\Big]\underset{x}{\overset{1}{\text{——}}}\Big[\underset{x+\delta x}{\overset{k-j-1}{\text{———}}}\Big]\underset{y}{\overset{1}{\text{——}}}\Big[\underset{y+\delta y}{\overset{n-k}{\text{———}}}\underset{+\infty}{}$$

$X_i \leq x$ pour : $j-1$ des $X_i$ , $x \leq X_i \leq x+\delta x$ pour l'un des $X_i$, $x+\delta x \leq X_i \leq y$ pour $k-j-1$ des $X_i$, $y \leq X_i \leq y+\delta y$ pour l'un des $X_i$, et $X_i > y+\delta y$ pour $n-k$ restes des $X_i$. On peut écrire

$$\begin{aligned}\mathbb{P}\left(x \leq X_{j,n}, \leq x+\delta x, y \leq X_{k,n} \leq y+\delta y\right) = & \frac{n!}{(j-1)!\,(k-j-1)!\,(n-k)!} \\ & \times \left[F(x)\right]^{j-1}\left[F(x+\delta x)-F(x)\right] \\ & \times \left[F(y)-F(x+\delta x)\right]^{k-j-1} \\ & \times \left[F(y+\delta y)-F(y)\right]\left[1-F(y+\delta y)\right]^{n-k} \\ & + O\left((\delta x)^2\,\delta y\right) + O\left(\delta x\,(\delta y)^2\right),\end{aligned}$$

où $O\left((\delta x)^2\,\delta y\right)$ et $O\left(\delta x\,(\delta y)^2\right)$ sont des limites d'ordre plus supérieur, qui correspondent aux probabilités de l'événement d'avoir plus d'un $X_i$ dans l'intervalle $(x, x+\delta x]$ et un $X_i$ dans l'intervalle $(y, y+\delta y]$, et de l'événement d'avoir un $X_i$ dans l'intervalle $(x, x+\delta x]$ et plus d'un $X_i$ dans l'intervalle $(y, y+\delta y]$, respectivement.

La densité du couple $(X_{j,n}, X_{k,n})$ est donc

$$\begin{aligned}f_{(X_{j,n},X_{k,n})}(x,y) &= \lim_{\delta x\to 0,\delta y\to 0}\frac{\mathbb{P}\left(x \leq X_{j,n}, \leq x+\delta x, y \leq X_{k,n} \leq y+\delta y\right)}{\delta x \delta y} \\ &= \frac{n!}{(i-1)!\,(j-i-1)!\,(n-j)!} \\ &\quad \times f(x)\,f(y)\left[F(x)\right]^{j-1}\left[F(y)-F(x)\right]^{k-j-1}\left[1-F(y)\right]^{n-k}, \\ &\qquad -\infty < x < y < +\infty. \qquad (1.8)\end{aligned}$$

**Corollaire 1.1.1 (Densité jointe du minimum et maximum)**

*D'après le cas général de la densité jointe de $(X_{j,n}, X_{k,n})$ (1.8), on a :*

$$f_{(X_{1,n},X_{n,n})}(x,y) = n(n-1)\,f(x)\,f(y)\left[F(y)-F(x)\right]^{n-2},\ -\infty < x < y < +\infty. \quad (1.9)$$

**Corollaire 1.1.2 (Densité jointe de $n$ statistiques d'ordre)**

*Soit $X_1, \cdots, X_n$ une suite de va's (i.i.d) de densité $f$, alors la densité jointe de la statistique d'ordre lui associée $(X_{1,n}, \leq \cdots \leq, X_{n,n})$ est donné par :*

$$f_{(X_{1,n},X_{2,n},\cdots,X_{n,n})}(x_1, x_2, \cdots, x_n) = n! \prod_{i=1}^{n} f(x_i), \ -\infty < x_1 < x_2 < \cdots < x_n < +\infty. \tag{1.10}$$

**Proposition 1.1.5 (Distribution d'un couple de statistique d'ordre)**

*La fonction de répartition de $(X_{j,n}, X_{k,n})$ est donnée, dans deux cas, comme suit :*

$1^{ier}$ *cas :* $x \geq y$

$$F_{\left(X_{j,n},X_{k,n}\right)}(x, y) = \mathbb{P}(X_{j,n} \leq x, X_{k,n} \leq y) = \mathbb{P}(X_{k,n} \leq y) = F_{X_{k,n}}(y). \tag{1.11}$$

$2^{ème}$ *cas :* $x < y$

$$\begin{aligned} F_{\left(X_{j,n},X_{k,n}\right)}(x, y) &= \mathbb{P}(X_{j,n} \leq x, X_{k,n} \leq y) = \mathbb{P}(\textit{au moins } j \textit{ des } X_i \textit{ sont inférieur à } x, \\ &\qquad \textit{au moins } k \textit{ des } X_i \textit{ sont inférieur à } y) \\ &= \sum_{s=k}^{n} \sum_{r=j}^{s} \mathbb{P}(\textit{exactement } r \textit{ des } X_i \textit{ sont inférieur à } x, \\ &\qquad \textit{exactement } s \textit{ des } X_i \textit{ sont inférieur à } y) \\ &= \sum_{s=k}^{n} \sum_{r=j}^{s} \frac{n!}{r!\,(s-r)!\,(n-s)!} [F(x)]^r [F(y) - F(x)]^{s-r} [1 - F(y)]^{n-s}. \end{aligned} \tag{1.12}$$

Les références de base utilisée dans cette section est le livre de Arnold et al (1992) [1], le livre de David et Nagaraja (2003) [7] et celui de Shahbaz et al (2016) *[29].*

## 1.2 Approche des maxima par blocs

L'une des approches classiques pour extraire et analyser des événements extrêmes, dans une série de données, est la méthode des maxima par blocs (ou Block Maxima Approach). Cette méthode consiste à diviser une série temporelle en $n$ sous-ensembles, appelés blocs, de taille fixe $k$. Pour chaque bloc, on extrait la valeur maximale, considérée comme représentative de l'événement extrême survenu durant cette période. Ces maxima sont ensuite modélisés à l'aide de la distribution des valeurs extrêmes généralisée ($GEV$), qui regroupe les lois de Gumbel, Fréchet et Weibull.

En particulier la taille des blocs $k$ doit être assez haut que possible pour garantir des blocs avec suffisamment d'observations afin d'assurer une convergence adéquate de la distribution limite des maxima. Néanmoins, un nombre suffisant de maxima indépendants $n$ est nécessaire pour garantir une bonne inférence, ce qui conduit à un compromis entre le biais et la variance. En pratique, la taille des blocs est souvent déterminée par le contexte d'applications, par exemple selon que les données soient journalières, mensuelles ou annuelles, etc. La figure 3.1 illustre une telle partition des données.

## 1.3 Convergence du maximum renormalisé

Soit $X_1, \cdots, X_n$ une suite de va's $(i.i.d)$ de fonction de répartition commune $F$. Dans la pratique, la distribution $F$ n'est généralement pas connue, et il est nécessaire d'approximer la distribution du maximum donnée dans la proposition 1.3, par une distribution adaptée. En faisant tendre $n$ vers l'infini, on a (voir Coles (2001) [6]) :

$$\lim_{n\to\infty} \mathbb{P}\left(X_{n,n} \leq x\right) = \lim_{n\to\infty} F^n\left(x\right) = \begin{cases} 0, & \text{si} \quad F\left(x\right) < 1, \\ 1, & \text{si} \quad F\left(x\right) = 1. \end{cases} \tag{1.13}$$

Toutefois, ce résultat n'apporte que peu d'informations sur le comportement du maximum, car il conduit à une loi dégénérée (elle prend les valeurs 0 et 1 seulement).

En théorie des valeurs extrêmes, pour obtenir une loi limite non dégénérée, l'idée est de considérer non pas le maximum en tant que tel dans l'équation 1.13, mais une renormalisation de ce dernier. La renormalisation la plus simple consiste à faire une transformation linéaire, analogue à celle utilisée dans le Théorème Central Limite ($TCL$).

En effet, le $TCL$ montre que la distribution de la somme d'une suite de va's ($i.i.d$), de moyenne $\mu$ et de variance $\sigma^2$ finie, $S_n = \sum_{i=1}^{n} X_i$ converge vers une loi normale standard quand $n$ tend vers $\infty$, sous des conditions de normalisation affine appropriées. Plus précisément, pour tout $x \in \mathbb{R}$, on a :

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2} dt. \tag{1.14}$$

Afin d'éviter la dégénérescence mentionnée ci-dessus, les premiers résultats concernant les lois limites des valeurs extrêmes, après normalisation convenable, ont été obtenus par Fisher et Tippett en 1928 [14], et formalisés par Gnedenko en 1943 [15] à travers le théorème suivant, qui joue un rôle fondamental en théorie des valeurs extrêmes.

**Théoreme 1.3.1 (Fisher Tippett)**

*Soit $X_1, \cdots, X_n$ une suite de va's ($i.i.d$) de fonction de répartition $F$. Supposons qu'il existe deux suites de constantes $(a_n)_{n\geq 1} > 0$ et $(b_n)_{n\geq 1}$ réelles, telles que*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{X_{n,n} - b_n}{a_n} \leq x\right) = \lim_{n\to\infty} F_{X_{n,n}}(a_n x + b_n) = \lim_{n\to\infty} F^n(a_n x + b_n) = H(x), \ \forall x \in \mathbb{R} \tag{1.15}$$

*Où $H$ est une loi non dégénérée. Alors $H$ appartient nécessairement à l'une des familles suivantes :*

***Type I : Gumbel :*** $\Lambda(x) = \exp\{-\exp(-x)\}, \ x \in \mathbb{R}.$

***Type II : Fréchet :*** $\Phi_\alpha(x) = \begin{cases} \exp\{-x^{-\alpha}\} & si \quad x > 0, \\ 0 & si \quad x \leq 0. \end{cases} \quad \alpha > 0,$

$$\textbf{\textit{Type III : Weibull :}}\ \Psi_\alpha(x) = \begin{cases} \exp\left\{-(-x)^{-\alpha}\right\} & si \quad x < 0, \\ 1 & si \quad x \geq 0. \end{cases} \quad \alpha < 0,$$

*Où les lois limites possibles pour le maximum* $\Lambda$, $\Phi$ *et* $\Psi$ *sont appelées les distributions standards ou traditionnelles des valeurs extrêmes.*

**Exemple 1.3.1 (Loi exponentielle)**

*Soit* $X$ *une va suit la loi exponentielle standard de fonction de répartition*

$$F(x) = 1 - \exp(-x).$$

*Prenons* $a_n = 1$ *et* $b_n = \log(n)$, *alors* $\dfrac{X_{n,n} - b_n}{a_n}$ *tend asymptotiquement vers la loi de Gumbel. En effet :*

$$\begin{aligned} \lim_{n\to\infty} F_{X_{n,n}}(a_n x + b_n) &= \lim_{n\to\infty} F_{X_{n,n}}(x + \log(n)) = \lim_{n\to\infty} F^n(x + \log(n)) \\ &= \lim_{n\to\infty} (1 - \exp(-(x + \log(n))))^n = \lim_{n\to\infty} \left(1 - \frac{\exp(-x)}{n}\right)^n \\ &= \lim_{n\to\infty} \exp\left(n \log\left(1 - \frac{\exp(-x)}{n}\right)\right) \approx \exp(-\exp(-x)) \\ &= \Lambda(x). \end{aligned}$$

*Car :* $\lim\limits_{n\to\infty} \left(1 - \frac{x}{n}\right)^n = \exp(x)$.

## 1.4 Distributions des valeurs extrêmes généralisée

von Mises en 1954 [12] et Jenkinson en 1955 [20] ont introduit une famille paramétrique dite famille des lois des valeurs extrêmes généralisée ($GEV$). Cette distribution permet d'unifier les trois types de distributions extrêmes identifiés dans le théorème de Fisher

Tippet, comme établit le résultat suivant :

$$H_\theta(x) = \begin{cases} \exp\left\{-\left(1+\gamma\dfrac{x-\mu}{\sigma}\right)^{-1/\gamma}\right\} & , \quad \gamma \neq 0,\ 1+\gamma\dfrac{x-\mu}{\sigma} > 0, \\ \exp\left\{-\exp-\left(\dfrac{x-\mu}{\sigma}\right)\right\} & , \quad \gamma = 0,\ x \in \mathbb{R}. \end{cases} \tag{1.16}$$

Où $\theta := (\gamma, \mu, \sigma) \in \ominus \subset \mathbb{R}^2 \times \mathbb{R}_+$ sont respectivement les paramètres de forme (encore appelé indice des valeurs extrêmes), le paramètre de localisation et d'échelle.

La fonction de densité associée est définie par :

$$h_\theta(x) = \begin{cases} \frac{1}{\sigma}\left(1+\gamma\dfrac{x-\mu}{\sigma}\right)^{-1/\gamma-1} H_{\gamma,\mu,\sigma}(x) & , \quad \gamma \neq 0,\ 1+\gamma\dfrac{x-\mu}{\sigma} > 0, \\ \frac{1}{\sigma}\exp\left(-\left(\dfrac{x-\mu}{\sigma}\right)\exp\left[-\left(\dfrac{x-\mu}{\sigma}\right)\right]\right) & , \quad \gamma = 0,\ x \in \mathbb{R}. \end{cases} \tag{1.17}$$

La distribution $GEV$ standard est définie par la fonction de répartition suivante :

$$H_\gamma(x) = \begin{cases} \exp\left\{-\left[1+\gamma x\right]^{-1/\gamma}\right\} & , \quad \gamma \neq 0\ 1+\gamma x > 0, \\ \exp\left\{-\exp\left[-x\right]\right\} & , \quad \gamma = 0\ x \in \mathbb{R}. \end{cases} \tag{1.18}$$

Les lois de valeurs extrêmes généralisées correspondent à une translation et un changement d'échelle près aux lois de valeurs extrêmes. Nous avons alors les correspondances suivantes

$$\Lambda(x) = H_0(x). \tag{1.19}$$

$$\Phi_\alpha(x) = H_{1/\alpha}(\alpha(x-1)),\ x \in \mathbb{R}. \tag{1.20}$$

$$\Psi_\alpha(x) = H_{-1/\alpha}(\alpha(x+1)),\ x \in \mathbb{R}. \tag{1.21}$$

La figure 1.1 ci-dessous, illustre la forme des trois types de distributions extrêmes.

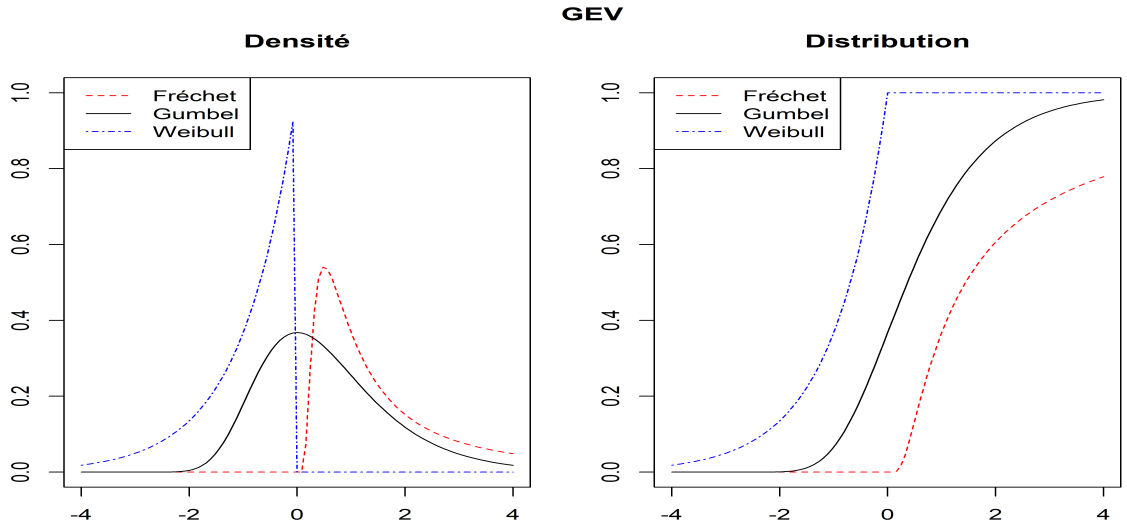


FIG. 1.1 – Densités et Distributions de Lois des Valeurs Extrêmes

## 1.5 Fonctions à variation régulière

Dans cette section, nous abordons la classe des fonctions, qui joue un rôle fondamental dans de nombreuses applications à travers l'ensemble des mathématiques. Nous allons introduire quelques notions générales sur ces fonctions.

**Définition 1.5.1 (Fonctions à variation régulière, de Bingham et al** (1984) **[3])**

*Une fonction mesurable* $g : \mathbb{R}^+ \to \mathbb{R}^+$ *est dite à variation régulière d'indice* $\alpha \in \mathbb{R}$ *et on note* $g \in RV_\alpha$*, si et seulement si'*

$$\lim_{t\to\infty} \frac{g(tx)}{g(t)} = x^\alpha, \ \forall x{>}0.$$

**Définition 1.5.2 (Fonction à variation lente )**

*Une fonction* $L$ *mesurable et positive sur* $]0, +\infty[$ *est à variations lentes, si pour tout* $x > 0$*, on a [9]*

$$\lim_{t\to\infty} \frac{L(tx)}{L(t)} = 1.$$

**Remarque 1.5.1**

*Toute fonction à variations régulières d'indice* $\alpha \in \mathbb{R}$*, peut s'écrire [22] :*

$$g(x) = x^{\alpha} L(x), \text{ avec } L(x) \in RV_0$$

**Théoreme 1.5.1 (Représentation de Karamata)**

*Si* $L \in RV_0$*, alors elle peut être écrite sous la forme :*

$$L(x) = c(x) \exp\left( \int_A^x \frac{\Delta(u)}{u} du \right), \ x \geq A,$$

*avec* $A > 0$*, où* $c(x) \to c_0 > 0$ *et* $\Delta(u) \to 0$ *quand* $x \to \infty$*.*

Cette représentation des fonctions à variations lentes est appelée représentation de Karamata, comme l'indiquent Bingham et al. (1984) [3]. Si la fonction $c(x)$ est constante, alors la fonction $L(x)$ est qualifiée de normalisée.

## 1.6 Domaines d'attraction

**Définition 1.6.1 (Domaine d'attraction)**

*On dit qu'une distribution* $F$ *appartient au domaine d'attraction de* $H_\gamma$*, et on note* $F \in \mathcal{D}(H_\gamma)$ *si la distribution du maximum renormalisée converge vers* $H_\gamma$*. Autrement dit, s'il existe des suites réelles* $a_n > 0$ *et* $b_n$ *tels que*

$$\lim_{n \to \infty} F^n (a_n x + b_n) = H_\gamma(x). \tag{1.22}$$

Dans la suite, on introduit les notations suivantes :

– La fonction de survie (ou queue de distribution) d'une va de fonction de répartition $F$ est

$$\bar{F}(x) = 1 - F(x). \tag{1.23}$$

– Le point terminal de la fonction $F$ est

$$x_F = \sup \{x \in \mathbb{R} : F(t) < 1\}. \tag{1.24}$$

– L'inverse généralisée (ou fonction des quantiles) de $F$ est définie par

$$F^{\leftarrow}(s) = \inf \{x \in \mathbb{R} : F(x) \geq s\}, \ 0 < s < 1. \tag{1.25}$$

– La fonction queue est définie par

$$U(t) = F^{\leftarrow}(1 - 1/t) = \left(1/\bar{F}\right)^{\leftarrow}(t), \ 1 < t < \infty. \tag{1.26}$$

**Proposition 1.6.1 (Caractérisation de $\mathcal{D}(H_\gamma)$)**

*On a $F \in \mathcal{D}(H_\gamma)$ si et seulement si*

$$n\bar{F}(xa_n + b_n) \to -\log H_\gamma(x) \ \ \textit{quand } n \to \infty. \tag{1.27}$$

*Pour certaine suite $(a_n, b_n)_{n\geq 1}$, avec $a_n > 0$ et $b_n \in \mathbb{R}$. On a alors la convergence en loi $(a_n^{-1}(M_n - b_n), n \geq 1)$ vers une va de fonction de répartition $H_\gamma$.*

Il faut bien noter que le paramètre de forme $\gamma$ conditionne le type de la loi des valeurs extrêmes. Nous présentons dans ce qui suit, les domaines d'attractions dans les trois cas correspondant au signe du paramètre $\gamma$.

### 1.6.1 Domaine d'attraction de Fréchet

Cas où $\gamma > 0$ : Ce cas correspond au domaine d'attraction de Fréchet, noté $D(\Phi_\gamma)$. Les lois appartenant à ce domaine d'attraction sont caractérisées par une queue à décroissance lente (polynomiale) à l'infini, et un point terminal $x_F = +\infty$. Elles sont dites aussi lois à queues lourdes (non-exponentielle). Une caractérisation de ce domaine d'attraction est donnée par le théorème suivant (pour la démonstration, voir Gnedenko (1943) [15]) :

**Théoreme 1.6.1 (Caractérisation du $\mathcal{D}(\Phi_\gamma)$)**

*La fonction de répartition $F$ appartient au $D(\Phi_\gamma)$ avec $\gamma > 0$, si et seulement si :*

$$\bar{F}(X) = x^{-\gamma} L(x), \tag{1.28}$$

*où la fonction $L$ est à variation lente.*

*Dans ce cas avec les suites de normalisation $a_n = F^{\leftarrow}\left(1 - \frac{1}{n}\right)$ et $b_n = 0$, la suite $\left(a_n^{-1} X_{n,n}\right)_{n\geq 1}$ converge en loi vers une va de fonction de répartition $\Phi_\gamma$ quand $n \to \infty$.*

### 1.6.2 Domaine d'attraction de Weibull

Cas où $\gamma < 0$ : Le domaine d'attraction dans ce cas est celui de Weibull, noté $D(\Psi_\gamma)$. Les lois de ce domaine sont bornées à droite, et par conséquent, le point terminal $x_F$ est fini. Une caractérisation d'appartenance à ce domaine d'attraction est donnée par le théorème suivant (pour la démonstration, voir Gnedenko (1943))[15] :

**Théoreme 1.6.2 (Caractérisation du $\mathcal{D}(\Psi_\gamma)$)**

*La fonction de répartition $F$ appartient au domaine d'attraction de la loi de Weibull de paramètre $\gamma > 0$ si et seulement si $x_F < +\infty$ et*

$$\bar{F}\left(x_F - \frac{1}{x}\right) = x^{-\gamma} L(x), \tag{1.29}$$

*où la fonction $L$ est à variation lente. Dans ce cas, un choix possible pour les suites $a_n$ et $b_n$ est $a_n = x_F - F^{\leftarrow}\left(1 - \frac{1}{n}\right)$ et $b_n = x_F$, la suite $\left(a_n^{-1}\left(X_{n,n} - x_F\right)\right)_{n\geq 1}$ converge en loi vers une va de fonction de répartition $\Psi_\gamma$ quand $n \to \infty$.*

### 1.6.3 Domaine d'attraction de Gumbel

Cas où $\gamma = 0$ : La loi présente dans la queue une décroissance de type exponentielle, ce qui permet de caractériser dans ce cas, le domaine d'attraction de Gumbel $D(\Lambda)$. Ce dernier est délicatement traitable, car il n y a pas de lien direct entre la queue de la loi et les fonctions à variations lentes définies par (Delmas et Jourdain (2006)) [10]. von Mises (1936) [24] a donné une caractérisation simple pour le domaine d'attraction de Gumbel, formulée par le biais du théorème suivant.

**Théoreme 1.6.3 (Caractérisation I du $\mathcal{D}(\Lambda)$)**

*La fonction $F$ appartient au $D(\Lambda)$ avec $\gamma = 0$, si et seulement s'il existe une certaine $z < x_F$ tels que $\bar{F}$ a la représentation suivante*

$$\bar{F}(x) = c(x) \exp\left\{ - \int_z^x \frac{g(t)}{a(t)} dt \right\}, \ z < x < x_F. \tag{1.30}$$

*où $g$ et $c$ sont des fonctions mesurable tels que $c(x) \to c > 0$, $g(x) \to 1$ quand $x \to x_F$, et $a(x)$ est une fonction positive absolument continue avec la densité á ayant $\lim_{x \to x_F} á(x) = 0$. Dans ce cas, nous pouvons choisir $a_n = a(b_n)$ et $b_n = F^{\leftarrow}\left(1 - \frac{1}{n}\right)$.*

*Un choix possible pour la fonction $a$ est*

$$a(x) = \int_{a_n}^{x_F} \frac{\bar{F}(t)}{\bar{F}(x)} dt = E(X - x \,|X > x), \ x < x_F. \tag{1.31}$$

*Cette fonction s'appelle habituellement la moyenne des excès (mean-excess function).*

**Théoreme 1.6.4 (Caractérisation II du $\mathcal{D}(\Lambda)$)**

*La fonction $F$ appartient au $D(\Lambda)$ s'il existe une fonction positive $\tilde{a}$, appelée fonction auxiliaire telle que :*

$$\lim_{x \to x_F} \frac{\bar{F}(x + t\tilde{a}(x))}{\bar{F}(x)} = \exp(-t), \ t \in \mathbb{R}. \tag{1.32}$$

*Un choix possible pour $\tilde{a}$ est la fonction $a$ donnée par (1.31).*

**Proposition 1.6.2**

*Soit $X$ une va positive, alors les affirmation suivantes sont équivalentes (voir Embrechts et al. (1997)) [13] :*

1. $X \sim \Phi_\gamma$,
2. $\ln X^\gamma \sim \Lambda$,
3. $-X^{-1} \sim \Psi_\gamma$.

Dans le tableau ci-dessous 1.1, on présente les domaine d'attraction de certaine lois :

| **Domaine d'attraction** | Gumbel ($\gamma = 0$) | Weibull ($\gamma < 0$) | Fréchet ($\gamma > 0$) |
|---|---|---|---|
| **Lois** | Normale<br>Exponentielle<br>Gamma<br>Lognormale<br>Weibull | Beta<br>Uniforme | Cauchy<br>Pareto<br>Student<br>Loggamma<br>Burr |

TAB. 1.1 – Domaine d'attractions de quelques lois usuelles

## 1.7 Estimation des paramètres

### 1.7.1 Estimation paramétrique

Plusieurs méthodes d'estimation des paramètres $\theta = (\gamma, \mu, \sigma)$ de la distribution $GEV$ *(1.16)* sont disponibles dans la littérature. Dans cette section, nous nous concentrerons sur les plus populaires : la méthode du maximum de vraisemblance ($MLE$ : Maximum Likelihood Estimation) et celle des moments de probabilités pondérés ($PWM$ : Probability-Weighted Moment).

#### Méthode du maximum de vraisemblance

L'estimation par la méthode maximum de vraisemblance donne des résultats asymptotiques et efficaces, les estimateurs obtenus convergent sous certaines conditions vers les

vraies valeurs des paramètres. Cette méthode consiste à choisir comme estimateur de $\theta$ la valeur qui maximise la vraisemblance sur un espace de paramètres $\ominus \subset \mathbb{R}^2 \times \mathbb{R}_+$. Soit $(X_1, X_2, \cdots, X_n)$ un échantillon de va's pour lesquelles la distribution $GEV$ est appropriée. La vraisemblance s'exprime de la manière suivante :

$$L(\theta; x_1, \cdots, x_n) = \prod_{i=1}^{n} h_\theta(x_i) \mathbb{1}_{\{1+\gamma(x_i-\mu)/\sigma>0\}}, \tag{1.33}$$

Il est souvent plus facile de calculer $\hat{\theta}$ par maximisation de la fonction log-vraisemblance au lieu de la vraisemblance elle-même. La fonction log-vraisemblance est donnée par :

$$l(\theta; x_1, \cdots, x_n) = \log L(\theta; x_1, \cdots, x_n).$$

Par conséquent, dans le cas $\gamma \neq 0$, la fonction log-vraisemblance est égale à

$$\begin{aligned} l(\theta; x_1, \cdots, x_n) &= \sum_{i=1}^{n} \log h_\theta(x_i) \mathbb{1}_{\{1+\gamma(x_i-\mu)/\sigma>0\}} \\ &= -n \log \sigma - \left(\frac{1}{\gamma} + 1\right) \sum_{i=1}^{n} \log\left(1 + \gamma \frac{x_i - \mu}{\sigma}\right) \\ &\quad - \sum_{i=1}^{n} \left(1 + \gamma \frac{x_i - \mu}{\sigma}\right)^{-1/\gamma}. \end{aligned} \tag{1.34}$$

L'estimateur du maximum de vraisemblance correspond alors

$$\hat{\theta}_n = \hat{\theta}_n(x_1, \cdots, x_n) = \arg\max_{\theta \in \ominus} l(\theta; x_1, \cdots, x_n). \tag{1.35}$$

Si $l(\theta; x_1, \cdots, x_n)$ admet des dérivées partielles par rapport à $\gamma, \mu$ et $\sigma$ (respectivement), alors le $MLE$ est la solution du système d'équations suivant

$$\frac{\partial l(\theta; x_1, \cdots, x_n)}{\partial \theta} = 0. \tag{1.36}$$

Dans le cas où $\gamma = 0$, la fonction log-vraisemblance est égale à

$$l(\theta; x_1, \cdots, x_n) = -n \log \sigma - \sum_{i=1}^{n} \exp\left(-\frac{x_i - \mu}{\sigma}\right) - \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma}. \tag{1.37}$$

En dérivant cette fonction relativement aux deux paramètres $\mu$ et $\sigma$ (respectivement), nous obtenons le système des équations à résoudre suivant

$$\begin{cases} n - \sum_{i=1}^{n} \exp\left(-\frac{x_i - \mu}{\sigma}\right) = 0, \\ n + \sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma}\right) \left(\exp\left\{-\frac{x_i - \mu}{\sigma}\right\} - 1\right) = 0. \end{cases}$$

Il convient toutefois de souligner qu'il n'existe généralement pas de solution explicite aux équations non linéaires issues de la maximisation de la log-vraisemblance. Ainsi, le recours à des méthodes numériques et à des algorithmes d'optimisation est nécessaire pour estimer les paramètres. En pratique, le calcul de ces estimateurs ne présente pas de difficulté majeure.

Cependant, la régularité des estimateurs en particulier leur efficacité asymptotique et leur normalité n'est pas systématiquement garantie. Smith en 1987 [30] a montré qu'il suffit que le paramètre de forme $\gamma > -1/2$ pour que les conditions de régularité de l'estimateur du maximum de vraisemblance soient satisfaites. Pour un traitement plus détaillé, on pourra consulter l'ouvrage de Castillo et al. (2005) [5].

### Méthode des moments de probabilités pondérés

Les moments de probabilités pondérés constituent une généralisation des moments classiques d'une distribution de probabilité. Cette notion a été introduite par Greenwood et al. (1979) [16]. Les moments de probabilités pondérés sont définis par :

$$M_{p,r,s} = E\left[X^p \{F(X)\}^r \{1 - F(X)\}^s\right], \text{ où } p, r, s \in \mathbb{R}. \tag{1.38}$$

Les moments de probabilités pondérés sont susceptibles d'être les plus utiles quand l'inverse de la distribution peut être écrit sous une forme fermée, pour cela nous pouvons écrire

$$M_{p,r,s} = \int_0^1 \{F^{\leftarrow}\}^p F^r \{1-F\}^s \, dF, \tag{1.39}$$

et ceci souvent la manière la plus commode d'évaluer ces moments. Le cas spécifique de l'estimation par la méthode des moments de probabilités pondérés pour la distribution $GEV$ est étudié intensivement en Hosking et al. (1985) [17]. Au cas où $\gamma \neq 0$ , plaçant $p = 1$, $r = 0, 1, 2, \cdots$ et $s = 0$, ils rendraient pour la distribution $GEV$

$$M_{1,r,0} = E\left[X^1 \{F(X)\}^r\right] = \int_0^1 H_\theta^{\leftarrow}(y)\, y^r dy, \tag{1.40}$$

où $r \in \mathbb{N}$ et pour $0 < y < 1$,

$$H_\theta^{\leftarrow}(y) = \begin{cases} \mu - \dfrac{\sigma}{\gamma}\left(1 - (-\log y)^{-\gamma}\right) & si \quad \gamma \neq 0, \\ \mu - \sigma \log(-\log y) & si \quad \gamma = 0. \end{cases} \tag{1.41}$$

Par conséquent, les moments de probabilités pondérés pour la distribution $GEV$ deviennent

$$M_{1,r,0} = \frac{1}{r+1}\left\{\mu - \frac{\sigma}{\gamma}\left[1 - (r+1)^\gamma \Gamma(1-\gamma)\right]\right\}, \ \gamma < 1, \tag{1.42}$$

où $\Gamma(.)$ désigne la fonction gamma, $\Gamma(t) := \int_0^{+\infty} x^{t-1} e^{-x} dx, t \geq 0$.

Soit $(X_1, X_2, \cdots, X_n)$ un échantillon de $n$ va's de distribution $GEV$, avec les statistiques d'ordre associées $X_{1,n}, X_{2,n}, \cdots, X_{n,n}$. L'estimateur des $PWM$ de $\theta$ est la solution du

système des équations suivant, obtenues à partir l'équation *(1.42)*, avec $r \in \mathbb{N}$,

$$M_{1,0,0} = \mu - \frac{\sigma}{\gamma}\left(1 - \Gamma\left(1-\gamma\right)\right), \tag{1.43}$$

$$2M_{1,1,0} - M_{1,0,0} = \frac{\sigma}{\gamma}\Gamma\left(1-\gamma\right)\left(2^{\gamma}-1\right), \tag{1.44}$$

$$\frac{3M_{1,2,0} - M_{1,0,0}}{2M_{1,1,0} - M_{1,0,0}} = \frac{3^{\gamma}-1}{2^{\gamma}-1}. \tag{1.45}$$

Après avoir remplacé $M_{1,r,0}$ par son estimateur sans biais (tel que proposé par Landwehr et al.(1979) [21]), on obtient

$$\hat{M}_{1,r,0} = \frac{1}{n}\sum_{i=1}^{n}\left(\prod_{i=1}^{n}\frac{(j-l)}{(n-l)}\right)X_{j,n}, \tag{1.46}$$

ou par l'estimateur consistant qui est asymptotiquement équivalent

$$\tilde{M}_{1,r,0} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{j}{n+1}\right)^{r}X_{j,n}. \tag{1.47}$$

Notons que pour obtenir $\hat{\gamma}$, l'équation *(1.45)* doit être résolus numériquement. Après, l'équation *(1.44)* peut ensuite être utilisée pour obtenir $\hat{\sigma}$

$$\hat{\sigma} = \frac{\hat{\gamma}\left(2\hat{M}_{1,1,0} - \hat{M}_{1,0,0}\right)}{\Gamma\left(1-\hat{\gamma}\right)\left(2^{\hat{\gamma}}-1\right)}. \tag{1.48}$$

En fin, donné $\hat{\gamma}$ et $\hat{\sigma}$ , $\hat{\mu}$ peut être obtenu à partir l'équation (1.43)

$$\hat{\mu} = \hat{M}_{1,0,0} + \frac{\hat{\sigma}}{\hat{\gamma}}\left(1 - \Gamma\left(1-\hat{\gamma}\right)\right), \tag{1.49}$$

voir Beirlant et al. (2004) [2], pour plus de détails.

### 1.7.2 Estimation semi-paramétrique

Des méthodes statistiques semi-paramétriques adaptées à ce contexte ne nécessitent pas la connaissance de la distribution complète, mais uniquement des informations portant sur les queues de celle-ci. Les estimateurs classiques dans ce cadre reposent sur les plus grandes valeurs de l'échantillon mais sans prendre trop de valeurs de ce dernier, c'est-à-dire les statistiques d'ordre $X_{n-k+1,n} \leq \cdots \leq X_{n,n}$,où $k$ est une suite intermédiaire d'entiers dépendant de la taille de l'échantillon $n$, telle que

$$k = k(n) \to \infty \text{ et } k/n \to \infty \text{ quand } n \to \infty.$$

#### Estimateur de Hill

Cet estimateur a été introduit par Hill en 1975 [19], il est limité au cas de Fréchet $\gamma > 0$, l'estimateur de Hill est probablement l'estimateur le plus étudié dans la littérature. Il est défini par

$$\hat{\gamma}^H_{k(n)} = \frac{1}{k\,(n) - 1} \sum_{i=n-k(n)+2}^{n} \log X_{i,n} - \log X_{n-k(n)+1,n},$$

ou encore par

$$\hat{\gamma}^H_{k(n)} = \frac{1}{k\,(n)} \sum_{i=1}^{k(n)} \log X_{n-i+1,n} - \log X_{n-k(n),n.}$$

Avant d'énoncer les résultats sur le comportement asymptotique de l'estimateur de Hill, on doit imposer la condition de la fonction à variation régulière de second ordre.

**Définition 1.7.1 (Fonction à variation régulière du second ordre)**

*On dit que la queue de* $F \in \mathcal{D}\left(\Phi_{1/\gamma}\right)$, $\gamma > 0$, *est a variation régulière du second ordre à l'infinie si la condition suivante est satisfaite : Il existe un certain paramètre* $\rho \leq 0$, *et une fonction* $A$ *satisfaisant* $\lim_{t\to\infty} A\,(t) = 0$ *et ne changeant pas son signe près de* $\infty$, *telles que pour tout* $x > 0$

$$\lim_{t\to\infty} \frac{\left(U\,(tx)\,/U\,(t)\right) - x^{\gamma}}{A\,(t)} = x^{\gamma} \frac{x^{\rho} - 1}{\rho}. \tag{1.50}$$

*Si* $\rho = 0$, $x^\rho - 1/\rho$ *s'interprète comme* $\log x$.

**Théoreme 1.7.1 (Propriétés asymptotique de $\hat{\gamma}^H_{k(n)}$)**

*Soit* $(X_1, X_2, \cdots, X_n)$ *un échantillon de* $n$ *va's de fonction de réparation* $F \in \mathcal{D}\left(H_{1/\gamma}\right)$, $\gamma > 0$ *.Supposons* $k \to \infty$ *et* $k/n \to \infty$ *quand* $n \to \infty$.

(1) ***Consistance faible :***

$$\hat{\gamma}^H_{k(n)} \xrightarrow{p} \gamma \text{ quand } n \to \infty.$$

(2) ***Consistance forte :*** *Si* $k/\log\log n \to \infty$ *quand* $n \to \infty$, *alors*

$$\hat{\gamma}^H_{k(n)} \xrightarrow{p.s.} \gamma \text{ quand } n \to \infty.$$

(3) ***Normalité asymptotique :*** *Supposons que* $F$ *satisfaisant la condition (1.50), avec* $\sqrt{k}A\left(n/k\right) \to \lambda$ *quand* $n \to \infty$, *alors*

$$\sqrt{k}\left(\hat{\gamma}^H_{k(n)} - \gamma\right) \xrightarrow{d} \mathcal{N}\left(\frac{\lambda}{1-\rho}, \gamma^2\right) \text{ quand } n \to \infty.$$

Ce dernier résultat permet de calculer des intervalles de confiance pour $\gamma$. Par exemple, à un niveau de confiance de $(1-\alpha)\,\%$, on a pour $\lambda = 0$

$$\gamma \in \left[\hat{\gamma}^H_{k(n)} - q_{1-\alpha/2}\frac{\hat{\gamma}^H_{k(n)}}{\sqrt{k\left(n\right)}}; \hat{\gamma}^H_{k(n)} + q_{1-\alpha/2}\frac{\hat{\gamma}^H_{k(n)}}{\sqrt{k\left(n\right)}}\right],$$

où $q_{1-\alpha/2}$ est le quantile d'ordre $(1-\alpha/2)$ d'une loi normale centrée réduite.

## Estimateur de Pickands

L'estimateur de Pickands repose sur l'utilisation des statistiques d'ordre. Il présente l'avantage d'être applicable quelque soit le domaine d'attraction de la distribution $F$. Cet esti-

mateur a été introduit par Pickands en 1975 [26]. Il est défini par la statistique

$$\hat{\gamma}^{P}_{k(n)} := \frac{1}{\log 2} \log \left( \frac{X_{n-k(n)+1,n} - X_{n-2k(n)+1,n}}{X_{n-2k(n)+1,n} - X_{n-4k(n)+1,n}} \right). \tag{1.51}$$

Pickands a prouvé la consistance faible de son estimateur. La convergence forte ainsi que la normalité asymptotique ont été établies par Dekkers et de Haan (1989) [11].

**Théoreme 1.7.2 (Propriétés asymptotique de $\hat{\gamma}^{P}_{k(n)}$)**

*Soit $X_1, X_2, \cdots, X_n$ une suite de va's (i.i.d) de fonction de répartition $F \in \mathcal{D}(H_\gamma)$, où $\gamma \in \mathbb{R}$. Si $k(n) \to \infty$ et $k(n)/n \to 0$ quand $n \to \infty$, alors*

(1) ***Consistance faible :***

$$\hat{\gamma}^{P}_{k(n)} \xrightarrow{p} \gamma \text{ quand } n \to \infty.$$

(2) ***Consistance forte :*** *Si $k/\log\log n \to \infty$ quand $n \to \infty$, alors*

$$\hat{\gamma}^{P}_{k(n)} \xrightarrow{p.s.} \gamma \text{ quand } n \to \infty.$$

(3) ***Normalité asymptotique :*** *Sous des conditions additionnelles sur la suite intermédiaire $k = k(n)$ et la fonction $F$, on a :*

$$\sqrt{k}\left(\hat{\gamma}^{P}_{k(n)} - \gamma\right) \to \mathcal{N}\left(0, \frac{\gamma^2\left(2^{2\gamma+1}+1\right)}{\left\{2\left(2^{\gamma}-1\right)\log 2\right\}^2}\right) \text{ quand } n \to \infty.$$

## 1.8 Estimation des quantiles extrêmes

Les estimateurs des quantiles extrêmes du *GEV* (i.e. les quantiles d'ordre $1-p$) peuvent être obtenues en inversant la fonction de distribution $H_\theta$ donnée par (1.16) et remplaçant

$\theta = (\gamma, \mu, \sigma)$ par $\hat{\theta} = (\hat{\gamma}, \hat{\mu}, \hat{\sigma})$ le $MLE$ ou Les $PWM$. Ils se présentent comme suit

$$\begin{aligned} \hat{x}_p &= H_{\hat{\theta}}^{\leftarrow}(1-p) \\ &= \begin{cases} \hat{\mu} - \dfrac{\hat{\sigma}}{\hat{\gamma}}\left(1 - (-\log(1-p))^{-\hat{\gamma}}\right) & si \quad \gamma \neq 0, \\ \hat{\mu} - \hat{\sigma}\log(-\log(1-p)) & si \quad \gamma = 0. \end{cases} \end{aligned} \tag{1.52}$$

Dans la terminologie courante $\hat{x}_p$ est le niveau de retour associé à la période de retour $T = 1/p$, car avec un degré raisonnable de précision, le niveau $\hat{x}_p$ devrait être dépassé en moyenne une fois tous les $1/p$ ans. Plus précisément, $\hat{x}_p$ est dépassé par le maximum annuel d'une année donnée avec une probabilité $p$, pour plus de détails, voir Coles (2001) [6].

– Si $F \in D(H_\theta)$, en utilisant la relation (1.27) pour les grands seuils $u = a_n x + b_n$, nous obtenons l'estimateur de la queue de la distribution

$$\hat{\bar{F}}(u) = \frac{1}{n}\left(1 + \hat{\gamma}\frac{u - \hat{b}_n}{\hat{a}_n}\right)^{-1/\hat{\gamma}}, \tag{1.53}$$

où $\hat{\gamma}$, $\hat{a}_n$ et $\hat{b}_n$ sont des estimateurs basés sur les $k$ plus grande statistiques d'ordre, de l'indice $\gamma$ et de constantes de normalisation $a_n$ et $b_n$ (respectivement). Dans le cas où les quantiles extrêmes sont à l'intérieur des données (i.e. $p \geq 1/n$). Ils peuvent être estimés par

$$\hat{x}_p := \hat{a}_n \frac{(np)^{-\hat{\gamma}} - 1}{\hat{\gamma}} + \hat{b}_n. \tag{1.54}$$

Les constants de normalisation $\hat{a}_n$ et $\hat{b}_n$ ont une très grande variance, car elles sont fondées sur des quantiles élevés de $X$. Afin de pallier ce problème (Dekkers et de Haan (1989) [12] ; Dekkers et al. (1986) [11]) proposent d'utiliser les $k$ plus grandes valeurs de l'échantillon afin d'estimer la queue de la distribution. Pour $x$ suffisamment grand,

$$\hat{\bar{F}}(u) = \frac{k}{n}\left(1 + \hat{\gamma}\frac{u - \hat{b}_{n/k}}{\hat{a}_{n/k}}\right)^{-1/\hat{\gamma}}, \tag{1.55}$$

on en déduit le cas où les quantiles extrêmes sont hors des données (i.e. $p < 1/n$)

$$\hat{x}_p := \hat{a}_{n/k} \frac{(np/k)^{-\hat{\gamma}} - 1}{\hat{\gamma}} + \hat{b}_{n/k}. \tag{1.56}$$

– Les estimateurs des quantiles extrêmes associés aux estimateurs semi paramétrique que nous présentons s'écrivent sous cette forme. Il reste donc à donner des estimations pour les constants de normalisation $a_{n/k}$ et $b_{n/k}$.

L'estimateur du quantile d'ordre $(1-p)$ associé à l'estimateur de Hill est donné par

$$\hat{x}_p^H := X_{n-k,n} \left( \frac{k}{np} \right)^{\hat{\gamma}^H_{k(n)}}, \tag{1.57}$$

où $\hat{b}_{n/k} = \hat{a}_{n/k}/\hat{\gamma}^H_{k(n)} = X_{n-k,n}$.

L'estimateur du quantile d'ordre $(1-p)$ lié à l'estimateur de Pickands est de la forme suivante

$$\hat{x}_p^P := X_{n-k+1,n} + \frac{(np/k)^{-\hat{\gamma}^P_{k(n)}} - 1}{1 - 2^{-\hat{\gamma}^P_{k(n)}}} \left( X_{n-k+1,n} - X_{n-2k+1,n} \right). \tag{1.58}$$

où $\hat{a}_{n/k} = \frac{\hat{\gamma}^P_{k(n)}}{1-2^{-\hat{\gamma}^P_{k(n)}}} \left( X_{n-k+1,n} - X_{n-2k+1,n} \right)$ et $\hat{b}_{n/k} = X_{n-k+1,n}$. Les propriétés asymptotiques de cet estimateur sont discutées par Dekkers et de Haan (1989) [12].

# Chapitre 2

# Distrubutions GPD

Contrairement à l'approche des maxima par blocs, qui ne retient que la valeur maximale observée dans chaque bloc, l'approche des excès au-dessus d'un seuil ($POT$ : Peaks Over Threshold) exploite l'ensemble des observations dépassant un seuil prédéfini, ce qui permet de mieux utiliser l'information contenue dans données extrêmes. Dans ce chapitre, on s'intéresse aux distributions de Pareto généralisées ($GPD$ : Generalized Pareto distribution), un outil fondamental de la théorie des valeurs extrêmes pour modéliser les excès au-delà d'un seuil. Après une présentation de leur cadre théorique, nous introduirons les méthodes de la sélection du seuil, les méthodes d'estimation des paramètres, ainsi que les quantiles extrêmes.

## 2.1 Approche POT

L'approche des maxima par blocs, fondée sur la distribution $GEV$, peut être réductrice, du fait que l'utilisation d'un seul maximum conduit à une perte d'information, en ignorant les autres valeurs extrêmes de l'échantillon.

Pour pallier ce problème, une alternative possible dans la modélisation des évènements extrêmes est la méthode des excès au-dessus d'un seuil ($POT$). Cette approche, complémentaire à l'analyse des maxima, présente l'avantage d'être plus efficace du point de vue

statistique, puisqu'elle repose sur un échantillon de taille plus importante en exploitant toutes les observations excédant un seuil réel, noté $u$ ni trop faible pour ne pas prendre en considération des valeurs non extrêmes, ni trop élevé pour avoir suffisamment d'observations, il est inférieur au point terminal ($u < x_F$).

Cette méthode a été introduite par Pickands (1975) [25], basé sur la distribution de Pareto Généralisées ($GPD$). Elle a été étudiée par divers auteurs, notamment Smith (1987) [29], Davison et Smith (1990) [8] et Reiss et Thomas (1997) [26].

### 2.1.1 Modélisation des excès

Soit $(X_1, X_2, \cdots, X_n)$ un échantillon de va's. On définit un seuil $u \in \mathbb{R}$, $N_u = card\{i : i = 1, \cdots, n\ ;\ X$ et $Y_i = X_i - u > 0$ pour $0 \leq i \leq N_u$, où $N_u$ est le nombre de dépassements du seuil $u$ par les $X_i$ et $Y_i$ sont les excès correspondants. La figure 2.1 donne une représentation graphique de l'approche $POT$.

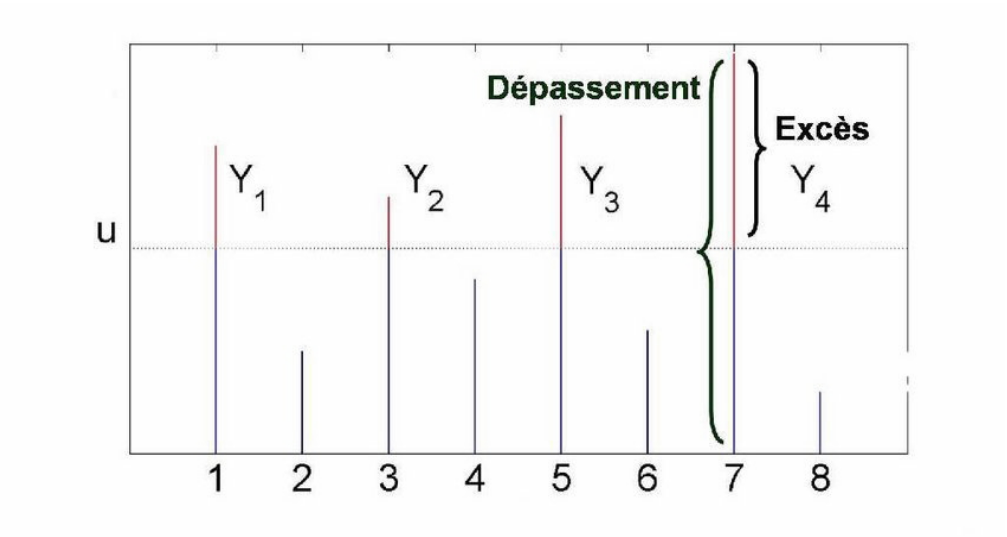


FIG. 2.1 – Schéma illustratif de l'approche POT

La méthode $POT$ repose sur l'approximation de la loi des excès au-dessus d'un seuil $u$ pour une variable aléatoire réelle $X$. Plus précisément, elle s'intéresse à la loi conditionnelle de la variable aléatoire $X - u$, étant donné que $X > u$. L'objectif est alors de déterminer une

loi de probabilité permettant d'approximer cette distribution conditionnelle.

### 2.1.2 Distribution des excès

**Définition 2.1.1 (Fonction de répartition des excès)**

*La fonction de répartition des excès au-dessus d'un seuil $u$, notée $F_u$, est définie par [6] [13] :*

$$F_u(y) = \mathbb{P}(X - u \leq y \,|X > u) = 1 - \frac{\bar{F}(u+y)}{\bar{F}(u)},\ 0 < y < x_F - u. \tag{2.1}$$

*Donc pour tout $y \in \mathbb{R}$, on a*

$$F_u(y) = \begin{cases} 0 & si \quad y \leq 0, \\ 1 - \dfrac{\bar{F}(u+y)}{\bar{F}(u)} & si \quad 0 < y < x_F - u, \\ 1 & si \quad y \geq x_F - u. \end{cases}$$

*Ou de manière équivalente :*

$$\bar{F}_u(y) = 1 - F_u(y) = \mathbb{P}(X - u > y \,|X > u) = \frac{\bar{F}(u+y)}{\bar{F}(u)},\ 0 < y < x_F - u.$$

**Définition 2.1.2 (Fonction moyenne des excès)**

*Soit $X$ une va réelle ayant une fonction de répartition $F$ et une fonction de survie $\bar{F}$. On appelle fonction moyenne des excès de $X$ par rapport au seuil $u$, notée $e(u)$, l'espérance conditionnelle de l'excès de $X$ au-delà de $u$, définie pour tout $u < x_F$ par :*

$$e(u) = E[X - u \,|X > u] = \frac{1}{\bar{F}(u)} \int_u^{x_F} \bar{F}(t)dt,\ u < x_F. \tag{2.2}$$

**Définition 2.1.3 (Fonction moyenne des excès empirique)**

*Soit* $(X_1, X_2, \cdots, X_n)$ *un échantillon de taille* $n \in \mathbb{N}^*$ *issu de la va* $X$*, et soit* $F_n$ *la fonction de répartition empirique associée. On appelle fonction moyenne des excès empirique de* $X$ *par rapport au seuil* $u < x_F$*, basée sur l'échantillon* $(X_1, X_2, \cdots, X_n)$*, la fonction* $e_n(u)$ *définie par :*

$$e_n(u) = \frac{1}{\bar{F}_n(u)} \int_u^{\infty} \bar{F}_n(t)\, dt = \frac{1}{N_u} \sum_{i=1}^{n} (X_i - u)\, \mathbb{1}_{\{X_i > u\}}, \text{ avec } X_{1,n} < u < X_{n,n}. \tag{2.3}$$

*où* $N_u = card\{i : i = 1, \cdots, n\ ;\ X_i > u\}$ *le nombre des observations qui excèdent* $u$ *et* $\frac{0}{0} = 0$ *(conventionnellement).*

## 2.2 Distribution de Pareto Généralisée

**Définition 2.2.1 (Distribution de Pareto Généralisée)**

*La fonction de répartition de la loi de Pareto généralisée standard est définie pour* $\gamma \in \mathbb{R}$ *par*

$$G_\gamma(x) = \begin{cases} 1 - (1 + \gamma x)^{-1/\gamma} & si \quad \gamma \neq 0, \\ 1 - \exp(-x) & si \quad \gamma = 0, \end{cases} \tag{2.4}$$

*où* $x \in \mathbb{R}_+$ *si* $\gamma \geq 0$*, et* $x \in [0, -1/\gamma[$ *si* $\gamma < 0$*.*

La $GPD$ standard peut être étendue à une famille plus générale, en remplaçant l'argument $x$ par $(x - \mu)/\sigma$, où $\mu \in \mathbb{R}$ et $\sigma > 0$ sont les paramètres de localisation et d'échelle (respectivement). La $GPD$ $G_{\gamma,\mu,\sigma}(x)$ est en fait $G_\gamma\left(\frac{x-\mu}{\sigma}\right)$. La $GPD$ avec un paramètre de localisation nulle et un paramètre d'échelle $\sigma > 0$, joue un rôle important dans la modélisation des valeurs extrêmes. Pour économiser la notation, nous dénoterons [9]

$$G_{\gamma,\sigma}(x) = \begin{cases} 1 - \left(1 + \frac{\gamma x}{\sigma}\right)^{-1/\gamma} & si \quad \gamma \neq 0, \\ 1 - \exp(-x/\sigma) & si \quad \gamma = 0, \end{cases} \tag{2.5}$$

où

$$x \in D(\gamma, \sigma) = \begin{cases} [0, \infty) & si \quad \gamma \geq 0, \\ [0, -\sigma/\gamma] & si \quad \gamma < 0. \end{cases}$$

**Définition 2.2.2 (Densité de la loi de Pareto Généralisée)**

*La dérivée de la fonction de répartition $G_{\gamma,\sigma}$ donne la fonction de densité de probabilité suivante :*

$$g_{\gamma,\sigma}(x) = \begin{cases} \sigma^{-1}\left(1 + \gamma\frac{x}{\sigma}\right)^{-1/\gamma - 1} & si \quad \gamma \neq 0, \\ \sigma^{-1}\exp(-x/\sigma) & si \quad \gamma = 0, \end{cases} \tag{2.6}$$

*où $x \in D(\gamma, \sigma)$.*

La figure 2.2 ci-dessous, illustre la forme de densité et la fonction de répartition de lois de Pareto Généralisées pour différentes valeurs de $\gamma$.

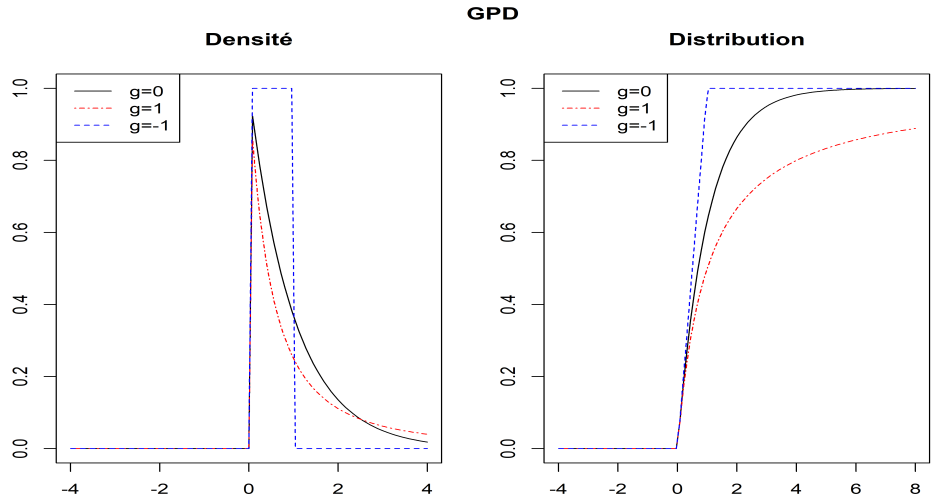


FIG. 2.2 – Densité et fonction de répartition de de lois de Pareto Généralisées

Les travaux de Balkema et de Haan (1974) [9] et Pickands (1975) [25] ont abouti au théorème suivant, qui joue un rôle fondamental dans l'analyse des valeurs extrêmes. Ce théorème assure que la loi des excès pour un seuil élevé (proche du point terminal) peut-être approchée par une $GPD$.

**Théoreme 2.2.1 (Balkema-de Haan-Pickands)**

*Une fonction de répartition $F$ appartient au domaine d'attraction de loi des valeurs extrêmes $H_\gamma$, si et seulement s'il existe une fonction positive $\sigma(u)$ et un réel $\gamma$, telle que : [9]*

$$\lim_{u \to x_F} \sup_{0 \leq y \leq x_F - u} \left| F_u(y) - G_{\gamma,\sigma(u)}(y) \right| = 0. \tag{2.7}$$

Si $y \in [0, \infty)$ pour $\gamma \geq 0$ et $y \in [0, -\sigma(u)/\gamma]$ pour $\gamma < 0$, alors selon ce théorème, si $F$ vérifie le théorème de Fisher et Tippet, il existe une fonction positive $\sigma(u)$ et un réel $\gamma$ telle que la loi des excès $F_u$ peut être uniformément approximée par une distribution $GPD$. Ainsi, l'indice des valeurs extrêmes, donné dans le théorème de Fisher et Tippet, est le même que celui de la loi des excès. La preuve de ce théorème doit être trouvé dans [13].

**Remarque 2.2.1** *Selon le signe de $\gamma$, nous avons les cas suivants : [26]*

1. *$\gamma$>0 : distribution de type Pareto usuelle,*
2. *$\gamma$<0 : distribution de type Beta bornée : (loi de Pareto de type II)*
3. *$\gamma = 0$ : distribution de type exponentielle*

**Exemple 2.2.1**

*Pour le cas de la loi de Pareto, la fonction de répartition est donnée par :*

$$F(x) = 1 - \left(\frac{c}{x}\right)^{\alpha}, \ \textit{avec} \ \ c > 0 \ \textit{et} \ \alpha > 0.$$

*Alors*

$$F_u(y) = 1 - \frac{\bar{F}(u+y)}{\bar{F}(u)} = 1 - \left(1 + \frac{1}{u}y\right)^{-\alpha}.$$

*Ceci correspond à $\gamma = 1/\alpha$ et $\sigma(u) = u\gamma$ dans (2.5).*

## 2.3 Sélection du seuil

Le choix du seuil est crucial dans la modélisation des excès. Un seuil trop élevé réduit la taille de l'échantillon, ce qui augmente la variance des estimateurs des paramètres de la loi des excès. À l'inverse, un seuil trop bas inclut des valeurs non extrêmes, entraînant un biais dans l'estimation. Il faut donc un compromis : le seuil doit être suffisamment grand pour que l'approximation asymptotique soit valable, sans être excessif. Plusieurs méthodes ont été proposées dans la littérature pour estimer le seuil $u$. Parmi celles-ci, nous présentons les méthodes graphiques, qui sont les plus couramment utilisées dans la pratique.

### MRL-plot

Le MRL-plot (Mean Residual Life plot) ou le graphe de la durée de vie moyenne résiduelle, est un outil diagnostique introduit par Davision et Smith [8], utilisé pour évaluer la validité du choix du seuil $u$, tout en garantissant l'existence de l'espérance [4]. Le MRL-plot porte aussi le nom mef-plot (mean excess function plot) ou le graphe de la fonction moyenne des excès. Ce graphe repose sur la fonction moyenne des excès $e(u)$ (2.2).

Sous l'hypothèse que les excès au-delà d'un seuil élevé suivent une $GPD$ avec paramètre de forme $\gamma \leq 1$, cette espérance devient :

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

Pour un seuil $u$ suffisamment grand ($u > u^*$), la fonction $e(u)$ devient linéaire en $u$

$$e(u) = E(X - u \mid X > y) = \frac{\sigma(u^*)}{1-\gamma} + \frac{\gamma}{1-\gamma}u, \ \gamma \leq 1.$$

oú $\sigma(u^*)$ est le paramètre d'échelle associé au seuil $u$.

Supposons que nous disposions d'un ensemble d'observations $X_1, X_2, \cdots, X_n$. La fonction

moyenne des excès empiriques sous la transformation affine s'écrit, pour $x \geq u$ , comme suit :

$$e_n(u) = \hat{e}(u) = \frac{\hat{\sigma}(u)}{1-\gamma} + \frac{\hat{\gamma}}{1-\gamma}u,\ \hat{\sigma}(u) + \hat{\gamma}u > 0.$$

Nous traçons alors graphiquement la fonction empirique des excès $\widehat{e}(u)$ en fonction de $u$ et sélectionnons le plus petit seuil $u$ à partir duquel $\widehat{e}(u)$ devient approximativement linéaire pour tout $x > u$.

**Définition 2.3.1 (MRL-plot)**

*Le MRL-plot est le tracé du nuage de point*

$$\{(u, e_n(u)),\ X_{1,n} < u < X_{n,n}\}.$$

*En observant ce graphe, on cherche la région où la courbe devient approximativement linéaire, cela suggère un seuil $u^*$ à partir duquel la GPD fournit une bonne approximation des excès [13].*

Trois cas peuvent alors se présenter :

1. Si la fonction moyenne des excès empirique présente une pente positive à partir d'un certain seuil, alors les données suivent une $GPD$ avec un paramètre $\gamma$ positif.
2. Si la pente est horizontale, cela suggère une loi exponentielle.
3. Si la pente est négative, la distribution est bornée à droite, indiquant une distribution à queue légère.

## tc-plot

Le tc-plot (Threshold Choice Plot) ou le graphe de stabilité des paramètres d'échelle et de forme, est un outil graphiques largement utilisés pour la sélection du seuil dans l'approche $POT$ [28]. Il permet de déterminer un seuil optimal en ajustant les données à une distribution $GPD$ pour différents seuils. L'idée est d'analyser la stabilité des paramètres

estimés, à savoir le paramètre de forme $\gamma$ et le paramètre d'échelle $\sigma$, en fonction du seuil choisi $u$.

Cette technique est implémentée dans le logiciel R via des packages spécialisés. Ces derniers offrent des outils objectifs permettant de guider le choix du seuil optimal, en examinant la stabilité des paramètres estimés lorsque le seuil varie. Le graphique associé montre l'évolution des estimations de $\gamma$ et $\sigma$ en fonction des seuils candidats $u^*$. Une plage de seuils où les paramètres restent relativement constants indique la validité du modèle $GPD$. Ainsi, les seuils au-dessus desquels cette stabilité est observée sont ceux pour lesquels l'ajustement du modèle $GPD$ devient approprié.

## 2.4 Estimation des paramètres de la GPD

### 2.4.1 Méthode du maximum de vraisemblance

L'estimation par la méthode du maximum de vraisemblance fournit des estimateurs asymptotiquement efficaces sous certaines conditions régulières.[29]. Supposons que notre échantillon des excès $(X_1, X_2, \cdots, X_{N_u})$ est $i.i.d.$ suivant une densité $g_{_{\gamma,\sigma}}$, donnée dans l'équation (2.6).

L'expression de la fonction de vraisemblance est donnée par :

$$L_{\gamma,\sigma}\left(x_1, \cdots, x_{N_u}\right) = \prod_{i=1}^{N_u} g_{\gamma,\sigma}\left(x_i\right),$$

où $(x_1, x_2, \cdots, x_{N_u})$ sont les réalisations de $(X_1, X_2, \cdots, X_{N_u})$.

La fonction log-vraisemblance, dans le cas où $\gamma \neq 0$, s'exprime comme suit :

$$l_{\gamma,\sigma}\left(x_1, \cdots, x_{N_u}\right) = \log L_{\gamma,\sigma}\left(x_1, \cdots, x_{N_u}\right) = -N_u \log \sigma - \left(\frac{1}{\gamma} + 1\right) \sum_{i=1}^{N_u} \log\left(1 + \frac{\gamma}{\sigma} x_i\right). \tag{2.8}$$

L'annulation des dérivées partielles de cette fonction par rapport à $\gamma$ et $\sigma$ (respectivement),

conduit au système d'équations de maximisation, à partir desquelles nous calculons les estimateurs du Maximum de vraisemblance $(\hat{\gamma}_{N_u}, \hat{\sigma}_{N_u})$ :

$$\begin{cases} \dfrac{1}{N_u}\sum_{i=1}^{N_u} \log\left(1+\gamma\dfrac{x_i}{\sigma}\right) = \gamma, \\ \dfrac{1}{N_u}\sum_{i=1}^{N_u} \dfrac{x_i/\sigma}{1+\gamma x_i/\sigma} = \dfrac{1}{1+\gamma}, \end{cases} \tag{2.9}$$

Pour $\gamma = 0$, la fonction log-vraisemblance égale à

$$l_{\gamma,\sigma}(x_1, \cdots, x_{N_u}) = -N_u \log\sigma - \frac{1}{\sigma}\sum_{i=1}^{N_u} x_i. \tag{2.10}$$

En dérivant cette fonction, par rapport à $\sigma$, on obtient

$$\hat{\sigma} = \frac{1}{N_u}\sum_{i=1}^{N_u} x_i = \bar{X}. \tag{2.11}$$

Cette méthode présente l'avantage de fournir des estimateurs ayant de bonnes propriétés asymptotiques. Toutefois, lorsque $\gamma \neq 0$ les estimateurs ne sont pas explicites, car ils résultent de la résolution d'un système d'équations non linéaires. Ce dernier néanmoins résolu par des algorithmes numérique.

Lorsque $\gamma > -1/2$ ; Smith en 1987 [29] a montré la normalité asymptotique des estimateurs du maximum de vraisemblance

$$\sqrt{N_u}\begin{pmatrix} \hat{\gamma}_{N_u} - \gamma \\ \hat{\sigma}_{N_u}/\sigma_{N_u} - 1 \end{pmatrix} \xrightarrow{d} \mathcal{N}_2\left(0, (1+\gamma)\begin{pmatrix} 1+\gamma & -1 \\ -1 & 2 \end{pmatrix}\right) \text{ quand } N_u \to \infty,$$

où $\mathcal{N}_2(\mu, \sum)$ désigne la distribution normale bivariée de vecteur moyen $\mu$ et de matrice de covariance $\sum$. Ce résultat asymptotique est fondamental car il permet notamment de construire des intervalles de confiance pour les estimateurs du maximum de vraisemblance.

### 2.4.2 Méthode des moments de probabilités pondérés

Dans certains cas, les moments classiques peuvent ne pas exister ou être mal définis, notamment lorsqu'on travaille avec des distributions à queue lourde comme la $GPD$. Pour pallier cette difficulté, Hosking et Wallis en (1987) [18] ont notamment proposé une approche fondée sur les moments de probabilités pondérés pour l'estimation des paramètres de la $GPD$.

$$\varpi_r = \mathbb{E}\left[X\bar{G}^r_{\gamma,\sigma}(X)\right] = \int_0^1 x\bar{G}^r_{\gamma,\sigma}dG_{\gamma,\sigma}(x) = \frac{\sigma}{(r+1)(r+1-\gamma)},\ r \in \mathbb{N}. \tag{2.12}$$

avec $X$ est une va de fonction de distribution $G_{\gamma,\sigma}$. Pour $r = 0, 1$, on obtient

$$\gamma = 2 - \frac{\varpi_0}{\varpi_0 - 2\varpi_1} \text{ et } \sigma = \frac{2\varpi_0\varpi_1}{\varpi_0 - 2\varpi_1}. \tag{2.13}$$

Pour calculer les estimateurs des moments de probabilités pondérés $\hat{\gamma}$ et $\hat{\sigma}$ des paramètres de la $GPD$, on remplace $\omega_0$ et $\omega_1$ par leurs estimateurs empirique, on obtient

$$\hat{\gamma} = 2 - \frac{\hat{\varpi}_0}{\hat{\varpi}_0 - 2\hat{\varpi}_1} \text{ et } \hat{\sigma} = \frac{2\varpi_0\hat{\varpi}_1}{\hat{\varpi}_0 - 2\hat{\varpi}_1}. \tag{2.14}$$

## 2.5 Estimation de la queue de la distribution

Une fois les paramètres de la $GPD$ sont estimés par l'une des méthodes ci-dessus. Une telle formulation est donnée par l'égalité suivante

$$\bar{F}(x) = \bar{F}_u(x-u)\,\bar{F}(u),\ u < x < x_F. \tag{2.15}$$

La queue conditionnelle $\bar{F}_u$ de $F$ peut être estimée par

$$\hat{\bar{F}}_u(x-u) = \bar{G}_{\hat{\gamma}(u),\hat{\sigma}(u)}(x-u) = \left(1 + \hat{\gamma}(u)\frac{x-u}{\hat{\sigma}(u)}\right)^{-1/\hat{\gamma}(u)},\ u < x < x_F, \tag{2.16}$$

ainsi que $\bar{F}(u)$ est estimée par la probabilité empirique d'exceedance

$$\widehat{\bar{F}}(u) = \bar{F}_n(u) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i > u\}} = \frac{N_u}{n}, \ u < x_F. \tag{2.17}$$

L'estimateur de la queue de la distribution est donc

$$\widehat{\bar{F}}(x) := \frac{N_u}{n}\left(1 + \hat{\gamma}(u)\frac{x-u}{\hat{\sigma}(u)}\right)^{-1/\hat{\gamma}(u)}, \ u < x < x_F. \tag{2.18}$$

## 2.6 Estimation des quantiles extrêmes

L'estimateur des quantiles aux ordres élevés au-dessus du seuil $u$ ($x_p > u$) est obtenu en inversant l'expression de l'estimateur de la queue de la distribution (2.18)

$$\hat{x}_p := u + \frac{\hat{\sigma}(u)}{\hat{\gamma}(u)}\left(\left(\frac{N_u}{np}\right)^{\hat{\gamma}(u)} - 1\right), \ p < \frac{N_u}{n}, \tag{2.19}$$

avec $\hat{\sigma}_u$ et $\hat{\gamma}_u$, les estimateurs des paramètres de la loi GPD et $N_u$, le nombre d'excès.

Cette expression figure, dans Davison et Smith (1990) [8] et Embrechts et al. (1997) [13].

Le seuil $u$ est souvent choisi égal à une des statistiques d'ordre $X_{1,n} \leq X_{2,n} \leq \cdots \leq X_{n,n}$. Si l'on choisit comme seuil $u = X_{n-k,n}$ la $(k+1)$-ième plus grande observation, alors $N_u = k$ et l'estimateur des quantiles aux ordres élevés se réécrit de la manière suivante

$$\hat{x}_p^{(POT)} = X_{n-k,n} + \frac{\hat{\sigma}^{(POT)}}{\hat{\gamma}^{(POT)}}\left(\left(\frac{k}{np}\right)^{\hat{\gamma}^{(POT)}} - 1\right), \ p < \frac{k}{n}, \tag{2.20}$$

où $\hat{\gamma}^{(POT)}$, $\hat{\sigma}^{(POT)}$ sont les estimateurs résultants de $\gamma$ et $\sigma$ (respectivement).

# Chapitre 3

# Application sous R

Dans ce chapitre, nous présentons une étude pratique fondée sur des données réelles issues d'un contexte hydrologique, extraites à partir du package `ismev` du logiciel statistique R (version 4.5). Cette série est utilisée afin d'appliquer les deux approches principales de la modélisation des valeurs extrêmes, présentées dans les chapitres précédents. Il s'agit, d'une part, de l'approche des maxima par blocs, en ajustant la distribution $GEV$ à des maxima annuels extraits de la série, et d'autre part, de l'approche $POT$, qui vise à modéliser les observations excédant un certain seuil à l'aide de la $GPD$. L'ensemble des résultats obtenus repose sur l'utilisation des packages `fBasics, extRemes`, `evd`, `POT`, `evir`,, `dplyr`, `ggplot2`, `fExtremes`, `lmome`, `lmomc, nortest et ismev` du logiciel R.

## 3.1 Description des données

On considère la série "`rain`" qui représente des précipitations journalières (en $mm$ : millimètres) enregistrées dans une station météorologique située dans le sud-ouest de l'Angleterre. La période d'observation s'étale de 1914 à 1962, couvrant ainsi près de 48 années.

Dans un premier temps, nous effectuons une analyse statistique descriptive de notre série de données.

Les résultats obtenus à l'aide du logiciel R sont résumés dans le tableau 3.1 ci-dessous.

| | |
|---|---|
| Minimum | 0 $(mm)$ |
| $Qu_1$ (premier quartile) | 0 $(mm)$ |
| Médiane | 0.5 $(mm)$ |
| Moyenne | 3.4761 $(mm)$ |
| $Qu_3$ (troisième quartile) | 4.3 $(mm)$ |
| Maximum | 86.6 $(mm)$ |
| Variance | 39.997 |
| Skewness | 3.2895 |
| Kurtosis | 17.1373 |

TAB. 3.1 – Caractéristiques statistiques des données

D'après ce tableau ainsi que la figure 3.1, l'ensemble des précipitations journalières (de taille $n = 17531$ observation) varie entre une valeur minimale de 0 $(mm)$, correspondant à de nombreux jours secs, jusqu'à une valeur maximale de 86.6 $(mm)$, enregistrée le 04 Juin 1929. La moyenne des précipitations journalières est de 3.4761 $(mm)$, ce qui traduit une distribution dominée par des précipitations faibles à modérées. Cette faible moyenne reflète la fréquence élevée des jours secs, caractérisés par des valeurs nulles comme le montre l'histogramme des données 3.2. La variance élevée des précipitations journalières traduit une forte fluctuation entre sécheresse et pluies intenses.

Le coefficient d'asymétrie (skewness) indique une distribution asymétrique, liée à la survenue d'événements pluvieux rares mais intenses. Enfin, le coefficient d'aplatissement (kurtosis), avec une valeur élevée de 17.1373, souligne une forte concentration des données dans les queues de la distribution, confirme la présence d'événements extrêmes.

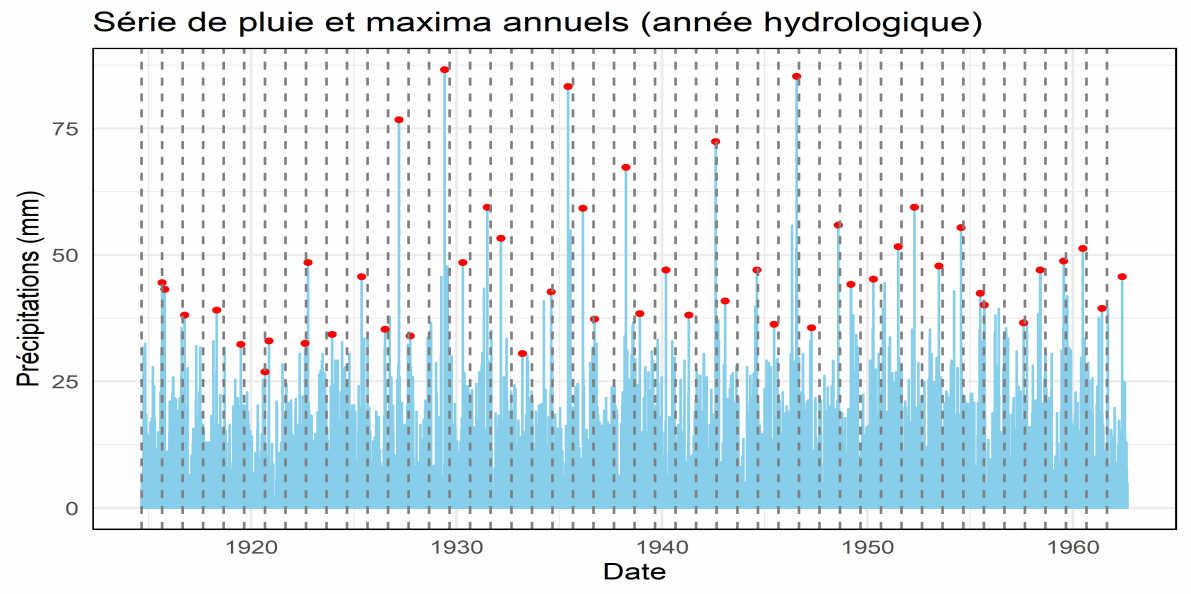


FIG. 3.1 – Série des précipitations journalières en Angleterre couvrant la période du 1914 au 1962

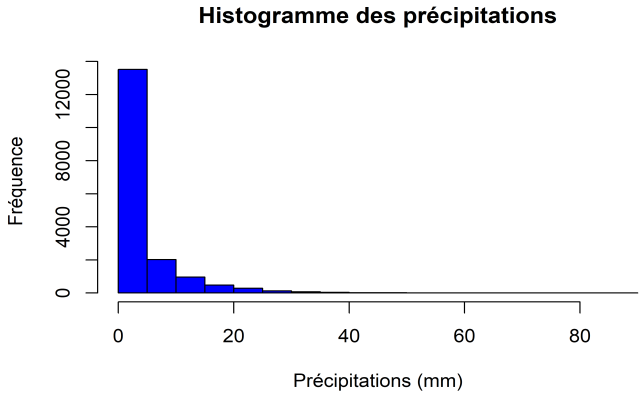


FIG. 3.2 – Histogramme de la série des précipitations journalières

### 3.1.1 Ajustement par une loi normale

Pour vérifier si les données suivent une loi normale, nous avons appliqués le test d'Anderson-Darling, en posant les hypothèses suivantes :

$$\begin{cases} H_0 : \text{L'échantillon suit une loi normale.} \\ H_1 : \text{L'échantillon ne suit pas une loi normale.} \end{cases}$$

La p-value obtenue est inférieure à $2.2 \times 10^{-16}$, ce qui est inférieur à tout seuil $\alpha$ courant $(0.05, 0.01, 0.001, \cdots)$, nous rejetons donc l'hypothèse nulle avec une très forte certitude. Ainsi, nous concluons que l'échantillon ne suit pas une loi normale.

## 3.2 Modélisation via une distribution *GEV*

### 3.2.1 Série des maxima annuels

Avant de pouvoir modéliser les valeurs extrêmes à l'aide d'une loi de probabilité, il est nécessaire d'extraire les maxima annuels à partir des données. Pour ce faire, l'approche des maxima par blocs est appliquée aux données présentées dans la section 3.1. Elle consiste à diviser la série en blocs annuels, définis selon l'année hydrologique (du $1^{er}$ septembre au 31 août). Comme le montre la figure 3.2, on retient, pour chaque bloc, la valeur maximale de précipitations journalières, représentant l'événement extrême de cette période. La figure 3.3 illustre la série des maxima annuels ainsi obtenue à partir les données initiales.

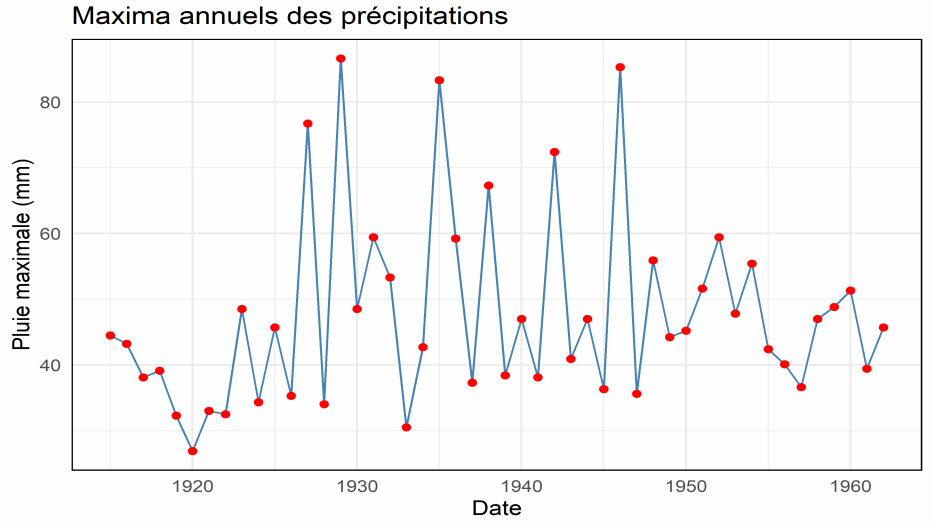


FIG. 3.3 – Série des maxima annuels des précipitations journalières

L'analyse statistique de la série des maxima annuels, de taille 48 observations, révèle une variabilité notable, avec des valeurs comprises entre un minimum de 26.9 ($mm$) et un maximum de 86.6 ($mm$). La moyenne s'établit à 47.58 ($mm$), légèrement supérieure à la médiane (44.85 $mm$), ce qui indique une légère asymétrie à droite confirmée par un coefficient de skewness de 1.21. Cette asymétrie suggère la présence de quelques années avec des précipitations fortes. La variance de 206.26 confirme une dispersion significative autour de la moyenne.

L'étape suivante consiste à ajuster une distribution $GEV$ sur ces maxima, dans le but d'estimer les paramètres $\mu$, $\sigma$ et $\gamma$.

### 3.2.2 Estimation des paramètres

#### Estimation paramétrique

Deux méthodes sont utilisées pour estimer les paramètres $\mu$, $\sigma$ et $\gamma$ de la distribution $GEV$, à savoir la méthode de maximum de vraisemblance ($MLE$) et la méthode des moments de probabilités pondérés ($PWM$).

Le tableau 3.2 présente les estimateurs obtenus pour chaque paramètre ainsi que leurs intervalles de confiance à 95%.

| | **MLE** | **PWM** |
|---|---|---|
| $\hat{\mu}$ | 40.6847044<br>$[37.678825, 43.69058]$ | 40.5279960<br>$[37.65730, 43.74861]$ |
| $\hat{\sigma}$ | 9.3664138<br>$[7.046682, 11.6861451]$ | 9.4364470<br>$[7.092485, 12.189706]$ |
| $\hat{\gamma}$ | 0.1429752<br>$[-0.087720, 0.37367]$ | 0.1483429<br>$[-0.263317, 0.0311443]$ |

TAB. 3.2 – Estimateurs des paramètres de la distribution GEV selon les méthodes MLE et PWM, avec intervalles de confiance à 95

On remarque que les deux méthodes donnent des résultats similaires, ce qui indique une bonne stabilité des estimations. Le paramètre de forme $\gamma$ est positif dans les deux cas, cela suggère que la distribution des maxima suit une loi de type Fréchet, caractérisée par une queue lourde. Cependant, les intervalles de confiance associés à ce paramètre incluent également des valeurs négatives ainsi que la valeur nulle. Cette observation indique que l'incertitude autour de l'estimation du paramètre de forme reste importante, ce qui ne permet pas de conclure de manière définitive sur le type exact de la distribution. Elle pourrait aussi correspondre à une distribution de type Gumbel ($\gamma = 0$) ou même Weibull ($\gamma < 0$).

En résumé, les deux méthodes suggèrent une tendance vers des événements extrêmes avec une queue lourde, mais les intervalles de confiance appellent à la prudence dans l'interprétation de la forme exacte de la distribution.

#### Estimation semi-paramétrique

Des estimateurs semi-paramétriques tels que celles de Hill et Pickands ont été utilisées pour estimer l'indice des valeurs extrêmes $\gamma$. Selon les résultats obtenus sous R, les estimateurs de Hill $\hat{\gamma}^H = 0.1352498$ et de Pickands $\hat{\gamma}^P = 0.6664439$ sont tous deux strictement positives, ce qui confirme une distribution à queue lourde de type Fréchet, comme suggéré par les méthodes d'estimation paramétriques.

### 3.2.3 Validation du modèle ajusté

La validation du modèle ajusté repose sur plusieurs outils graphiques permettant d'évaluer la qualité de l'ajustement de la distribution $GEV$ aux données observées. La figure 3.4 ci-dessous présente :

**1. Le QQ-plot :** Les points suivent globalement la diagonale, montrant que les quantiles empiriques sont proches de ceux du modèle, notamment pour les valeurs extrêmes.

**2. La densité ajustée :** La courbe de la densité empirique (ligne continue) est presque superposée sur la densité théorique *GEV* (ligne discontinue), ce qui indique un bon ajustement global du modèle aux données.

**3. Courbe des niveaux de retour :** La courbe montre une croissance régulière des quantiles extrêmes avec la période de retour, accompagnée de bandes de confiance raisonnables, confirmant la cohérence du modèle pour prédire les événements rares.

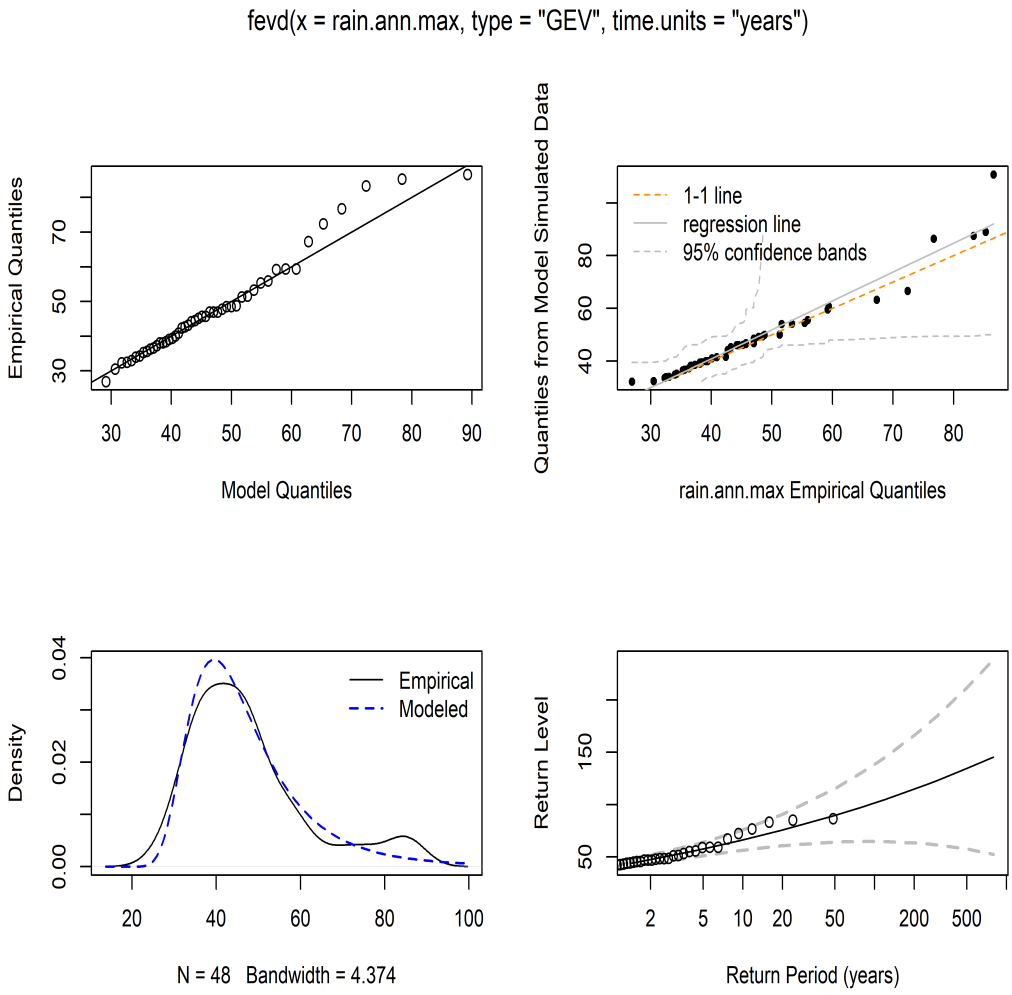


FIG. 3.4 – Graphiques de diagnostic de l'ajustement avec la distribution *GEV*

Ces diagnostics visuels indiquent que le modèle *GEV* fournit un ajustement satisfaisant aux données observées.

### 3.2.4 Estimation des quantiles extrêmes

L'estimation des quantiles extrêmes (ou niveaux de retour) constitue un outil essentiel pour anticiper les risques hydrologiques majeurs. Le tableau 3.3 regroupe les quantiles extrêmes estimés pour différentes périodes de retour (10, 20, 50 et 100 ans), à l'aide des deux méthodes d'estimation $MLE$ et $PWM$. Les intervalles de confiance à 95% sont également indiqués pour chaque valeur estimée.

| **Périodes** | **MLE** | **PWM** |
|---|---|---|
| 10 | 65.54301<br>[56.67333, 74.41268] | 65.74069<br>[57.80253, 73.48204] |
| 20 | 74.81010<br>[61.31762, 88.30259] | 75.74051<br>[64.47500, 85.23318] |
| 50 | 87.91828<br>[65.42912, 110.40745] | 90.36748<br>[73.14026, 102.97961] |
| 100 | 98.63615<br>[66.85346, 130.41884] | 102.72576<br>[79.70053, 118.83180] |

TAB. 3.3 – Niveau de retour pour différentes période de retour et intervalles de confiance (GEV)

On remarque que les quantiles extrêmes augmentent pour des périodes de retour élevées, ce qui correspond au comportement attendu des événements rares : plus un événement est rare, plus son intensité estimée est élevée. De plus, les résultats obtenus par les deux méthodes sont très proches, ce qui confirme la cohérence des estimations. En revanche, les intervalles de confiance s'élargissent avec la période de retour, ce qui indique une incertitude croissante dans l'estimation des événements très rares. Cela souligne l'importance de rester prudent dans l'interprétation des valeurs extrêmes.

## 3.3 Modélisation via une $GPD$

### 3.3.1 Sélection du seuil

La première étape de l'approche $POT$ consiste à déterminer un seuil optimal pour ajuster la distribution $GPD$. Dans cette section, la sélection du seuil $u$ est effectuée à l'aide des

deux outils graphiques abordés dans la section 2.3 du chapitre 2, à savoir le MRL-plot et le tc-plot.

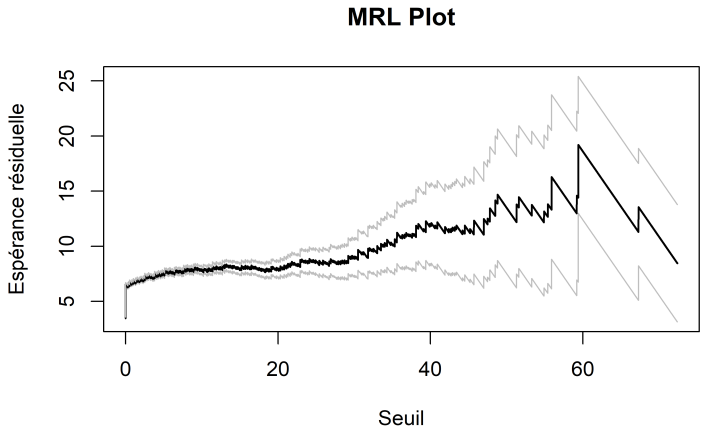


FIG. 3.5 – Résultats graphiques de la sélection du seuil appliquée aux précipitations journalières en Angleterre (MRL-plot)

Le graphique MRL-plot 3.1 montre un comportement approximativement linéaire à partir du seuil $u = 30\ mm$, ce qui nous permet de considérer cette valeur comme un seuil optimal pour la $GPD$. De plus la pente positive observée indique un paramètre de forme $\gamma$ positif.

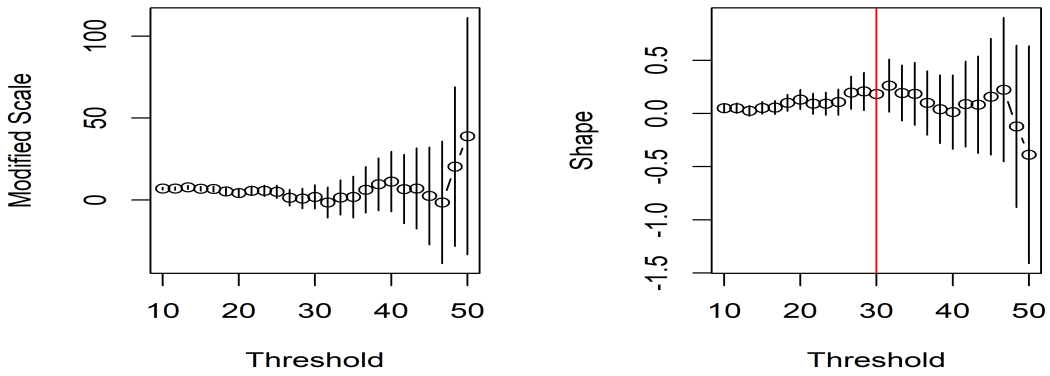


FIG. 3.6 – Résultats graphiques de la sélection du seuil appliquée aux précipitations journalières en Angleterre (TC-plot) paramètre de forme (à gauche) et d'échelle (à droite) pour la $GPD$

Dans le graphique de stabilité des paramètres (tc-plot) obtenu, nous observons qu'à partir du seuil $u = 30$ $(mm)$, une ligne horizontale peut traverser l'ensemble des barres de confiance. Cela suggère que ce seuil constitue un choix approprié, car au-delà de cette valeur, les estimations des paramètres restent stables.

En résumé, les deux figures suggère un seuil optimal de $u = 30$ $(mm)$, à partir duquel une série des excès est extraite.

### 3.3.2 Série des excès

La figure 3.7, illustre la série des excès $Y = X - u$ des précipitations journalières au-delà du seuil de 30 $(mm)$, de taille 152 observations.

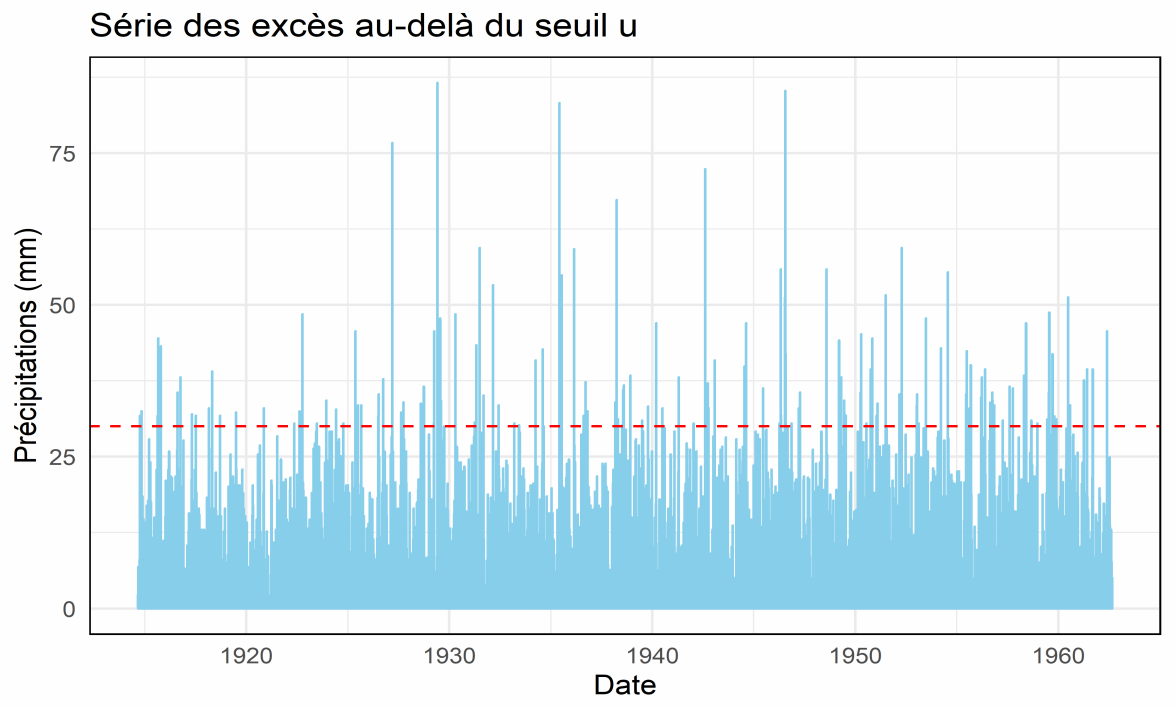


FIG. 3.7 – Série des excès au-delà du seuil $u = 30$ $mm$ (ligne horizontale), extraite à partir des précipitations journalières observées en Angleterre

Les caractéristiques statistiques de cette série sont résumés dans le tableau 3.4.

| | |
|---|---|
| Minimum | 0.2 $(mm)$ |
| $Qu_1$ (premier quartile) | 2 $(mm)$ |
| Médiane | 5.3 $(mm)$ |
| Moyenne | 9.08 $(mm)$ |
| $Qu_3$ (troisième quartile) | 12.03 $(mm)$ |
| Maximum | 56.6 $(mm)$ |
| Variance | 115.48 |
| Skewness | 2.32 |
| Kurtosis | 6.14 |

TAB. 3.4 – Caractéristiques statistiques de la série des excès

L'analyse descriptive montre que ces excès sont généralement faibles, avec un minimum de 0.2 $(mm)$, indiquant que certaines précipitations dépassent à peine le seuil (par exemple, 30.2 $mm$). La moyenne est supérieure la médiane, ce qui suggère une distribution asymétrique à droite, amplifiée par quelques excès très importants comme le maximum observé de 56.6 $(mm)$ (correspondant à une pluie totale de 86.6 $mm$ en une journée). La variance élevée traduit une forte dispersion autour de la moyenne, tandis que les coefficients de skewness et de kurtosis indiquent une distribution très asymétrique avec une queue lourde. Ces caractéristiques justifient le recours à des modèles de valeurs extrêmes, notamment la $GPD$, pour modéliser ces données.

### 3.3.3 Estimation des paramètres

De même que pour la distribution $GEV$, les paramètres $\sigma$ et $\gamma$ de La $GPD$ sont estimés ici par les deux méthodes $MLE$ et $PWM$. Les résultats obtenus sont présentés dans le tableau 3.5, ci-dessous :

L'ajout de nouvelles données conduit à une modification de l'estimation du paramètre de forme $\gamma$ par rapport à celle obtenue pour la distribution $GEV$. Cependant, ce paramètre reste strictement positif, suggérant une distribution à queue lourde. Toutefois, l'intervalle de confiance à 95%, associé à ce paramètre inclut la valeur nulle, ce qui n'exclut pas le choix d'une distribution exponentielle.

| | **MLE** | **PWM** |
|---|---|---|
| $\hat{\sigma}$ | 7.440252<br>$[5.56158139, 9.3189230]$ | 7.3486370<br>$[5.85957, 9.516246]$ |
| $\hat{\gamma}$ | 0.184498<br>$[-0.01385379, 0.3828497]$ | 0.1910539<br>$[-0.002733531, 0.3151875]$ |

TAB. 3.5 – Estimateurs des paramètres de la GPD selon les méthodes MLE et PWM, avec intervalles de confiance à 95

### 3.3.4 Validation du modèle ajusté

Comme pour l'ajustement avec la distribution $GEV$, des graphiques de diagnostic ont été utilisés pour évaluer la qualité de l'ajustement 3.8. Ces derniers confirment que la $GPD$ s'ajuste bien aux données. Cela valide son utilisation pour modéliser la série des excès au-delà du seuil $u$.

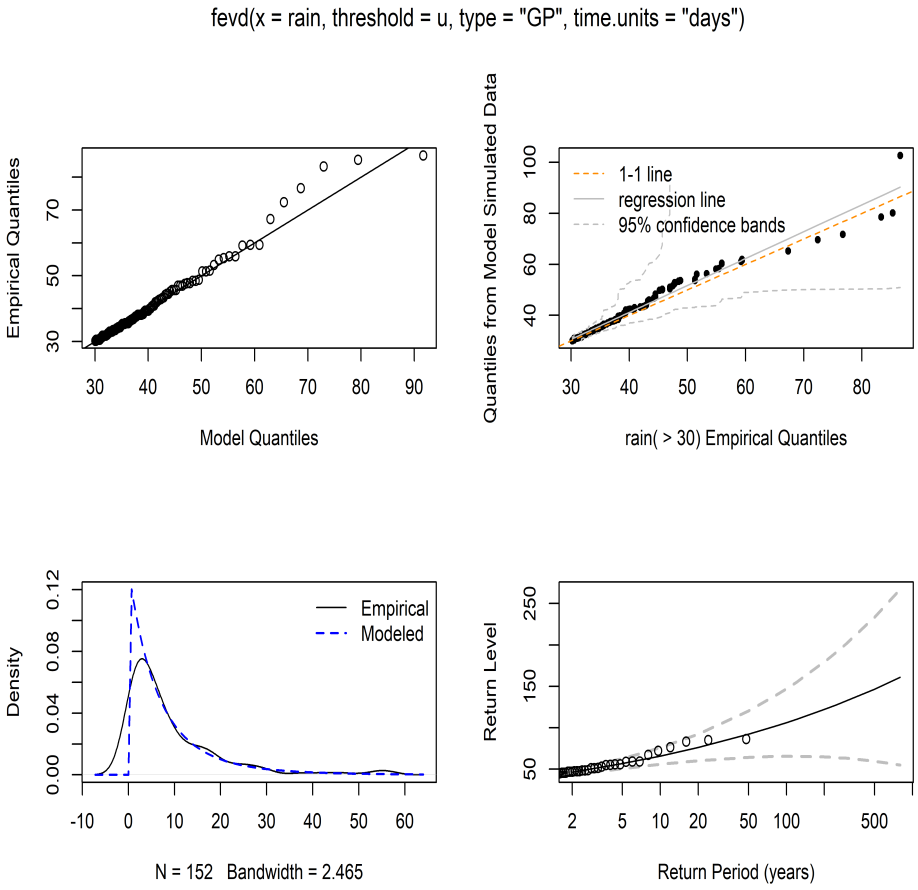


FIG. 3.8 – Graphiques de diagnostic de l'ajustement avec la $GPD$

### 3.3.5 Estimation des quantiles extrêmes

Dans le cadre de l'ajustement de la $GPD$, le tableau 3.6 présente les niveaux de retour estimés pour différentes périodes (10, 20, 50 et 100 ans), obtenus à l'aide des deux méthodes $MLE$ et $PWM$. Les intervalles de confiance à 95% sont également indiqués pour chaque estimation, permettant d'apprécier l'incertitude associée à ces valeurs.

| **Périodes** | **MLE** | **PWM** |
|---|---|---|
| 10 | 65.96142<br>[55.89296, 76.02989] | 65.74069<br>[57.80253, 73.48204] |
| 20 | 76.36880<br>[60.26535, 92.47226] | 75.74051<br>[64.47500, 85.23318] |
| 50 | 92.33677<br>[64.32393, 120.34961] | 90.36748<br>$C$[73.14026, 102.97961] |
| 100 | 106.34231<br>[65.62237, 147.06225] | 102.72576<br>[79.70053, 118.83180] |

TAB. 3.6 – Niveau de retour pour différentes période de retour et intervalles de confiance (GPD)

Les résultats montrent une croissance des excès attendus avec la période de retour : environ 66 $(mm)$ pour 10 ans, 76 $(mm)$ pour 20 ans jusqu'à 90 $(mm)$ à 92 $(mm)$ pour 50 ans. Pour 100 ans, les estimations dépassent 100 $(mm)$, ce qui dépasse le maximum observé. Cela indique une surestimation possible pour les longues périodes. Il est donc important de considérer la $GPD$ avec prudence pour des périodes de retour supérieures à 50 ans.

Les deux méthodes donnent des estimations similaires, avec des intervalles de confiance généralement large. L'élargissement progressif des intervalles de confiance reflète l'incertitude croissante liée à l'estimation des événements rares.

# Conclusion

Au cours de ce travail, nous nous sommes intéressés à la modélisation des valeurs extrêmes, dans le but de mieux comprendre et quantifier les phénomènes rares.

Dans un premier temps, nous avons présenté les distributions des valeurs extrêmes généralisées ($GEV$), qui permet de modéliser les maxima par blocs d'une série temporelle. Le deuxième chapitre a été consacré à l'approche basée sur la distribution de Pareto généralisé ($GPD$), adaptée à la modélisation des excès au-delà d'un seuil élevé, dans le cadre de l'approche $POT$. Le dernier chapitre de ce mémoire a été dédié à une application sur une série de données réelles : les précipitations journalières en Angleterre, enregistrées sur une longue période. L'ajustement des modèles $GEV$ et $GPD$ à ces données a permis d'estimer leurs paramètres, de réaliser des diagnostics de qualité d'ajustement, et d'obtenir des quantiles extrêmes (niveaux de retour), utiles pour la gestion du risque d'évènements extrêmes.

Les résultats obtenus confirment la pertinence de ces approches pour la modélisation des extrêmes hydrologique, notamment pour des périodes de retour inférieures à 50 ans. Néanmoins, leur mise en œuvre requiert une attention particulière quant au choix des paramètres (seuils, période d'observation, méthodes d'estimation), qui peuvent influencer significativement les résultats.

Ce travail ouvre ainsi des perspectives intéressantes, notamment l'extension à des modèles non stationnaires, ou encore l'intégration de covariables climatiques, afin d'obtenir une modélisation plus précise.

# Bibliographie


[1] Arnold, B.C., Balakrishan, N. and Nagaraja, H.N. (1992). A First Course in Order Statistics. Wiely, New York.

[2] Beirlant, J., Goegebeur, Y., Segers, J. and Teugels, J. (2004). Statistics of Extremes - Theory and Applications. Wiley.

[3] Bingham, N.H., Goldie, C.M. and Teugels, J.L. (1987). Regular Variation. Cambridge University Press, Cambridge.

[4] Cabanal-Duvillard, T., & Ionescu, V. (1997). Un théoreme central limite pour des variables aléatoires non-commutatives. Comptes Rendus de l'Académie des Sciences-Series IMathematics, 10, 1117-1120.

[5] Castillo, E., Hadi, A.S., Balakrishnan, N. and Sarabia, J.M. (2005). Extreme Value and Related Models with Applications in Engineering and Science. Wiley series in probability and statistics.

[6] Coles, S. (2001). An Introduction to Statistical Modelling of Extreme Values. Springer Series in Statistics.

[7] David.H.A.,Nagaraja.H.N. (2003). Order Statistics, Third Edition. Wiely.

[8] Davison, A.C. and Smith, R.L. (1990). Models for exceedances over high thresholds, Journal of the Royal Statistical Society.Series B (Methodological), **52,** N°3.

[9] de Haan, L. and Ferreira, A. (2006). Extreme value theory : an introduction. Springer Series in Operations Research and Financial Engineering, Boston.

[10] Delmas, J.-F., & Jourdain, B. (2006). Lois de valeurs extrêmes. In Modèles aléatoires : Applications aux sciences de l'ingénieur et du vivant, 303-341.

[11] Dekkers, A.L.M., Einmahl, J.H.J. and de Haan, L. (1989). A Moment Estimator for the Index of an Extreme Value Distribution. Annals of Statistics **17**, 1833-1855.

[12] Dekkers, A.L.M. and de Haan, L. (1989). On the Estimation of the Extreme Value Index and Large Quantile Estimation. Annals of Statistics **17**, 1795-1832.

[13] Embrechts, P., Kluppelberg, C. and Mikosch, T. (1997). Modelling Extremal Event for Insurance and Finance. Springer, Berlin.

[14] Fisher, R.A. and Tippett, L.H.C. (1928). Limiting Forms of the Frequency Distribution of the Largest or Smallest Member of a Sample. Proceedings of the Cambridge Philosophical Society **24**, 180-190.

[15] Gnedenko, B.V. (1943). Sur la Distribution Limite du Terme Maximum d'une Série Aléatoire. Annales de Mathématiques **44**, 423-453.

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

[17] Hosking, J.R.M.,Wallis, J.R. and Wood, E.F. (1985). Estimation of the Generalized Extreme Value Distribution by the Method of Probability-Weighted Moments. Technometrics **27**, 251-261.

[18] Hosking, J.R.M. and Wallis, J.R. (1987). Parameter and Quantile Estimation for the Generalized Pareto Distribution. Technometrics **29**, 339-349.

[19] Hill, B. (1975). A Simple General Approach to Inference About the Tail of a Distribution. Annals of Statistics **3**, 1163-1174.

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

[21] Landwehr, J.M., Matalas, N.C. and Wallis, J.R. (1979). Probability weighted moments compared with some traditional techniques in estimating Gumbel parameters and quantiles. Water Resources Research **15**, 1055–1064.

[22] Maric, V. (2000). Regular variation and differential equations. Springer Science & Business Media Vol. 1726.

[23] Meraghni, D. (2008). Modeling Distribution Tails. Doctorat thesis, Med Khider University, Biskra-Algeria.

[24] von Mises, R. (1936)"La distribution de la plus grande de n valeurs. Rev. Math. Union Interbalcanique **1**, 141-160.

[25] Pickands, J. (1975). Statistical Inference Using Extreme Order Statistics. Annals of Statistics **3**, 119-131.

[26] Reiss, R.D. and Thomas, M. (1997). Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields. Birkhäuser, Basel.

[27] Shahbaz, M. Q., Ahsanullah, M., Shahbaz, S. H., & Al-Zahrani, B. M. (2016).Ordered Random Variables : Theory and Applications. Atlantis Press.

[28] Scarrott, C. and MacDonald, A. (2012). A Review of Extreme Value Threshold Estimation and Uncertainty Quantication, REVSTATStatistical Journal, **10**, 33-60.

[29] Smith, R.L. (1987). Estimating Tails of Probability Distributions. Annals of statistics 15, 1174-1207.

[30] von Mises, R. (1954). La distribution de la plus grande de n valeurs. in (ed.), selected papers (vol. ii, pp. 271-294). providence, ri. American Mathematical Society.

# Annexe : Abréviations et Notations

| **Symbole** | | **Signification** |
|---|---|---|
| $GEV$ | : | Distribution des valeurs extrêmes généralisée. |
| $va's$ | : | Variable aléatoires. |
| $(X_1, X_2, \cdots, X_n)$ | : | Echantillons de taille $n$ de $va's$. |
| $(i.i.d)$ | : | Indépendante et identiquement distribuée. |
| $\min(X_1, X_2, \cdots, X_n)$ | : | Minimum de $X_1, \cdots, X_n$. |
| $\max(X_1, X_2, \cdots, X_n)$ | : | Maximum de $X_1, \cdots, X_n$. |
| $X_{1,n}, X_{2,n}, \cdots, X_{n,n}$ | : | Statistique d'ordre associées à $(X_1, \cdots, X_n)$. |
| $X_{k,n}$ | : | La $k^{ieme}$ statistique d'ordre. |
| $F$ | : | Fonction de répartition. |
| $\bar{F}$ | : | Fonction de survie. |
| $F^{-1}$ | : | Inverse généralisé de la fonction de répartition. |
| $F_n$ | : | Fonction de répartition empirique. |
| $F_u$ | : | Fonction de répartition des excès au-déla d'un seuil $u$ |
| $\mu$ | : | Espérance, ou moyenne d'une $va$. |
| $TCL$ | : | Théorème Centrale Limite. |
| $\mathbb{R}$ | : | Ensemble des valeurs réelles. |
| $x_F$ | : | Point terminal de la distribution $F$. |

| | | |
|---|---|---|
| $TCL$ | : | Théorème centrale limite. |
| $\Lambda$ | : | Loi Gumbel. |
| $\Phi$ | : | Loi de Fréchet. |
| $\Psi$ | : | Loi de Weibull. |
| $u$ | : | Seuil. |
| $RV_\alpha$ | : | Variation régulière à $\infty$ avec l'indice $\alpha$. |
| $L$ | : | Fonction à variation lente. |
| $POT$ | : | Peaks Over Threshold. |
| $N_u$ | : | Nombre de dépassements du seuil $u$. |
| $S_n$ | : | Somme arithmétique. |
| $D(.)$ | : | Domaine d'attraction. |
| $GPD$ | : | Distribution de Pareto Généralisée. |
| $PWM$ | : | Méthode des moments de probabilité pondérés. |
| $MLE$ | : | Méthode de maximum de vraisemblance. |

## Résumé

Ce mémoire porte sur la modélisation statistique des valeurs extrêmes, un outil essentiel pour l'analyse et l'estimation des phénomènes rares, notamment en hydrologie. Il explore les fondements théoriques des distributions GEV et GPD, appliquées respectivement aux maxima par blocs et aux excès au-dessus d'un seuil (approche POT). Une étude pratique sur les précipitations journalières en Angleterre, réalisée à l'aide du logiciel R, a permis d'estimer les paramètres, d'évaluer les ajustements et de calculer des quantiles extrêmes. Les résultats confirment la pertinence de ces modèles pour les événements rares.



**Mots-clés :** Approche des maxima par blocs, Approche POT, Distribution GEV, Distribution GPD, Estimation des paramètres, Sélection du seuil, Valeurs extrêmes.



## ملخص

تتناول هذه المذكرة النمذجة الإحصائية للقيم المتطرفة، وهي أداة أساسية لتحليل وتقدير الظواهر النادرة، لا سيما في مجال الهيدرولوجيا. حيث تستعرض الأسس النظرية لتوزيع القيم القصوى المعمم (GEV) وتوزيع باريتو المعمم (GPD) اللذين يُطبقان على التوالي على القيم القصوى حسب الكتل والتجاوزات فوق العتبة (منهجية POT). أجريت دراسة تطبيقية على بيانات التساقطات اليومية في إنجلترا باستخدام برنامج R، مما أتاح تقدير المعلمات، موافقة النماذج وحساب قيم احتمالية متطرفة. تؤكد النتائج أهمية هذه النماذج في تمثيل الظواهر النادرة.



**الكلمات المفتاحية** : منهجية القيم القصوى حسب الكتل، منهجية POT، توزيع GEV، توزيع GPD، تقدير المَعلمات، اختيار العتبة، القيم المتطرف.



## Abstract

This thesis focuses on the statistical modelling of extreme values, an essential tool for the analysis and estimation of rare events, particularly in hydrology. It explores the theoretical foundations of the Generalized Extreme Value (GEV) and Generalized Pareto Distribution (GPD), applied respectively to block maxima and exceedance over a threshold (POT approach). A practical study on daily precipitation in England, conducted using the R software, enabled the estimation of parameters, assessment of model fit, and calculation of extreme quantiles. The results confirm the relevance of these models for rare events.



**Keywords:** Block maxima approach, POT approach, GEV distribution, GPD distribution, Parameter estimation, Threshold selection, Extreme values.