الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Saad Dahlab Blida
Faculté des Sciences
Département de Mathématiques

Mémoire de fin d'étude

en vue de l'obtention du diplôme de

Master en Mathématiques

Option

Modélisation Statistique et Stochastique

Thème

Sélection du seuil de la distribution de Pareto généralisée. Application aux données météorologiques

Réalisé par:

- Ali Cherif Nadjet
- Houari Ratiba

Soutenu le 02/10/2019

Devant le jury composé de :

Mr O.TAMI	Président	USDB
Mr A.RASSOUL	Examinateur	ENSH
Mr R.FRIHI	Promoteur	USDB

Année universitaire : 2018/2019

Remerciements

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

En second lieu, nous tenons à remercier notre encadreur Mr R.FRIHI, son précieux conseil et son aide durant toute la période du travail.

Nos vifs remerciements vont également aux membres du jury pour l'intérêt qu'ils ont porté à notre recherche en acceptant d'examiner notre travail Et de l'enrichir par leurs propositions.

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

ملخص

نظرية القيمة القصوى هي أداة مناسبة لاستقراء سلوك ذيول التوزيع استنادا إلى أكبر (أو أصغر) القيم التي تمت ملاحظتها. يتم نمذجتها من خلال توزيع القيم القصوى أو من خلال توزيع باريتو المعمم . والهدف من عملنا هو تقدير العتبة بطرق و أساليب مختلفة. والتي من خلالها تبدأ الملاحضات في ان تصبح حدية. وبالتالي اختيار العدد الأمثل من الاحصاءات، التي تعتبر حاسمة في تقدير الحد. ومن هنا يمكن تطبيق هذه الأساليب على بيانات الأرصاد الجوية.

abstract

Extreme Value Theory (ETV) is an appropriate tool to extrapolate the behavior of distribution tails from the largest (or smallest) observed values. these are modelled by the distribution of extreme values Generalized (DGEV) or by the Distribution of Pareto Generalized (GPD). The objective of our work is to estimate the threshold of the GPD Act by different methods. The estimation of the GPD law tail index is crucial in the modelling process, which depends largely on the number of extreme statistics observed. This number determines the values, which among the data, which are really extreme. In other words, it allows to define the threshold where observations begin to become extreme. Selection of the optimal number of extreme statistics crucial for the IVE estimation We applied these methods to meteorological data.

Résumé

la théorie de valeurs extrêmes (TVE) représente un outil approprié permettant d'extrapoler le comportement des queues de distributions à partir des plus grand (ou plus petits) valeurs observées. ces derniers sont modélisés par la distribution des valeurs extrêmes Généralisée (DGEV) ou par la Distribution de Pareto Généralisée (GPD). L'objectif de notre travail est d'estimer le seuil de la loi de GPD par différentes méthodes. L'estimation de l'indice de queue de la loi de GPD est primordiale dans le processus de modélisation qui dépend largement du nombre de statistiques d'ordre extrêmes observées. Ce nombre détermine les valeurs, qui parmi les données, qui sont réellement extrêmes. En d'autres termes, il permet de définir le seuil où les observations commencent à devenir extrêmes.
La sélection du nombre optimal de statistiques d'ordre extrêmes cruciale pour l'estimation de l'IVE Nous avons appliqué ces méthodes aux données météorologiques.

Table des matières

Liste des tableaux

Table des figures

Notations et Abréviations

TVE	Théorie des valeurs extrêmes.
GEV	Distribution des Valeurs Extrêmes Généralisée.
GPD	Distribution de Pareto Généralisée.
POT	L'approche par dépassements de seuil.
mrlplot	Mean residual life plot .
tcplot	Threshold stability plots
TCL	Théorème central limite.
IVE	Indice des valeurs extrêmes.
i.i.d	Indépendant et identiquement distribué.
EMV	Méthode du Maximum de Vraisemblance.
$F^{\leftarrow}$	Inverse généralisée de la fonction de répartition.
F^{-1}	Inverse de la fonction de répartition.
x_F	Point terminal.
F_n	Fonction de répartition empirique.
F	Fonction de distribution.
F_u	Distribution conditionnelle des excès.
$(X_1, ..., X_n)$	Échantillon de taille n de X.
$(X_{1,n}, ..., X_{n,n})$	Échantillon ordonné.
$\mathbb{R}$	ensemble de nombre réels
$X_{i,n}$	La i ème statistique d'ordre.
$\mathcal{R}\nu_\rho$	Variation régulière d'indice ρ
$L(x)$	Fonction à variation lente
ϕ	Loi de Gumbel
ψ	Loi de Fréchet
λ	Loi de weibull

Introduction Générale

L'apparition de valeurs extrêmes dans une série d'observations relatives à un certain phénomène témoigne de l'occurrence d'événements rares, qui malgré leur faible probabilité ont des répercussions (souvent négatives) sur les décideurs (individus ou institutions). D'où l'importance de la construction de modèles statistiques décrivant le mieux possible ces observations. A cet effet, la théorie de valeurs extrêmes (TVE) représente un outil approprié permettant d'extrapoler le comportement des queues de distributions à partir des plus grandes (ou plus petites) valeurs observées. Les modèles des valeurs extrêmes sont appliqués à une grande variété de problèmes tels l'environnement (vitesse du vent,précipitation,humidité et de températures,...), la pluviométrie, la finance et l'assurance et en hydrologie pour calculer la probabilité que la hauteur d'eau d'un fleuve dépasse un certain seuil. En télécommunications, physique,. . .etc.
La modélisation et des distributions à queues lourdes est fortement liée à un nombre réel appelé indice de queue ou indice des valeurs extrêmes (IVE) et dont la valeur gouverne le degré d'épaisseur des queues. L'estimation de cet indice est primordiale dans le processus de modélisation, dépend largement du nombre de statistiques d'ordre extrêmes observées. Ce nombre détermine les valeurs, qui parmi les données, qui sont réellement extrêmes. En d'autres termes, il permet de définir le seuil où les observations commencent à devenir extrêmes. La sélection du nombre optimal de statistiques d'ordre extrêmes cruciale pour l'estimation de l'IVE et permet d'améliorer la performance des estimateurs est alors notre but dans ce mémoire.
Le présent mémoire est alors une synthèse des travaux de recherches concernant les méthodes d'estimation du seuil de la loi GPD Le plan du La sélection du nombre optimal de statistiques d'ordre extrêmes est cruciale pour l'estimation de l'IVE et permet d'améliorer la performance des estimateurs est alors notre but dans ce mémoire.
Le présent mémoire est alors une synthèse des travaux de recherches concer-

nant les méthodes d'estimation du seuil de la loi GPD Le plan du travail est le suivant :
Dans le chapitre 1 , nous rappelons quelques éléments théoriques essentiels de la théorie des valeurs extrêmes (TVE). Il contient des rappels sur la statistique d'ordre. Nous avons présenté le comportement asymptotique du maximum d'un échantillon. Cette étude fait appel à la notion de fonctions à variations régulières. Nous avons présenté aussi la méthode des excès au-delà d'un seuil (POT).

Dans le chapitre 2 nous donnons une aperçu sur quelques méthodes d'estimation du seuil.
Dans le chapitre 3 une étude de simulation est réalisée, puis nous avons appliqué ces méthodes aux données météorologiques (température, humidité, précipitation et vitesse du vent) en utilisant le logiciel R et matlab
Nous avons clôturé ce travail par une conclusion générale.

Chapitre 1

Théorie des valeurs extrêmes

1.1 Introduction

La théorie des valeurs extrêmes (TVE) concerne les questions probabilistes et statistiques à des valeurs très élevées ou très faibles dans des suites des variables aléatoires et dans les processus stochastiques. La TVE a permis de modéliser la queue de distribution des pertes d'un portefeuille d'investissement.
La théorie des valeurs extrêmes est appliquée en hydrologie pour prévoir les crues, en démographie pour prévoir la distribution de probabilité de l'âge maximum que l'être humain pourra atteindre, en assurance pour prévoir les grands sinistres, en finance ou encore en météorologie.
L'objectif essentiel de ce chapitre est de présenter les définitions et les résultats principaux sur la théorie des valeurs extrêmes dans le cas unidimensionnel. On commence dans un premier temps par donner la notion et les propriétés des statistiques d'ordre dont lesquelles repose l'utilisation des lois des valeurs extrêmes.

1.2 Statistique d'ordre

Considérons n variables aléatoires réelles $X_1, X_2, ..., X_n$ indépendantes et identiquement distribuées (iid) définies sur l'espace$(\Omega, \mathfrak{B})$, d'une densité commune f et d'une fonction de répartition F.

Définition 1. *Rangeons les variables aléatoires $X_1, X_2, ..., X_n$ par ordre croissant de grandeur, on introduit la notation $X_{i,n}$ avec :*

$$X_{1,n} \leq X_{2,n} \leq ..... \leq X_{n,n} \tag{1.1}$$

$X_{i,n}$ est la i ème statistique d'ordre (ou statistique d'ordre)

Remarque 1. *L'échantillon ordonné*
$(X_{1,n} \leq X_{2,n} \leq ... \leq X_{n,n})$ n'est pas iid. Pour $i \neq j$, $X_{i,n}$ et $X_{j,n}$ sont dépendants l'une de l'autre ; en particulier si $i < j$ alors $X_{i,n} < X_{j,n}$.
Deux statistiques d'ordre sont intéressantes pour l'étude des événements extrêmes,ce sont :

$$X_{1,n} = min(X_1, X_2, ..., X_n).$$

qui est la plus petite statistique d'ordre ou statistique du minimum

et

$$X_{n,n} = max(X_1, X_2, ..., X_n).$$

qui est la plus grande statistique d'ordre ou statistique du maximum.
On s'intéresse au comportement de ces deux statistiques.
$F_{1,n}(x) = P(X_{1,n} < x)$ *et* $F_{n,n}(x) = P(X_{n,n} < x)$.

1.2.1 Lois des statistiques d'ordre

Lemme 1. *Arnold Barry et al. (1992) [6] La densité conjointe de statistique d'ordre* $X_{1,n} \leq ... \leq X_{n,n}$*est donnée par*

$$f_{(X_{1,n} \leq ... \leq X_{n,n})}(x_1, x_2, ..., x_n) = n! \prod_{i=1}^{n} f(x_i).$$

avec $x_1 \leq x_2, ..., \leq x_n$

Lemme 2. *Arnold Barry et al. (1992)[6] (Densité conjointe de deux statistique d'ordre) La densité conjointe de*$(X_{r,n} \leq X_{s,n})$ *avec* $r \leq s$ *est donnée par :*

$$f_{(X_{r,n} \leq X_{s,n})}(x, y) = n! \frac{[F(x)]^{r-1} f(x) [F(y) - F(x)]^{s-r-1} f(y) [1 - F(y)]^{n-s}}{(r-1)!(s-r-1)(n-s)!}$$

avec

$$-\infty < x < y < +\infty$$

La fonction de répartition $F_{(X_{r,n} \leq X_{s,n})}$

$$F_{(X_{r,n} \leq X_{s,n})}(x, y) = F_{X_{s,n}}(y)$$

pour $x \geq y$ *et*

$$F_{(X_{r,n} \leq X_{s,n})}(x, y) = \sum_{j=s}^{n} \sum_{i=r}^{j} n! \frac{[F(x)]^{i-1} f(x) [F(y) - F(x)]^{j-i-1} f(y) [1 - F(y)]^{n-j}}{(i-1)!(j-i-1)!(n-j)!}$$

pour $x < y$.
Notons par $F - i, n(x)$ *la distribution de* $X_{i,n}$ *David (1970)[19] et Balakrishnan et Cohen (1991)[37] ont démontré que l'expression de* $F_{i,n}(x)$ *est*

$$F_{i,n}(x) = P(X_{i,n} < x) = \sum_{r=i}^{n} C_n^k [F(x)]^r [1 - F(x)]^{n-r}$$

La fonction de densité est :

$$f_{i,n}(x) = \frac{n!}{(i-1)!(n-i)!}[F(x)]^{i-1}[1-F(x)]^{n-i}f(x)$$

En particulier, On peut conclure pour la statistique du maximum que la distribution et la densité sont :

$$F_{n,n}(x) = [F(x)]^n \tag{1.2}$$

et

$$f_{n,n}(x) = n[Fx]^{n-1}f(x)$$

De même pour la statistique du minimum, on a :

$$F_{1,n}(x) = 1 - [1-F(x)]^n \tag{1.3}$$

et

$$f_{1,n}(x) = n[1-F(x)]^{n-1}f(x)$$

Dans la suite de ce travail, on ne présente que les résultats concernant le maximum, puisque les résultats relatifs au minimum se déduisent de l'égalité suivante

$$\min(X_1, X_2, ..., X_n) = -\max(-X_1, -X_2, ..., -X_n)$$

1.2.2 Loi des valeurs extrêmes

La formule (1.2) montre que la loi de maximum est relié d'une manière principale avec $F(x)$, mais cette dernière n'est pas toujours connue, même si elle est connue, la loi du maximum n'est pas facile à calculer. Donc on s'intéresse à étudier les comportements asymptotiques du maximum en faisant tendre n vers l'infini. Un théorème a été proposé par Gnedenko (1943)[10] connu sous le nom de théorème de Fisher-Tippett-Gnedenko ou théorème de la valeur extrême qui donne la forme des lois limites pour le maximum normalisé. Les précédentes versions ont été énoncés par Fréchet (1927)[34] et par Fisher et Tippett (1928)[42]. Jenkinson (1955)[8] a donné l'expression générale. Pour plus de détails sur cet sujet, on peut se référer [29].

Définition 2. *On dit que deux variables aléatoires réelles X et Y des lois respectives F et H sont de même type, s'il existe des constantes réelles $a > 0$ et $b \in R$ tels que $Y = aX + b$ i.e $F(ax+b) = H(x)$*

Fisher et Tippett (1928)[42] ont démontré l'existence des suites de normalisation $a_n > 0$ et $b_n \in R$ et une loi non-dégénérée H telle que :

$$p\left\{\frac{X_{n,n} - b_n}{a_n} \leq x\right\} = (F(a_n x + b_n))^n \to H(x)$$

Le théorème suivant donne une caractérisation de la distribution limite du maximum $X_{n,n}$.

Théoreme 1. ***(Fisher-Tippett-Gnedenko)***[42]
S'il existe deux suites de constante de normalisation $(a_n)_{n\leq 1}, a_n > 0$ et $(b_n)_{n\leq 1}, b_n \in R$ et une loi non-dégénérée de loi H telles que $\frac{X_{n,n}-bn}{a_n} \xrightarrow{L} H$, alors H appartient à une des trois distributions standard des valeurs extrêmes suivantes :

1. *Fréchet :* $\phi_\alpha(x) = \begin{cases} 0 & x \leq 0 \\ & \alpha > 0 \\ exp(-x^{-\alpha}), & x>0 \end{cases}$

2. *Weibull :* $\psi_a(x) = \begin{cases} exp(-(-x)^\alpha) & x \leq 0 \\ & \alpha < 0 \\ 1, & x>0 \end{cases}$

3. *Gumbel :* $\Lambda_0(x) = \{exp(-exp(-x)), x \in \mathbb{R}\}$

Ce théorème montre que la loi limite des extrêmes à toujours la même forme. Les trois formules précédentes peuvent être combinées en une seule paramétrisation :

$$H_\xi(x) = \begin{cases} exp\left(-(1+\xi x)^{\frac{-1}{\xi}}\right) & \text{si } \xi \neq 0,\ 1+\xi x > 0 \\ exp(-exp(-x)) & \xi = 0,\ -\infty \leq x \leq +\infty \end{cases} \tag{1.4}$$

où H est une fonction de répartition non-dégénérée et ξ est un paramètre qui contrôle la lourdeur de la queue de loi appelé indice des valeurs extrêmes (ou indice de queue).
Cette loi est appelée loi de valeurs extrêmes généralisée (Generalized Extreme Value) que l'on note GEV. la forme la plus générale de la GEV est :

$$H_{\xi,u,\sigma}(x) = exp\left\{-\left(1+\xi\frac{x-u}{\sigma}\right)^{-1/\xi}\right\},\ \xi \neq 0, 1+\xi\frac{x-u}{\sigma} > 0 \tag{1.5}$$

u et σ sont respectivement les paramètres de localisation et d'échelle.
ξ est le paramètre de forme.

La figure (1.1) [20] présente les différentes formes de H_ξ possibles en fonction de ξ .

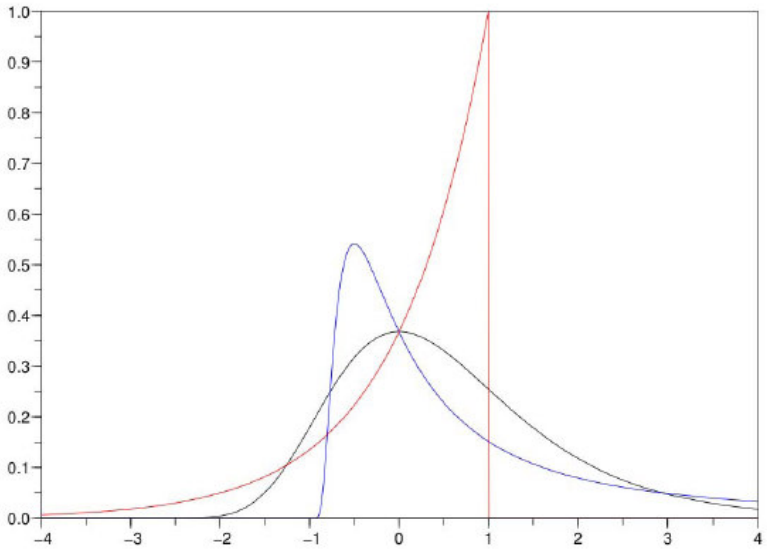


FIGURE 1.1: Exemples de densités associées à la loi des valeurs extrêmes (noir :$\xi = 0$,bleu :$\xi = 1$,rouge :$\xi = -1$)

1.3 Caractérisation des Domaines d'attraction

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

1. Lorsque $\xi = 0$ la loi des valeurs extrêmes présente une décroissance de type exponentiel dans la queue de la loi, on dit alors que la loi appartient au domaine d'attraction de Gumbel. C'est le cas des lois normale, exponentielle, etc ...

2. Le cas $\xi > 0$ correspond à la loi de Fréchet de paramètre $\alpha = 1/\xi$ dont la fonction de survie décroit comme une fonction puissance, et dans ce cas on dit que la loi appartient au domaine d'attraction de Fréchet. C'est le cas par exemple de la loi de Cauchy ou Pareto.

3. Le cas $\xi < 0$ correspondant à la loi de Weibull de paramètre $\alpha = -1/\xi$ et donc la loi appartient au domaine d'attraction de Weibull. c'est le cas par exemple de la loi uniforme ou béta.
   On peut trouver un classement de différents lois par domaine

- d'attraction dans Embrechts et al.(1997)[40] .

Voici un classement de quelques lois par domaine d'attraction dans le tableau 1.1 [20] .

TABLE 1.1: *Quelques lois et leurs domaines d'attraction*

domaines d'attraction	Fréchet ($\xi > 0$)	Weibull ($\xi < 0$)	Gumbel ($\xi = 0$)
lois	Cauchy Pareto Student	Uniforme Beta	Gumbel Exponentielle Lognormale Gamma Weibull Normal

Dans la suite, on va présenter quelques théorèmes de caractérisation des trois domaines d'attraction. Pour cela, on va faire un appel à quelques notions principales.

Point terminal :

Le point terminal d'une fonction F est :

$$x_F = sup\{x, F(x) \leq 1\}$$

Inverse généralisé :

l'inverse généralisée d'une fonction F est l'application définie par

$$F^{\leftarrow}(x) = inf\{x \in \mathbb{R}, F(x) \geq y\}$$

Fonction à variations régulières

On dit qu'un fonction $U(.)$ est à variations régulières à l'infini d'indice $\rho \in \mathbb{R}$, que l'on notera $U(.) \in \mathcal{R}\nu_\rho$, si U est positive à l'infini (i.e. s'il existe A tel que pour tout $x > A, U(x) > 0$) et si pour tout $\lambda > 0$,

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

– si $\rho = 0$, c'est-à-dire $U(.) \in \mathcal{R}\nu_0$, alors la fonction $U(.)$ est appelée fonction à variations lentes à l'infini, notée pour la suit $L(.)$

– si $\rho = \infty$, on parle de fonction à variations rapides à l'infini. On peut montrer facilement que tout fonction $U(.)$ à variations régulières d'indice

$\rho \in \mathbb{R}$ s'écrit,

$$U(x) = x^{\rho} L(x), L(x) \in \mathcal{R}\nu_0$$

Proposition 1. *(Resnick(1987))*[47]
Soit U est une fonction à variations régulières d'indice ρ, alors pour tout $0 < a < b$

$$\lim_{x\to\infty} \sup_{\lambda\in[a,b]} \left| \frac{U(\lambda x)}{U(x)} - \lambda^{\rho} \right| = 0.$$

Lemme 3. *(**Resnick** (1987))*[47]
– Si U est à variations régulières d'indice $\rho > 0$, alors $U^{\leftarrow}(x)$ est à variations régulières d'indice $1/\rho$.
– Si U est à variations régulières d'indice $\rho < 0$, alors $U^{\leftarrow}(1/x)$ est à variations régulières d'indice $-1/\rho$.

Théoreme 2. *(**Représentation de Karamata**)*
Toute fonction à variations lentes $L(.)$ s'écrite sous la forme :

$$L(x) = c(x) exp\left\{ \int_1^x \frac{\Delta(u)}{u} du \right\}. \tag{1.6}$$

où $c(x) \to c > 0$ et $\Delta(x) \to 0$ lorsque $x \to \infty$. Cette représentation des fonctions à une variations lentes est connue sous le nom de représentation de Karamata-Bingham et al. (1987)[36] .
Si la fonction $c(.)$ est constante, la fonction $L(.)$ est dite normalisée.

1.3.1 Domaine d'attraction de Fréchet

Gnedenko (1943)[10] a énoncé un résulta qui assure que toute fonction appartenant au domaine d'attraction de Fréchet est une fonction à variations régulières.

Théoreme 3. *Une fonction de répartition $F(.)$ appartient au domaine d'attraction de Fréchet si et seulement si elle s'écrit sous la forme :*

$$F(x) = 1 - x^{-1/\xi} \ell(x) \tag{1.7}$$

telle que $\ell(.)$ est une fonction à variations lentes.
Dans ce cas les suites de normalisation $(a_n)_n$ et $(b_n)_n$ sont données pour tout $n>0$ par $a_n = \overline{F}^{\leftarrow}(1/n)$ et $b_n = 0$.
On rappel que $\overline{F}$ est la fonction de survie définie par
$\overline{F}(x) = 1 - F(x)$

Bingham et al. (1987)[36] ont montré que l'équation (1.7) est équivalente à :

$$Q(\alpha) = \alpha^{-\xi}\ell(\alpha^{-1})$$

où $\ell(.) \in \mathcal{R}\nu_0$, $\alpha \in [0,1]$
et $Q(\alpha)$ *est la fonction quantile définie par*

$$Q(\alpha) = inf\{x, F(x) \geq \alpha\}$$

1.3.2 Domaine d'attraction de Weibull

Gnedenko (1943)et Resnick (1987) ont démontré que par un simple changement de variable dans la fonction de répartition, on peut passer du domaine d'attraction de Fréchet à celui de Weibull.

Théoreme 4. *Une fonction de répartition* $F(.)$ *appartient au domaine d'attraction de Weibull si et seulement si son point terminal* x_F *est fini et si la fonction de répartitionF*$(.)$ *définie par :*

$$F(x) = \begin{cases} 0 & si\ x < 0 \\ F(x_F - 1/x) & x \geq 0 \end{cases} \tag{1.8}$$

appartient au domaine d'attraction de Fréchet avec un indice des valeurs extrêmes
Ainsi, une fonction de répartition $F(.)$ *du domaine d'attraction de Weibull s'écrit :*

$$F(x) = 1 - (x_F - x)^{-1/\xi}L((x_f - x)^{-1}),\ L(.) \in \mathcal{R}\nu_0\ pour\ x \leq x_F$$

De manière équivalente, le quantile $Q(\alpha), \ell(.) \in \mathcal{R}\nu_0$ *associé s'écrit :*

$$Q(\alpha) = x_F - \alpha^{-\xi}\ell(1/\alpha) \tag{1.9}$$

Les suites de normalisation (a_n)*et* (b_n) *sont données par* $a_n = b_n - \overline{F}^{\leftarrow}(1/n)$ *et* $b_n = x_F$
Ce domaine d'attraction a été considéré par Falk (1995), Gardes (2010) pour donner un estimateur de point terminal d'une distribution.

1.3.3 Domaine d'attraction de Gumbel

Le domaine d'attraction de Gumbel contient les lois où la fonction de survie est à décroissance exponentielle, i.e. les lois à queues légères. Le théorème suivant qui a été démontré dans Resnick (1987)donne une caractérisation de ce domaine.

Théoreme 5. *Une fonction de répartition $F(.)$ appartient au domaine d'attraction de Gumbel si et seulement si il existe $t < x_F \leq \infty$ tel que*

$$\overline{F}(x) = c(x)exp\left\{-\int_t^x \frac{1}{a(u)}du,\ t < x \leq x_F\right\}. \tag{1.10}$$

où $c(x) \to c > 0$ lorsque $x \to x_F$ et $a(.)$ est une fonction positive et dérivable de dérivée $a'(.)$ telle que $a'(x) \to 0$ lorsque $x \to x_F$.
Dans ce cas, un choix possible pour les suites (a_n) et (b_n) pour tout $n > 0$ est :

$$a_n = q(1/n)\ et\ b_n = \frac{1}{\overline{F}(a_n)}\int_{a_n}^{x_F}\overline{F}(s)ds.$$

Dans la section ci dessous, et sous certaines hypothèses sur la loi $F(x)$, on va donner quelques exemples expliquant le comportement des lois limites de la GEV.

1.4 Estimation de l'indice de queue

Les deux estimateurs sans doute les plus populaires dans la littérature sont les estimateurs de Hill (1975)[12] et de Pickands (1975)[30] .
On note $X_{1,n}, ..., X_{n,n}$ les statistiques d'ordre associées à l'échantillon $X_1, ..., X_n$.
C'est-à-dire que l'on classe $X_1, ..., X_n$ par ordre croissant de sorte que :

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

On considère les k valeurs les plus grandes (ou les plus petites), k dépend a priori de n, même si on ne le mentionnera pas dans la notation, l'idée est d'avoir $k \to \infty$ lorsque $n \to \infty$, mais sans prendre « trop » de valeurs de l'échantillon, ce qui conduit à imposer $\frac{k}{n} \to 0$.
Incidemment, cela implique que se posera la question du choix optimal de k. En effet, il est indispensable de calculer ces estimateurs sur les queues de distribution, Choisir un k trop élevé engendre le risque de prendre en compte des valeurs qui ne sont pas extrêmes, à l'inverse, un sous-échantillon trop petit ne permet pas aux estimateurs d'atteindre leur niveau de stabilité.
Enfin, on retiendra que l'approche non paramétrique n'est envisageable que si l'on dispose d'un nombre important d'observations, dans le cas ou les échantillons sont de petite taille, on se tournera vers l'approche paramétrique.

1.4.1 L'estimateur de Pickands

Cet estimateur a été introduit en 1975 par James Pickands, pour toute $\xi \in \mathbb{R}$ l'estimateur de Pickands est défini par :

$$\hat{\xi}^{p}_{k,n} = \frac{1}{\log 2} \log \left(\frac{X_{n-k+1:n} - X_{n-2k+1,n}}{X_{n-2k+1,n} - X_{n-4k+1,n}} \right) \tag{1.11}$$

Il présente l'intérêt d'être valable quelle que soit la distribution des extrêmes (Gumbel, Weibull ou Fréchet). La représentation graphique de cet estimateur en fonction du nombre k d'observations considérées montre un comportement en général très volatil au départ, ce qui nuit à la lisibilité du graphique.
De plus, cet estimateur est très sensible à la taille de l'échantillon sélectionné,ce qui le rend peu robuste.
On peut noter qu'il est asymptotiquement normal, avec :

$$\sqrt{k} \frac{\hat{\xi}^{p}_{k,n} - \xi}{\sigma(\xi)} \to \mathcal{N}(0,1)$$

Lorsque $k \to \infty$ l'écart type asymptotique est donné par :

$$\sigma(\xi) = \frac{\xi \sqrt{2^{2\xi+1} + 1}}{2(2^{\xi} - 1) ln(2)}$$

1.4.2 Estimateur de Hill

L'estimateur de Hill n'est utilisable que pour les distributions de Fréchet (donc telles que $\xi > 0$) pour lesquelles il fournit un estimateur de l'indice de queue plus efficace que l'estimateur de Pickands. Il est défini par la statistique suivante[12] :

$$\hat{\xi}^{H}_{k,n} = \frac{1}{k} \sum_{i=1}^{k} \log \left(X_{n+i-1,n} - X_{n-k,n} \right) \tag{1.12}$$

On choisit k, $n \to \infty$ de sorte que $\frac{k}{n} \to 0$ alors on peut montrer que $\lim_{k \to +\infty} \hat{\xi}^{H}_{k,n} = \xi$ et l'estimateur de Hill est de plus asymptotiquement normal :

$$\sqrt{k} \frac{\hat{\xi}^{H}_{k,n} - \xi}{\xi} \to \mathcal{N}(0.1)$$

la convergence étant en loi.
Cet estimateur est l'estimateur du maximum de vraisemblance dans le cas particulier du modèle $S(x) = 1 - F(x) = Cx^{-1/\xi}$,on reconnaît ici une

distribution de Pareto d'indice $\alpha = \frac{1}{\xi}$
Dans le cas général du domaine de Fréchet, la fonction de survie est de la forme $S(x) = 1 - F(x) = x^{-1/\xi}L(x)$ avec L une fonction à variation lente. Cela induit un biais important sur l'estimateur de Hill, qui est donc en pratique d'un maniement délicat. Dans le cas général, la fonction L apparaît comme un paramètre de nuisance de dimension infinie, qui complique l'estimation.

1.4.3 L'estimateur des moments

C'est un estimateur proposé par Dekkers et al [7] . Comme pour l'estimateur de Hill.
Cet estimateur est défini par la statistique [25] :

$$\hat{\xi}^{M}_{k,n} = 1 + M^{(r)}_{n} - \frac{1}{2}\left(1 - \frac{(M^{(1)}_{n})^2}{M^{(2)}_{n}}\right)^{-1}, \; r = 1, 2 \tag{1.13}$$

avec $M^{(r)}_{n} = \frac{1}{k}\sum_{i=1}^{k}\left(\log X_{n-i+1,n} - \log X_{n-k,n}\right)^2$. et $M^{(1)}_{n}$ est l'estimateur de Hill $\hat{\xi} - k, n^{M}$.
Cet estimateur est convergent et asymptotiquement gaussien :

$$\sqrt{k}\frac{\hat{\xi}^{M}_{k,n} - \xi}{\sqrt{1+\xi^2}} \to \mathcal{N}(0,1)$$

1.4.4 Comparaison des différents estimateurs

La figure 1.2 nous montre l'erreur de l'estimation en fonction de la méthode.[35]

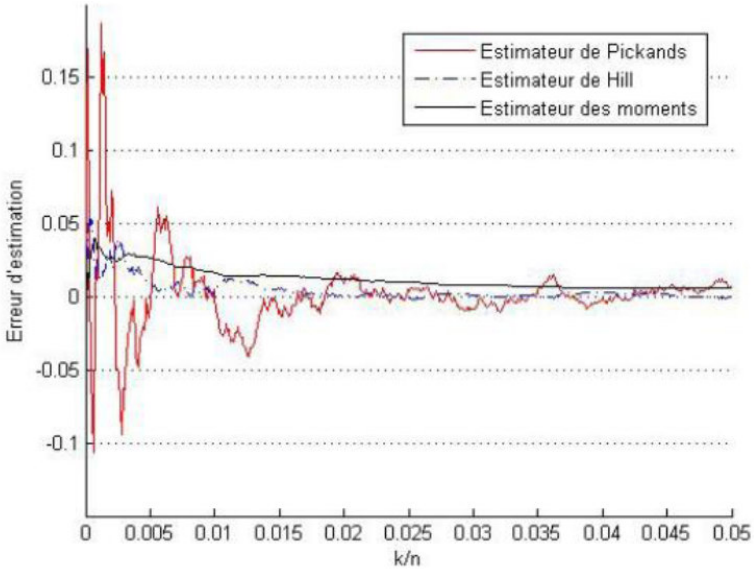


FIGURE 1.2: Comparaison de l'erreur d'estimation en fonction de la méthode .

On remarque que l'estimateur de Pickands est moins efficace dans l'estimation de l'indice de queue. Par contre, on observe une efficacité de l'estimateur de Hill sur ceux de Pickands et des moments.

1.5 Distribution des excès

La méthode des excès au-delà d'un seuil repose sur le comportement des valeurs observées au-delà d'un seuil donné. En d'autres termes, elle consiste à observer non pas le maximum ou les plus grandes valeurs mais toutes les valeurs des réalisations qui excèdent un certain seuil élevé. L'idée de base de cette approche consiste à choisir un seuil suffisamment élevé et à étudier les excès au-delà de ce seuil.
Cette méthode est initialement introduite par Pickands (1975) et étudiée par des auteurs tels que de Smith (1987) et Reiss et Thomas (2001).

On définit un seuil $u \in \mathbb{R}$, $N_u = card\{i : i = 1, ..., n;\ X_i > u\}$ et $Y_j = X_i - u > 0$ pour $0 \leq j \leq N_u$ où N_u est le nombre de dépassements du seuil u par les $X_{i \leq n}$ et les $Y_{j \leq N_u}$ sont les excès correspondants (Figure 1.3 [1]).

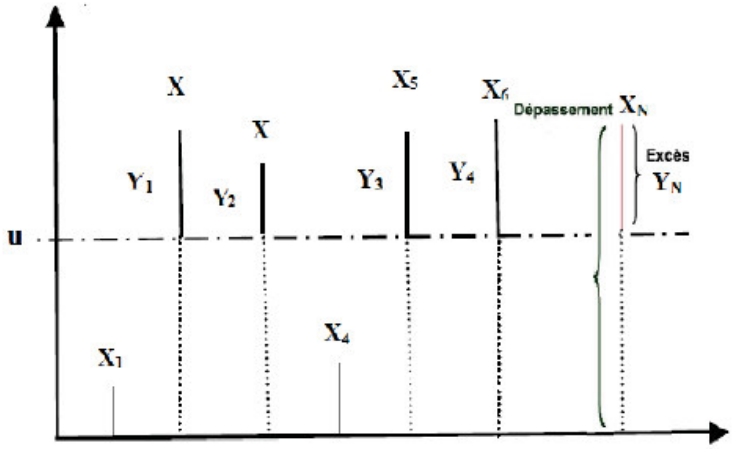


FIGURE 1.3: Dépassement de seuil (POT)

La loi conditionnelle des excès F_u par rapport au seuil u est :

$$F_u(y) = P(X - u \leq y/X > u) = \left(\frac{F(u+y) - F(u)}{1 - F(u)}\right) \quad pour\ 0 \leq y \quad (1.14)$$

ou de manière équivalente :

$$\overline{F_u}(y) = P(X - u > y|X > u) = -F_u(y) = \frac{\overline{F}(u+y)}{\overline{F}(u)},\ y \geq 0 \qquad (1.15)$$

1.5.1 Distribution de Pareto Généralisée (GPD)

La distribution de Pareto généralisée, joue un rôle essentiel dans la modélisation des excès.
Une distribution $G_{\xi,\sigma(u)}$ est dite de Pareto généralisée de paramètre $\xi \in R$ et $\sigma > 0$ si elle s'écrit :

$$G_{\xi,\sigma}(y) = \begin{cases} 1 - \left(1 + \frac{\xi}{\sigma}y\right)^{\frac{-1}{\xi}} & \text{si } \xi \neq 0 \\ 1 - exp(-\frac{y}{\sigma}) & \text{si } \xi = 0 \end{cases} \qquad (1.16)$$

Cette distribution est définie pour :

$$\begin{cases} y \geq 0 & \text{si } \xi \geq 0, \\ 0 \leq y \leq -\frac{\sigma}{\xi} & \text{si } \xi < 0 \end{cases}$$

Le paramètre ξ est le même d'une GVE .
ξ : paramètre de forme (queue).
σ : paramètre d'échelle.

Remarque 2. *Selon le signe de ξ, nous avons les cas suivants :*
$\xi > 0$: distribution de type Pareto à queue lourde.
$\xi < 0$: distribution de type Beta bornée au dessus de $u - \frac{\sigma_u}{\xi}$
$\xi = 0$: distribution de type exponentielle à queue légère.

Les figures (1.4) et (1.5) [9] représentent les fonctions de répartition et de densité des lois de Pareto pour un paramètre d'échelle fixé à 1. Ainsi, en faisant varier le paramètre de forme entre -1 et 1, on obtient différentes lois GPD comme suit, $\xi = -1$ pour la loi de Pareto II en rouge, $\xi = 0$ pour la loi exponentielle en noir et $\xi = 1$ pour la loi de Pareto en bleu. En remplaçant par exemple dans la relation (1.16) $\xi = 0$ et $\sigma_u = 1$; on obtient respectivement les fonctions de répartition et de densité suivantes :

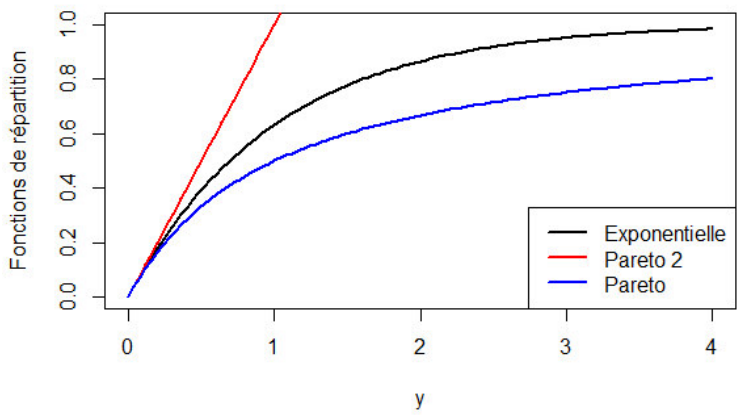


FIGURE 1.4: Fonctions de répartition des lois de Pareto Généralisée

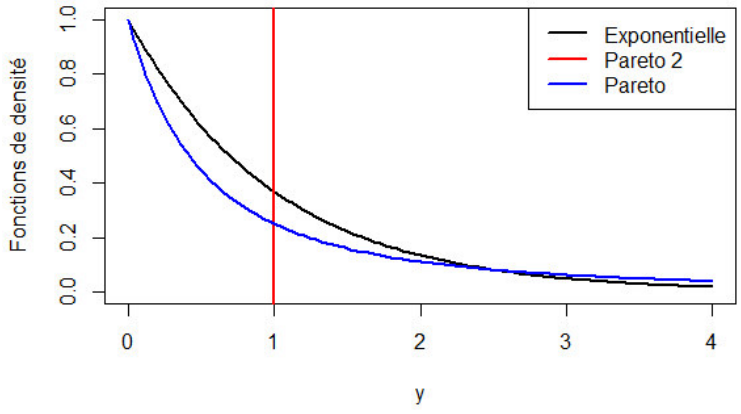


FIGURE 1.5: Fonctions de densité des lois de Pareto Généralisée

1.5.2 Théorème de Balkema-de Haan-Pickands

Le théorème de Pickands-Balkema-de Haan [30] ci-après, montre qu'on peut approchée la distribution des excès au delà d'un seuil u, sous certaines conditions de convergence, par une loi de Pareto généralisée que l'on note GPD, (Distribution de Pareto Généralisée).

Théoreme 6. *Une fonction de répartition F appartient au domaine d'attraction maximale de H_ξ, si et seulement si, il existe une fonction positive $\sigma(u)$ telle que :*

$$\lim_{u \to y_F} \sup_{0 \leq x \leq x_F - u} |F_u(y) - G_{\xi,\sigma(u)}(y)| = 0 \tag{1.17}$$

$F_u(y)$ est la fonction de répartition conditionnelle des excès au-dessus d'un seuil u, x_F est le point terminal de F, et $G_{\xi,\sigma(u)}(y)$ est la GPD.

On conclut de ce théorème, que si F vérifie le théorème de Fisher et Tippet, alors il existe une fonction $\sigma(.)$ positive et un réel ξ tels que la loi des excès F_u peut être uniformément approchée par une distribution de Pareto généralisée (GPD) notée $G_{\xi,\sigma(u)}$. Ainsi, l'indice des valeurs extrêmes donnée par le théorème de Fisher et Tippet est le même que celui de la loi des excès.

1.6 Estimation des paramètres de la GPD

1.6.1 Méthode du maximum de vraisemblance

Une fois le seuil optimal choisi, on construit une nouvelle série d'observations au dessus de ce seuil, et la distribution de ces données suit approximativement une distribution généralisée de Pareto.
La densité de la distribution GPD s'écrit [47] :

$$H_{\xi,\sigma}(x) = \begin{cases} \sigma^{1/\xi}(\sigma + \xi x)^{-\frac{1}{\xi}-1} & \text{si } \xi \neq 0 \\ \sigma^{-1} exp(-\frac{x}{\sigma}) & \text{si } \xi = 0 \end{cases} \tag{1.18}$$

Le logarithme de la fonction de vraisemblance que nous maximisons est de la forme :

$$\log L(\xi, \sigma, X_1, ..., X_n) = -n \log(\sigma) - (1 + \frac{1}{\sigma}) \sum_{i=1}^{n} \log(1 + \frac{\xi}{\sigma} X_i) \tag{1.19}$$

On pose $p = \frac{\xi}{\sigma}$, l'annulation des dérivées partielles des logarithmes de la fonction de vraisemblance conduit au système :

$$\begin{cases} \hat{\xi} = \frac{1}{n}\sum_{i=1}^{n} \log(1+pX_i) = \hat{\xi}(p) & \frac{1}{p} = \frac{1}{n}(1+\frac{1}{\hat{\xi}}) \\ \sum_{i=1}^{n} \frac{X_i}{1+pX_i} \end{cases}$$

L'estimateur du maximum de vraisemblance de (ξ, τ) est $(\hat{\xi}, \hat{\xi}(p), \hat{p})$ ou p est solution de :

$$\frac{1}{p} = \frac{1}{n}(1+\frac{1}{\xi})\sum_{i=1}^{n} \frac{\xi}{1+pX_i}$$

Cette dernière équation se résout numériquement de maniéré itérative pour autant que l'on dispose d'une valeur initiale $p-0$ pas trop éloignée de p. En pratique cette valeur initiale pourra être obtenue par la méthode des moments ou par la méthode des quantiles.

1.6.2 Méthode des moments pondérés

Définition 3. *On appelle moment pondéré d'ordre r, le moment défini par :*

$$u_r = \mathbb{E}\left[Z(\overline{H}_{\xi,\sigma}(Z))^r\right],\ r \in \mathbb{N}$$

Prenons par exemple Z suit la loi de Pareto généralisée de paramétré (ξ, σ). Le moment pondéré u_r est égal à :

$$u_r = \frac{\sigma}{(r+1)(r+1-\xi)},\ r = 0, 1$$

En résolvant l'équation précédente pour $r = 1$ et $r = 2$, on obtient :

$$\hat{\sigma} = \frac{2u_0u_1}{u_0 - 2u_1} \text{ et } \hat{\xi} = \frac{u_0}{u_0 - 2u_1}$$

On remplace u_0 et u_1 par leurs estimateurs des moments empiriques pondérés définis par :

$$M_r = \frac{1}{n}\sum_{j=1}^{n}\left(\prod_{l=1}^{r}\frac{n-j-l+1}{n-l}\right)X_{j,n} = \frac{1}{n}\sum_{j=1}^{n}(1-\frac{j}{n+1})X_{j,n}$$

o u u $X_{j,n}$ est la j ème statistique d'ordre.
On trouve un estimateur de ξ par la méthode des moments pondérés

$$\hat{\xi} = \frac{\frac{1}{n}\sum_{j=1}^{n}(4\frac{j}{n+1}-3)X_{j,n}}{\frac{1}{n}\sum_{j=1}^{n}(2\frac{j}{n+1}-1)X_{j,n}} \tag{1.20}$$

1.7 Sélection du seuil

Le choix du seuil doit établir un compromis entre biais et variance. Concrètement, le seuil doit être suffisamment grand pour pouvoir utiliser les résultats asymptotiques, mais pas trop élevé pour obtenir des estimations précises. Par contre le choix d'un seuil faible risque de déclarer abusivement des observations extrêmes, introduire un biais dans l'estimation et par conséquent, mal approximer la loi asymptotique. Dans ce sens, plusieurs méthodes de détection du seuil seront traitées dans le chapitre suivant.

Chapitre 2

Méthodes d'estimation du seuil

2.1 Introduction

La détermination du seuil est l'étape la plus délicate dans l'implémentation de l'approche POT, étant donné que la qualité du modèle en dépend. La convergence des excès vers une GPD passe par la détermination d'un seuil adéquat (pas très bas et pas très haut). Le choix du seuil doit être un compromis de sorte que le seuil déterminé soit suffisamment grand pour pouvoir utiliser les résultats asymptotiques, mais pas trop élevé afin d'obtenir des estimations précises. Cependant, le choix d'un seuil faible peut conduire à des incertitudes sur le nombre d'observations extrêmes et par conséquent produire des estimations biaisées et une mauvaise approximation de la loi asymptotique. Dans cette optique, plusieurs méthodes de détermination de seuils sont proposées dans la littérature. On distingue principalement deux approches, l'approche graphique et celle dite numérique. La plupart de ces méthodes sont subjectives et il est nécessaire de quantifier les incertitudes dues à ces méthodes.

2.2 Méthodes graphiques

Le Peak Over Threshold (POT) est une technique qui a été développée par des chercheurs , et elle est fréquemment utilisée dans la structure des valeurs extrêmes. L'approche POT consiste à ajuster un modèle paramétrique pour que ses excès au-dessus d'un seuil u soient assez élevés (de Zea Bermudez et Kotz, 2010). En d'autres termes, cette technique permet d'évaluer si le choix du seuil u est adéquat pour être représenté par un modèle asymptotique (e.g.exponentiel ou GPD). Afin de comprendre cette notion, la section suivante fera une brève introduction de quelques méthodes utilisant la technique POT.

2.2.1 Mean Residual life(MRL-plot))

Le graphe de la durée de vie moyenne résiduelle (MRlplot) introduite par Davison et Smith [26] utilise la méthode d'espérance des excès de GPD, $E(X-u|X>u) = \sigma_u/(1-\xi)$, comme diagnostic, définie pour $\xi < 1$ pour s'assurer de l'existence de la moyenne [16]. Pour tout $u > u^*$ plus élevé, l'attente devient

$$e(u) = E(x-u|x>u) = \frac{\sigma_{u^*}}{1-\xi} + \frac{\xi}{1-\xi}u,\ \xi < 1 \tag{2.1}$$

qui est linéaire en u^* avec $\xi/(1-\xi)$ et $\sigma_{u^*}/(1-\xi)$.

où u : indique le seuil.

ξ : indique le paramètre de forme.

σ_u : indique le paramètre d'échelle correspondant au seuil u.

$e(u)$ est la moyenne des excès au delà du seuil u.

Le graphe de la durée de vie moyenne résiduelle est le graphe des points $\{(u, e_n(u)), X_{1:n} < u < X_{n:n}\}$

Un estimateur empirique de cette fonction est donné par :

$$\hat{e}_n(u) = \frac{1}{N_u}\sum_{i=1}^{N_u}(x_i - u),\ xi > u,\ 0 < u < +\infty \tag{2.2}$$

où N_u : indique le nombre de dépassements par rapport à u.
Supposons données les observations $(X_1, X_2, ...X_n)$, on trace graphiquement $\hat{e}(u)$ en fonction de u et on choisit le plus petit u de manière à ce que $\hat{e}(u)$ soit approximativement linéaire pour tout $x > u$.
La fonction moyenne des excès empiriques sous la transformation affine s'écrit pour $x < x_f$, comme suit :

$$\hat{e}(u) = \frac{\hat{\sigma}_u + \hat{\xi}u}{1-\xi},\ \hat{\sigma}_u + \hat{\xi}u > 0 \tag{2.3}$$

Trois cas peuvent alors se présenter :

1. Si à un certain seuil, le mean excess plot est marqué par une pente positive. Alors les données suivent la distribution GPD avec un paramètre ξ positif (c'est à dire une distribution de pareto).

2. Si le graphe mean excess plot est horizontale. Alors les données suivent une distribution exponentielle..
3. Si le graphe mean excess plot est marquée par une pente négative. Alors les données suivent une distribution à queue légère.
   Le figures(2.1) présente le graphe de la durée de vie moyenne résiduelle de taille 3000.

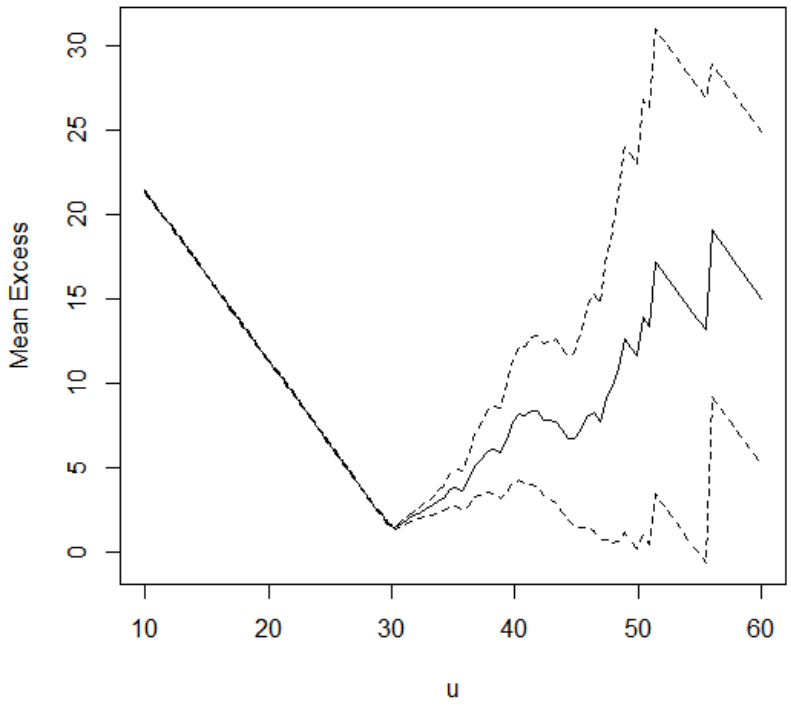


FIGURE 2.1: MRL-plot pour GPD considérés n=3000 et $\xi = 0.3$

sur la figure (2.1), on constate une linéarité entre 29 et 33. De ce fait, on peut affirmer que le seuil est compris entre 29 et 33

2.2.2 Graphes de Stabilité des paramètres de forme et d'échelle

Le Graphe de choix du seuil (tc-plot) [13] est un outils graphiques largement utilisés pour la sélection du seuil dans l'analyse POT.
Encore appelée « stable scale and shape parameters », cette méthode permet de déterminer un seuil requis en ajustant les données à une distribution de GPD en utilisant un seuil différent. La stabilité des paramètres (forme et échelle) peut alors être contrôlée et localisée.
Cette technique est implémentée dans le logiciel R avec des packages spécifiques. Ces packages disposent d'outils objectifs pour guider le choix du seuil adéquat en examinant tout simplement la stabilité des paramètres de forme et d'échelle ξ et σ. Ce graphe établit un lien direct entre les valeurs des paramètres estimés (ξ et σ) et les seuils potentiels u. Les paramètres estimés au-dessus des seuils sont ceux pour lesquels le modèle GPD devient valable.

La loi GPD satisfait à une propriété de stabilité de seuil, pour tout seuil supérieur $u^* > u$,les excès ultérieurs suivent également un GPD de même forme, mais de forme identique d'échelle décalée $\sigma_{u^*} = \sigma_u + \xi(u^* - u)$. L'indice u sur σ_u fait le seuil dépendance explicite, bien que dans le cas limite $\xi = 0$ ceci disparaisse.

« l'échelle modifiée » de reparamétérisation $\sigma^* = \sigma_{u^*} - \xi u^*$ est constant au-dessus de u, c'est-à-dire une fois que le GPD fournit une approximation de queue adéquate. Le paramètre de forme et d'échelle peut être orthogonalisé après Cox et Reid (1987)[15] avec la forme préférée habituellement $(\tilde{\sigma}_u, \xi)$, où $\tilde{\sigma}_u = \sigma_u(1 + \xi)$, car la forme est souvent un paramètre clé d'intérêt.
Les graphes établaient un lien direct entre les valeurs des paramètres estimés (ξ et σ) et les seuils potentiels u^*. Les paramètres estimés au-dessus des seuils sont ceux pour lesquels le modèle GPD devient valable.

Le figures(2.2) présente les graphes de stabilité dee paramétrer d'échelle et de forme.

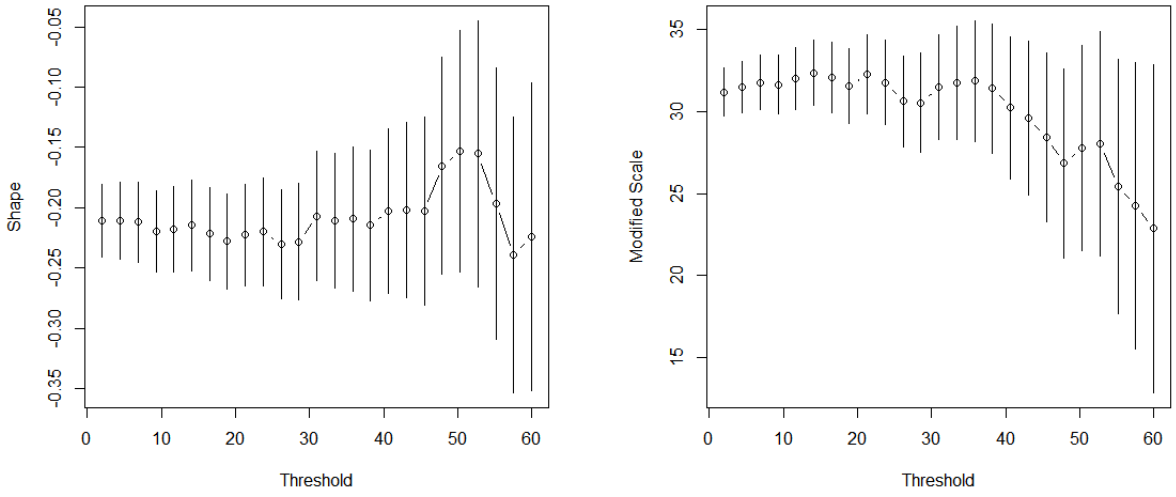


FIGURE 2.2: *Graphes de Stabilité des paramètres de forme (a gauche) et d'échelle (a droit) pour (**GPD**) considérés n=3000, $\xi = 0.3$)*

sur la figure (2.2), on remarque une stabilité sur l'intervalle [28-32] ,le seuil est situer autour de 30

2.2.3 Estimateur de Hill

9 l'estimateur de Hill ([12] et [27]) est donné par la forme empirique suivante :

$$\hat{\xi} = \frac{1}{k} \sum_{i=1}^{k} \log \left(X_{n+i-1,n} - X_{n-k,n} \right) \tag{2.4}$$

avec k, l'ordre statistique le plus élevé (le nombre des excès) et $\alpha = \frac{1}{\xi}$ est l'indice de la queue de distribution.
Cet estimateur intervient dans la construction du graphique
Hill-plot :Représentation de α en fonction de la statistique d'ordre $X_{1,n}$.Le Hill-plot nous permet de choisir un seuil élevé pour la construction d'un modèle (GPD).
Le Hill-plot est donc un outil à double utilité :

— L'estimation de l'indice de la queue de la distribution.
— L'estimation du seuil

Le graphique Hill-plot, nous permet d'avoir des estimations du paramètre en fonction de l'ordre statistique le plus élevé (nombre des excès), nous choisissons ainsi l'indice le plus stable.

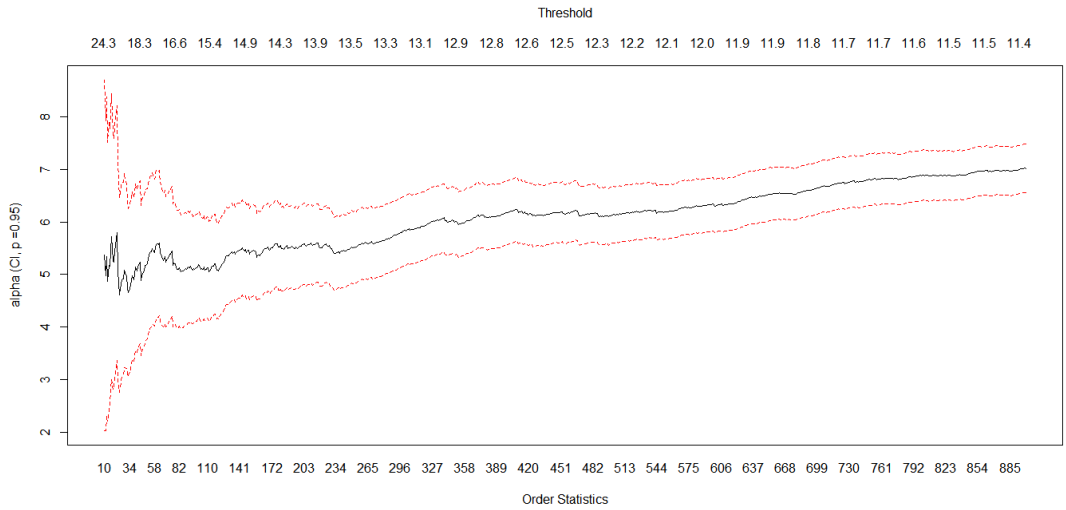


FIGURE 2.3: Hill-plot pour GPD(n=3000,u=10,$\sigma = 1, \xi = 0.3$)

Nous remarquons une **zone de stabilité** entre 141 et 234 excès. Au delà de 234 excès, l'estimateur n'est plus du tout stable.

Nous considèrerons donc que l'adéquation à une **GPD** débute au niveau du 234 ème excès, soit un **seuil à [13.9-15.0]**

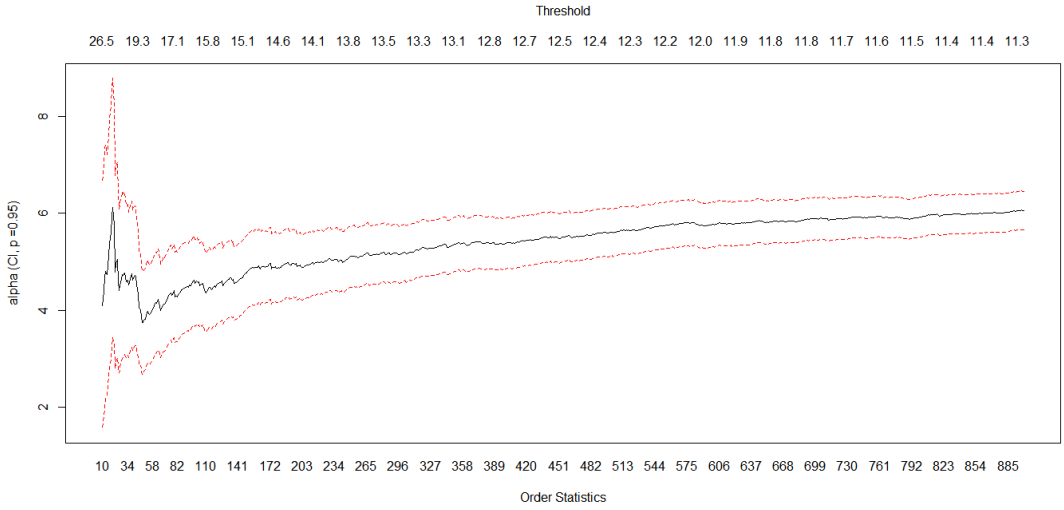


FIGURE 2.4: Hill-plot pour GEV(n=3000,u=10,$\sigma = 1, \xi = 0.3$)

Nous remarquons une **zone de stabilité** entre 172 et 220 excès. Au delà de 220 excès, l'estimateur n'est plus du tout stable. Nous considèrerons donc que l'adéquation à une **GEV** débute au niveau du 220 ème excès, soit un **seuil à [13.8-15.1]**

2.3 Méthode numérique

2.3.1 Procédure du Bootstrap

On distingue principalement deux approches : **l'approche paramétrique** qui est une méthode dans laquelle des hypothèses de base sont émises sur la distribution de l'échantillon de base et **l'approche non paramétrique**. Dans cette dernière approche, il peut y avoir des hypothèses sur la distribution de l'échantillon de base ; cependant, elle ne nécessite pas que la loi parent appartienne à une famille paramétrique.
La technique du bootstrap a été introduite par Efron (1979). C'est la méthode de réplication des échantillons la mieux fondée théoriquement. Elle consiste à créer, à partir d'un échantillon de base, un grand nombre d'échantillons par tirage aléatoire avec remise. Sur chaque échantillon, les statistiques auxquelles on s'intéresse sont calculées, ce qui permet d'approcher leur dispersion. De ce fait, on peut estimer la variance ou la loi des paramètres caractéristiques de la distribution de l'échantillon et construire

des intervalles de confiance, lorsque la distribution des paramètres est analytiquement complexe. Cette méthode semble donc adéquate pour la quantification des incertitudes dans la théorie des valeurs extrêmes.
Dans ce travail, on s'intéressera uniquement à l'approche non paramétrique pour quantifier les incertitudes sur les estimations.

Définition 4. *Le bootstrap est une technique de reéchantillonnage permettant de simuler la distribution d'un estimateur quelconque pour en apprécier le biais, la variance, l'erreur quadratique moyenne ou encore pour en estimer un intervalle de confiance, même si la loi théorique est inconnue.*

L'objectif de cette note est de décrire comment le bootstrap peut être utilisé pour résoudre les problèmes d'inférence statistique en relation avec l'estimation des paramètres. Nous présentons d'abord les méthodes de rééchantillonnage. Ensuite, nous examinons l'estimation de l'erreur-standard et du biais d'un estimateur.

a. **Approche non paramétrique du bootstrap**
Cette approche est généralement utilisée dans les situations où l'on ne peut pas faire l'hypothèse que la distribution des observations de certains paramètres appartient à une famille connue.

b. **Échantillon bootstrap** [44]
On a tiré n échantillon
$X = (x_1, x_2, ..., x_i, ..., x_n)$ d'une fonction de distribution inconnue F.
$F \rightarrow X = (x_1, x_2, ..., x_i, ..., x_n)$
On veut estimer un paramètre $\theta = t(F)$ a partir de X
On calcule un estimateur $\hat{\theta} = s(X)$
$\hat{F}$ est la fonction de distribution qui donne la probabilité $1/n$ a chaque x_i

Le principe de la méthode du bootstrap est de prélever une série d'échantillons aléatoires et simples avec remise de n observations dans l'échantillon initial, considéré comme une population.
Ces échantillons successifs seront notés :

$$X^* = (X_1^*, X_2^*, ..., X_b^*, ..., X_B^*),$$

suivant la loi $\hat{F}$.

Pour l'ensemble des B échantillons obtenus par bootstrap, les observations x_i n'apparaissent pas en nombre égal et on peut définir les proportions d'apparition P_i^* de chacune des observations, P_i^* étant égal au nombre de fois que l'observation x_i a été prélevée pour l'ensemble des B échantillons, divisé par le nombre total de prélèvements, qui est égal à n_B.

c. **Estimation de l'erreur-standard**[11]

Soit un paramètre θ de la population et soit :

$$\hat{\theta} = s(x_1, ..., x_n) = s(X)$$

une estimation de ce paramètre, obtenue à partir des données de l'échantillon initial X. Chaque échantillon obtenu par rééchantillonnage permet de calculer une répétition du bootstrap de l'estimation $\hat{\theta}$:

$$\hat{\theta_b}^* = s(X_b^*);\ b = (1, ..., B)$$

la fonction s étant la même que celle utilisée pour la définition de $\hat{\theta}$. Supposons qu'on s'intéresse à la moyenne, à la médiane et à la variance et qu'on se propose d'estimer ces trois paramètres à partir de l'échantillon X.

Si on utilise les estimateurs classiques, le paramètre $\hat{\theta}.$ s'écrit, successivement pour les trois paramètres considérés :

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\tilde{X} = \frac{1}{2}\left(x_{[\frac{n+1}{2}]} + x_{([\frac{n+1}{2}]+1)}\right)$$

et

$$\hat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$$

où $x_{[\frac{n+1}{2}]}$ et $x_{[\frac{n+1}{2}]+1}$ étant les observations de rangs $[\frac{n+1}{2}]$ et $[\frac{n+1}{2}]+1$ de échantillon initial.

Disposant des B répétitions, on peut déterminer la moyenne :

$$\hat{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*$$

et l'écart-type des $\hat{\theta}_b^*$:

$$\hat{\sigma}_{\hat{\theta}^*} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \hat{\theta}^*\right)^2}$$

L'écart-type $\hat{\sigma}_{\hat{\theta}^*}$ est une estimation de l'erreur-standard de l'estimateur du paramètre θ

Définition 5. *On appelle estimation bootstrap de l'écart-type* $\hat{\sigma}_F(\hat{\theta})$ *de* $\hat{\theta}$ *son estimation plug-in :*$\sigma_{\hat{F}}(\hat{\theta})$

Exemple 1. *À titre d'illustration, nous considérons le problème de l'estimation de diverses caractéristiques de la population, à partir d'un échantillon aléatoire et simple de 9 observations.*
La deuxième colonne du tableau suivant, notée X*, donne les premières observations de l'échantillon. Les colonnes suivantes donnent les premières et les dernières observations de 500 échantillons de 9 observations prélevés dans l'échantillon initial et notés* $(X_1^*, X_2^*, ..., X_{500}^*.)$

$$X = (12, 15, 42, 20, 32, 25, 36, 13, 40)$$

TABLE 2.1: *Échantillon initial et résultats de 500 rééchantillonnages (données partielles).*

Obs	X	X_1^*	X_2^*	X_2^*		X_{499}^*	X_{500}^*
1	15	40	42	13	...	20	36
2	32	20	15	40	...	12	42
3	36	27	32	12	...	25	12
...	..	...	..	...	...	...	...
9	25	13	40	25	..	36	42

TABLE 2.2: *Paramètres estimés pour l'échantillon initial*$(\hat{\theta})$ *et pour les trois premiers échantillons obtenus par rééchantillonnage* $(\hat{\theta}_1^*, \hat{\theta}_2^*$ *et* $\hat{\theta}_3^*)$ *,moyennes* $(\hat{\theta}^*)$*, et écarts-types* $(\hat{\sigma}_{\hat{\theta}^*})$ *;des paramètres estimés pour 500 rééchantillonnages*

Paramètre	Moyenne	Variance
$\hat{\theta}$	26.11111	123.4321
$\hat{\sigma}_{\hat{\theta}}$	3.927978	
$\hat{\theta}_1^*$	23.66667	25.00000
$\hat{\theta}_2^*$	19.11111	27.00000
$\hat{\theta}_3^*$	28.44444	24.33333
...	...	...
$\hat{\theta}_{499}^*$	27.66667	79.111111
$\hat{\theta}_{500}^*$	24.11111	77.580247
$\hat{\theta}^*$	26.15444	110.6078
$\hat{\sigma}_{\hat{\theta}^*}$	3.736612	

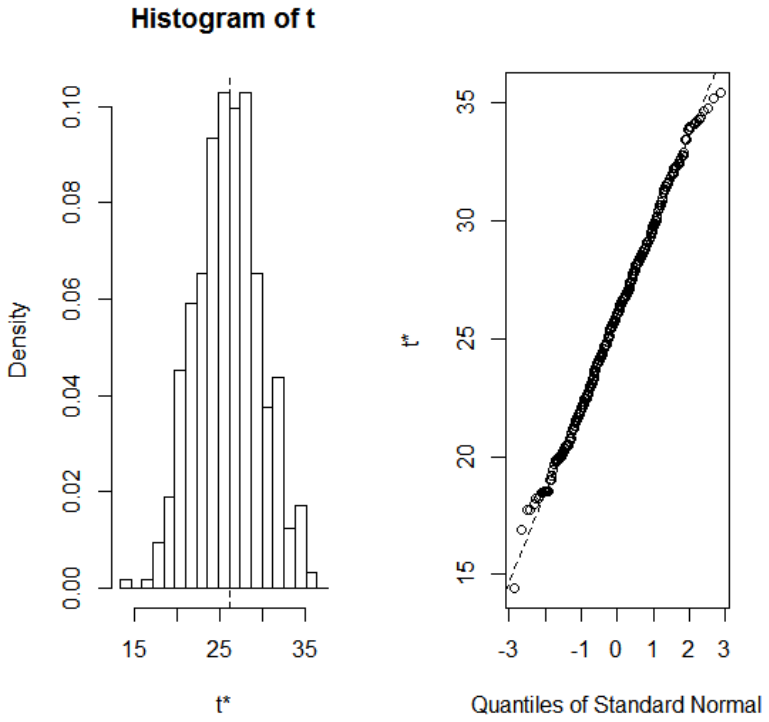


FIGURE 2.5: Distribution de la moyenne bootstrapée

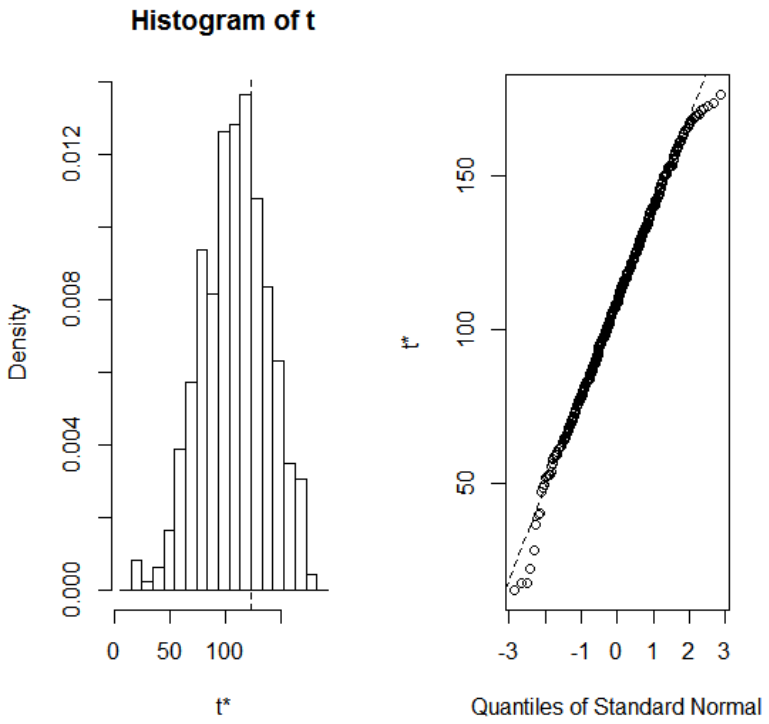


FIGURE 2.6: Distribution de la Variance Bootstrapée

Une approximation de l'estimateur bootstrap de l'écart-type de $\hat{\theta}$ est obtenue par une simulation (Monte-carlo) décrite dans l'algorithme ci-dessous. Pour un paramètre θ est un échantillon X donnés, on note $\hat{\theta} = s(X)$ l'estimation obtenue sur cet échantillon. Une réplication bootstrap de $\hat{\theta}$ est donnée par : $\hat{\theta}^* = s(X^*)$.

ALGORITHME : Estimation bootstrap de l'écart-type

1. Tirer B échantillons bootstrap $X_1^*, X_2^*, ..., X_B^*$ par tirage avec remise dans X.
2. Calculer la copie bootstrap $\hat{\theta}_b^* = s(X_b^*); b = 1, 2, ..., B$.
3. Calculer l'ecart-type de l'échantillon ainsi construit :

$$\hat{\sigma}_b^2 = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}_b^* - \hat{\theta}^*\right)^2$$

avec $\hat{\sigma}_b$ est l'approximation bootstrap de l'estimation plug-in recherchée de l'ecart-type de $\hat{\theta}$.

2.3.2 Methode de bootstrap double

Le bootstrap est une procédure qui consiste à choisir des échantillons aléatoires avec remplacement d'un ensemble de donnée et à analyser chaque échantillon de la même façon. Les méthodes bootstrap peuvent être utilisées pour une analyse entièrement paramétrique, semi-paramétrique et complètement non paramétrique. Hall (1990) et Danielsson et al. (2001) présentent la technique bootstrap et proposent la méthode bootstrap double [27] .
Soit

$$Q^{(i)}(n,k) = \frac{1}{k} \sum_{j=1}^{k} (log(X_{n-j+1}) - log(X_{n-k}))^i. \tag{2.5}$$

La méthode optimise k pour l'erreur quadratique moyenne de.

$$Q(n,k) = Q^{(2)}(n,k) - 2(Q^{(1)}(n,k))^2 \tag{2.6}$$

on peut voir que $\hat{\xi} = Q^{(1)}(n,k)$ est l'estimateur de Hill La méthode optimise $\hat{k^*}$ ou $u = X_{(n-k,n)}$ par les étapes suivantes :

1. Sélectionné uniformément $\epsilon \in (0, 1/2)$ et l'ensemble $n_1 = [n^{1-\epsilon}]$ Cela assure le Consistance si $n \to +\infty$.
   Évaluation $E[Q(n_1, r)^2 | X_1, ..., X_n]$ en ré-échantillonnant n_1 taille des échantillons bootstrap de la distribution empirique et le minimiser à r

$$\hat{k_1} = \arg\min_r (E[Q(n_1, r)^2 | X_1, ..., X_n])$$

2. l'ensemble $n_2 = [\frac{n_1^2}{n}]$ et minimiser $E[Q(n_1, r)^2 | X_1, ..., X_n]$ à r de la même manière qu'à la première étape.

$$\hat{k_2} = \arg\min_r(E[Q(n_2, r)^2 | X_1, ..., X_n])$$

3.

$$\hat{k} = \frac{\hat{k_1^2}}{\hat{k_1}}[\frac{log(k_1)^2}{(2log(n_1) - log(k_1))^2}]^{\frac{log(n_1)-log(k_1)}{log(k_1)}}$$

2.3.3 Metrique de Kolmogorov-Smirnov

Cette approche vise à minimiser la distance entre l'indice de la fonction de distribution empirique et la distribution Pareto ajustée avec le paramètre estimé de l'indice de queue. Daniel et al. (2016) ont proposé la distance pour les quantiles [27] .
Supposons que

$$P(X \geq x) = \bar{F}(x) = \mathbb{A}x^{-\frac{1}{\xi}} + o(x^{-\frac{1}{\xi}})$$

La fonction quantile peut être approximée par

$$q = \left[\frac{\bar{F}(x)}{\mathbb{A}}\right]^{-\xi}$$

La probabilité $\bar{F}(x)$ peut être remplacé par $\frac{j}{n}$,et ξ est estimé par l'estimateur de Hill pour certains k et $\mathbb{A}$ peut être estimée par $\frac{k}{n}(X_{n-k})^{\frac{1}{\xi}}$ Expert pour certains.Le quantile est ainsi estimé par

$$q(j, k) = \left[\frac{k}{j}(X_{n-k})^{\frac{1}{\xi}}\right]^{\xi}$$

L'optimal $\hat{k^*}$ pour l'estimateur de Hill qui réduit au minimum la distance entre le quantile empirique et le quantile estimé.

$$\hat{k^*} = \arg\min_k(\sup_{j\in\{1,...,S\}} |X_{(n-j,n)} - q(j, k)|)$$

Où $S > k$.
Alors le seuil optimal est $u^* = X_{(n-\hat{k^*},n)}$.

Chapitre 3

Simulations et Applications

3.1 Simulations

Le Matlab :est une langage de calcul scientifique de haut niveau et un environnement interactif pour le développement d'algorithme,la visualisation et l'analyse de donnée,ou encore le calcul numérique.en utilisant matlab,vous pouvez résoudre des problème scientifique.

Le logiciel R :est un logiciel de statistique créé par Ross Ihaka & Robert Gentleman. Il est à la fois un langage informatique et un environnement de travail : les commandes sont exécutées grâce à des instructions codées dans un langage relativement simple, les résultats sont affichés sous forme de texte et les graphiques sont visualisées directement dans une fenêtre propre. C'est un clone du logiciel S-plus qui est fondée sur le langage de programmation S, développée par les laboratoires Bell en 1988 . Ce logiciel sert à manipuler des données,à tracer des graphiques et à faire des analyses statistiques sur ces données. Le logiciel R constitue aujourd'hui un language de programmation intégré d'analyse statistique. Le site Internet http :// www. r-project. org, est la meilleure source d'informations sur ce logiciel.

Ce chapitre est reparti en deux parties, dans la première partie, nous allons faire une étude de simulation des méthodes étudier dans le chapitre président afin de comparer la performance de ces dernières.
Nous avons utilisé les Packages de R suivants :

- ☐ boot : pour le bootstrap dans les exemples basés sur des données simulées.
- ☐ evd, evir et ismev : pour la modélisation de GEVD et GPD.
- ☐ POT : pour la loi GPD

Dans la deuxième partie nous avons appliquée ces méthode sur des données météorologique.

3.1.1 Graphe de la durée de vie moyenne résiduelle

Il s'agit de chercher le seuil à partir du quel la fonction est linéaire.
Dans le logiciel R, elle est donnée par la fonction « MRL-plot».
La moyenne des excès peut être utilisée pour guider le choix du seuil adéquat u^*.
Nous avons simulé des échantillons de différents taille des lois GPD et GEV pour différents paramètres.L'étude de simulation est représentée dans les graphes suivants.

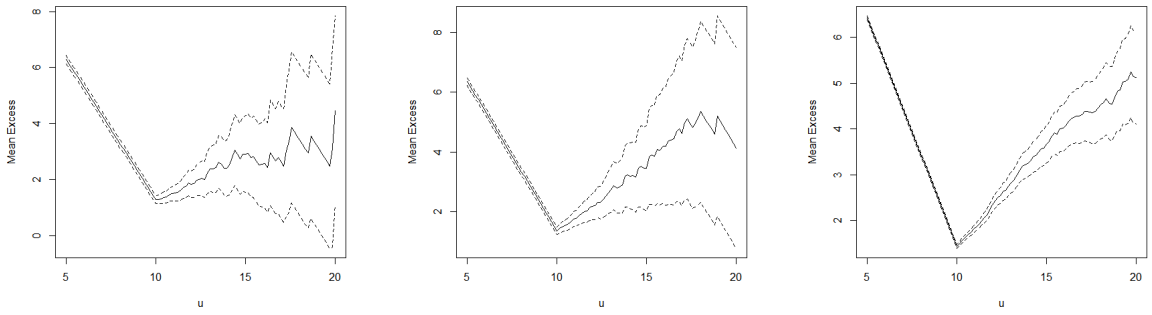


FIGURE 3.1: *mean residual life plot pour* $GPD(u = 10, \xi = 0.3, \sigma = 1)$ *et pour échantillon de taille (500,1000,10000)*

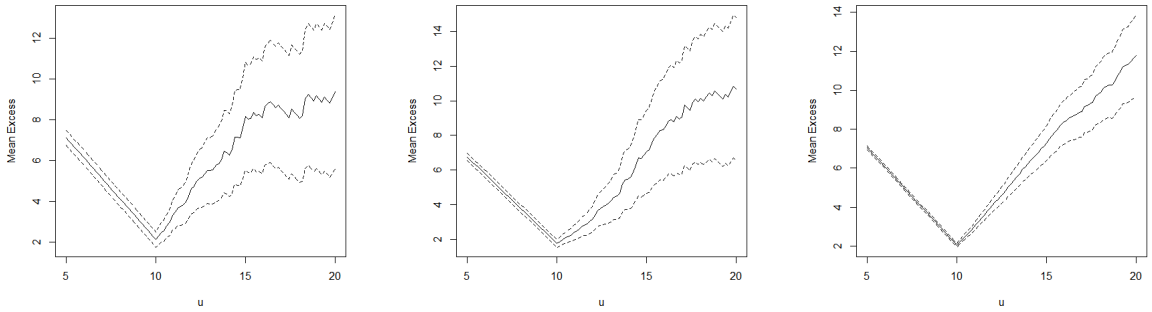


FIGURE 3.2: *mean residual life plot pour* $GPD(u = 10, \xi = 0.5, \sigma = 1)$ *et pour échantillon de taille (500,1000,10000)*

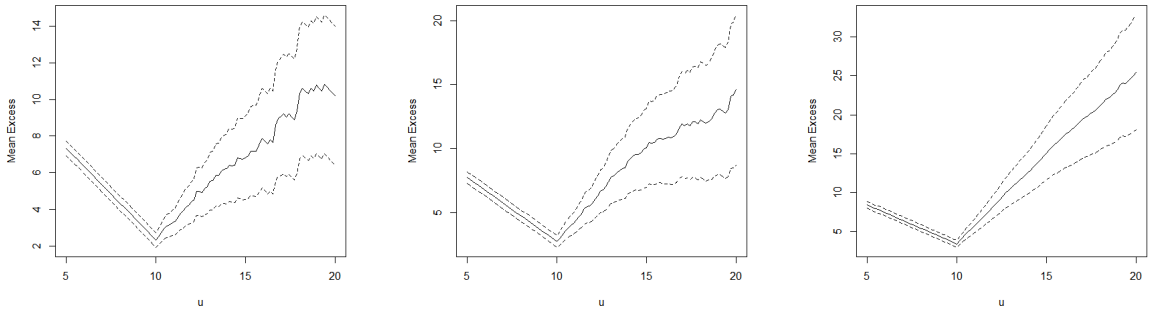


FIGURE 3.3: *mean residual life plot pour* $GPD(u = 10, \xi = 0.7, \sigma = 1)$ *et pour échantillon de taille (500,1000,10000)*

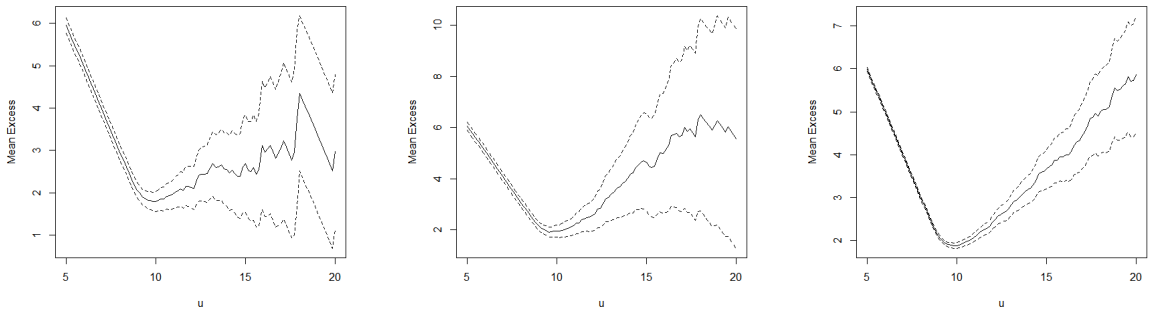


FIGURE 3.4: *mean residual life plot pour* $GEV(u = 10, \xi = 0.3, \sigma = 1)$ *et pour échantillon de taille (500,1000,10000)*

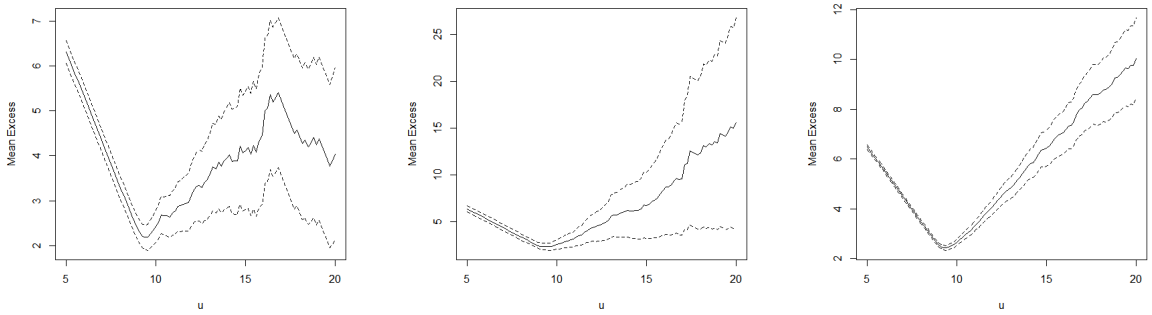


FIGURE 3.5: *mean residual life plot pour* $GEV(u = 10, \xi = 0.5, \sigma = 1)$ *et pour échantillon de taille (500,1000,10000)*

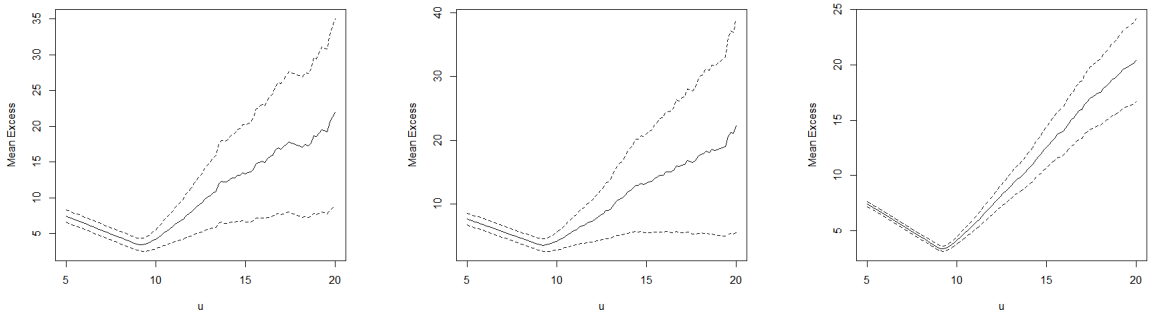


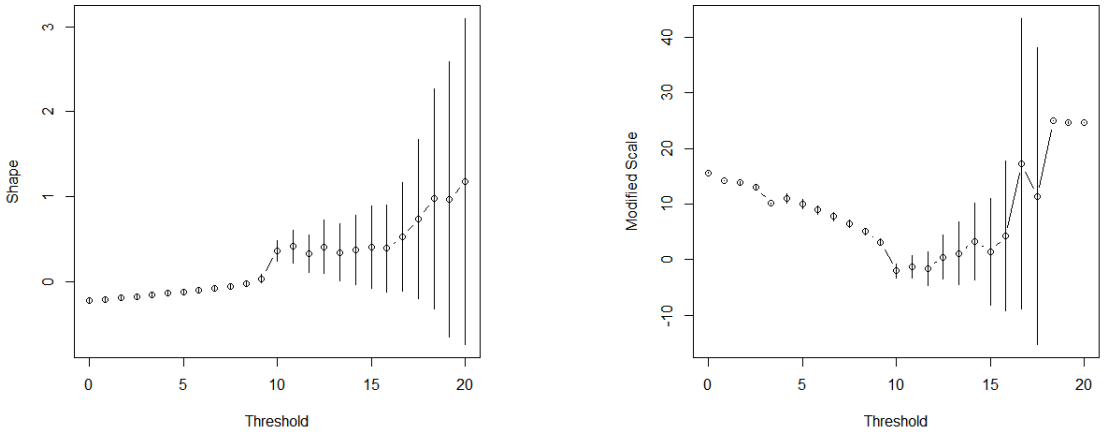
FIGURE 3.6: *mean residual life plot pour* $GEV(u = 10, \xi = 0.7, \sigma = 1)$ *et pour échantillon de taille (500,1000,10000)*

L'étude de simulation pour la sélection du seuil par la méthode MRL-plot,nous a révélé que la méthode estime bien le seuil.

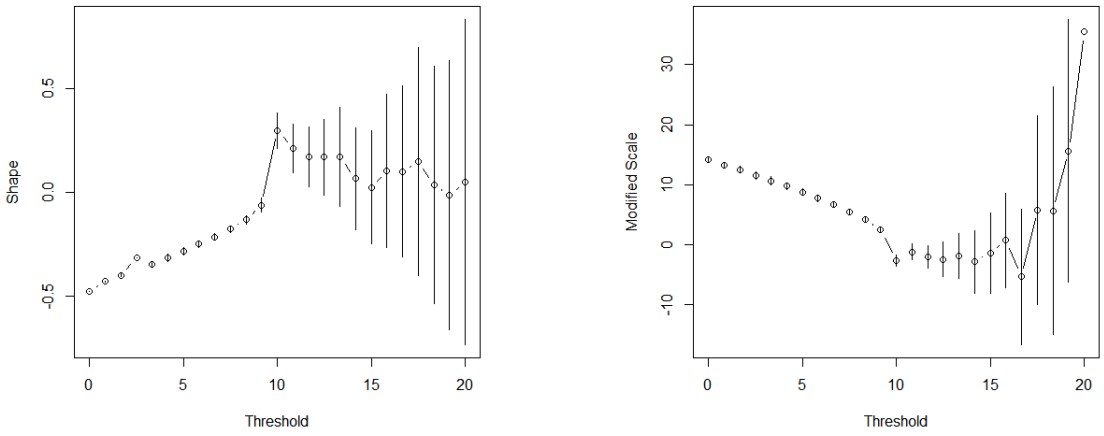
3.1.2 Graphe de Stabilité des paramètres de forme et d'échelle

Encore appelée « stable scale and shape parameters », cette méthode permet de déterminer un seuil requis en ajustant les données à une distribution de GPD en utilisant un seuil différent. La stabilité des paramètres (forme et échelle) peut alors être contrôlée et localisée.
Cette technique est implémentée dans le logiciel R avec le package (POT). Les graphes établaient un lien direct entre les valeurs des paramètres estimés (ξ et σ) et les seuils potentiels u^*. Les paramètres estimés au-dessus des seuils sont ceux pour lesquels le modèle GPD devient valable. L'étude de simulation est représentée dans les graphes suivants.

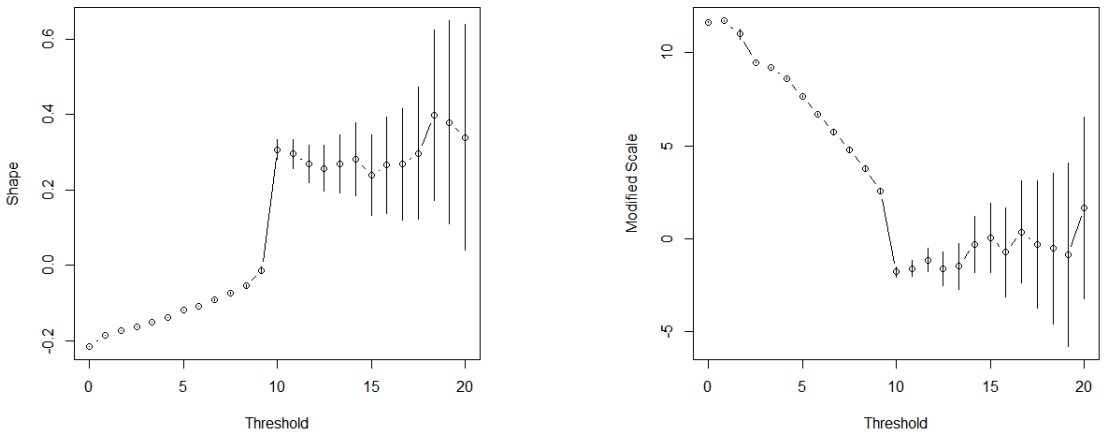
remarque
(a),(a') et (a") représentes les graphes de taille 500
(b),(b') et (b") représentes les graphes de taille 1000
(c),(c') et (c") représentes les graphes de taille 10000



(a)

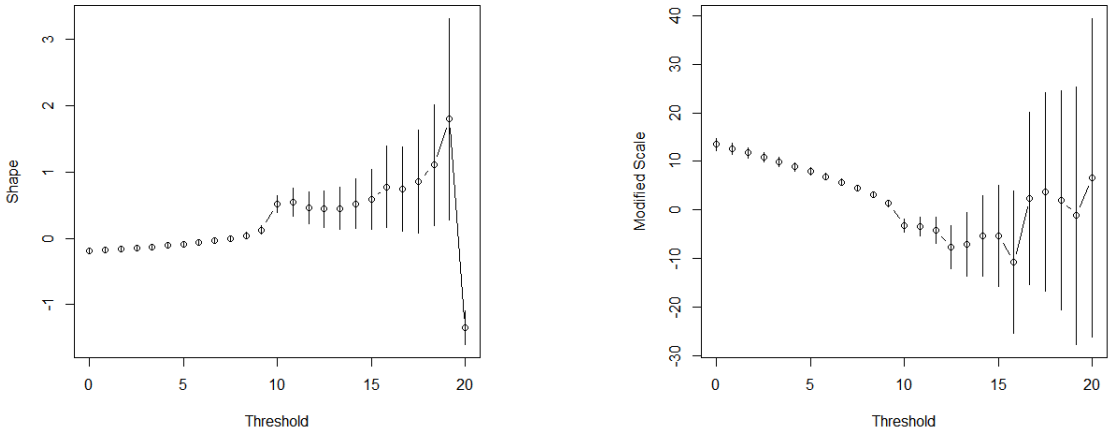


(b)

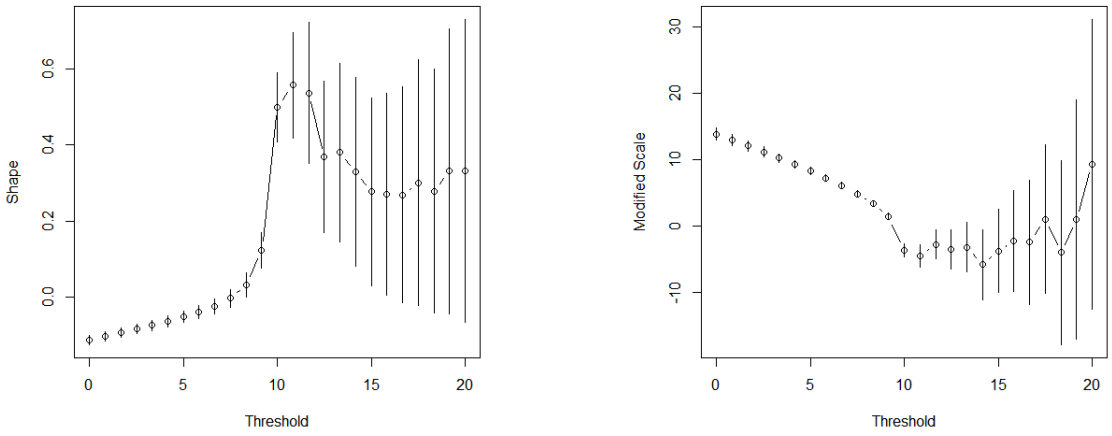


(c)

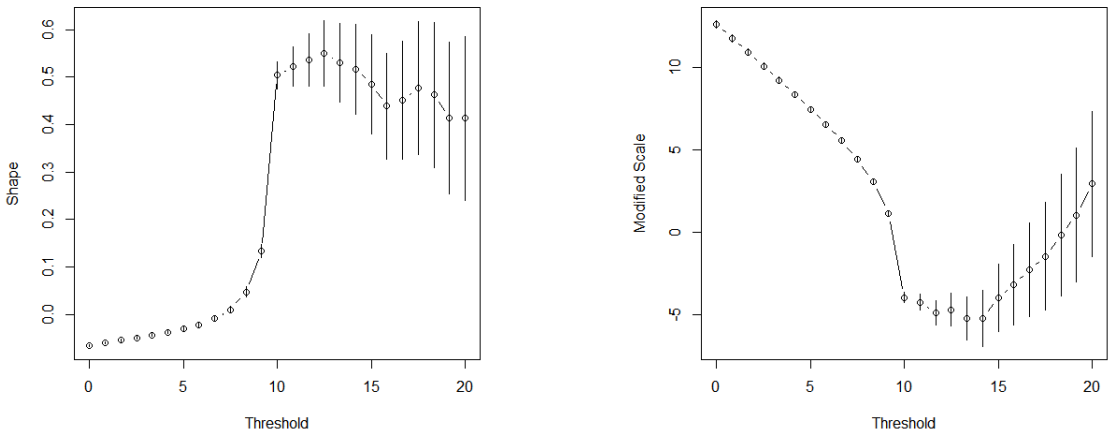
FIGURE 3.7: *Graphe de stabilité des paramètres de forme(a gauche) et d'échelle (a droite) modifié avec*($\xi = 0.3$ *)*



(a')

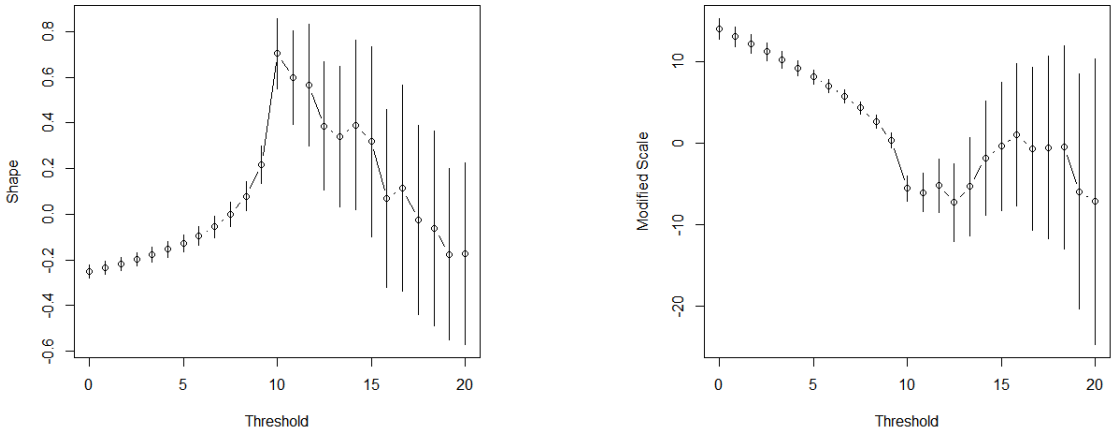


(b')

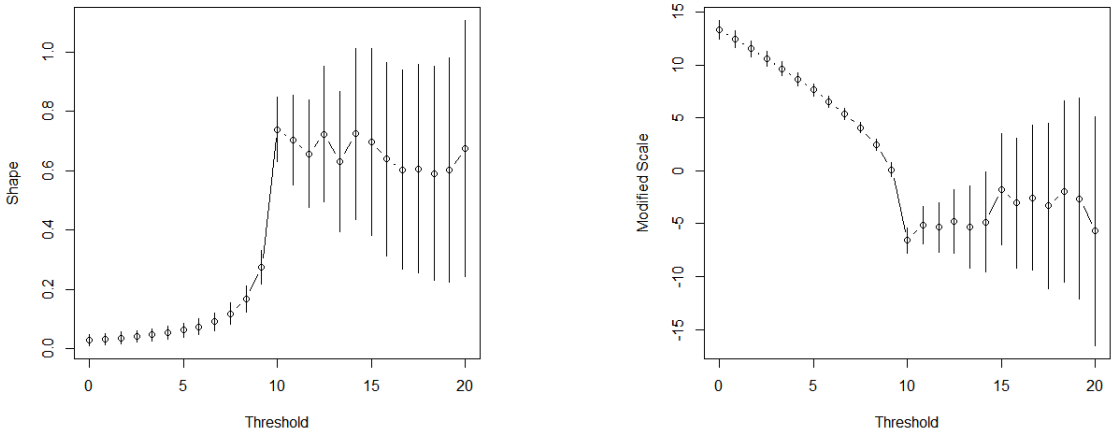


(c')

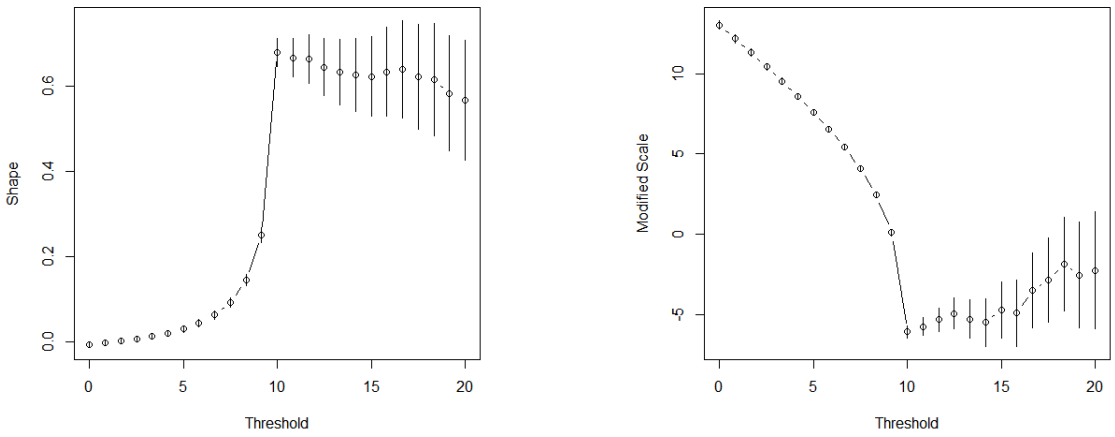
FIGURE 3.8: *Graphe de stabilité des paramètres de forme(a gauche) et d'échelle (a droite) modifié avec*($\xi = 0.5$ *)*



(a")



(b")



(c")

FIGURE 3.9: *Graphe de stabilité des paramètres de forme(a gauche) et d'échelle (a droite) modifié avec*($\xi = 0.7$ *)*

De même, nous avons remarqué que La stabilité linéaire combinée de ces représentations nous permet de prendre un seuil égal à 10 pour la modélisation.

3.1.3 Hill-plot

La méthode de Hill-plot est aussi validé par simulation. Le seuil sera le point de stabilité du graphe.

Nous avons simulé des échantillons de différents taille des lois GPD et GEV pour différents paramètres.L'étude de simulation est représentée dans les graphes suivants.

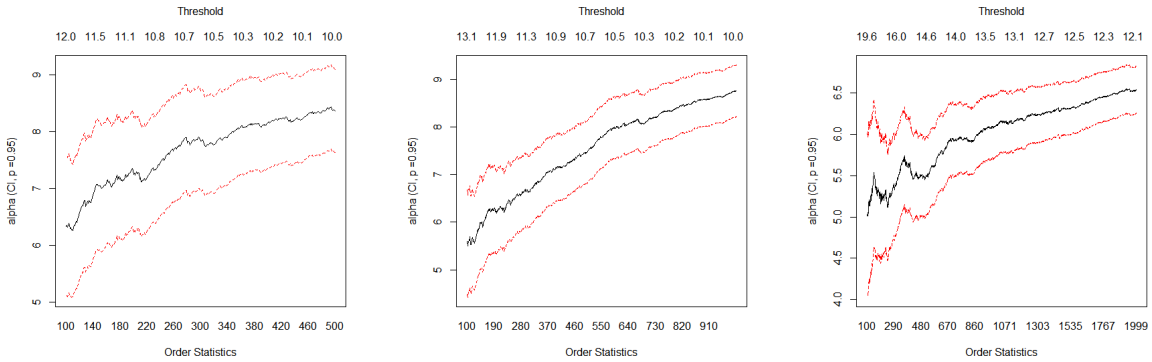


FIGURE 3.10: *Estimateur de Hill en fonction du seuil et du nombre d'excès pour (**G**PD) considérés (n=(500,1000,10000), $\xi = 0.3$)*

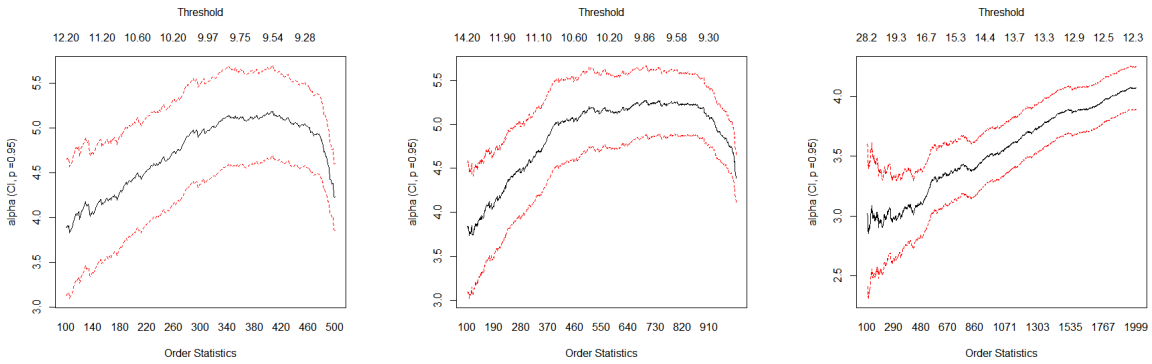


FIGURE 3.11: *Estimateur de Hill en fonction du seuil et du nombre d'excès pour (**G**PD) considérés (n=(500,1000,10000), $\xi = 0.5$)*

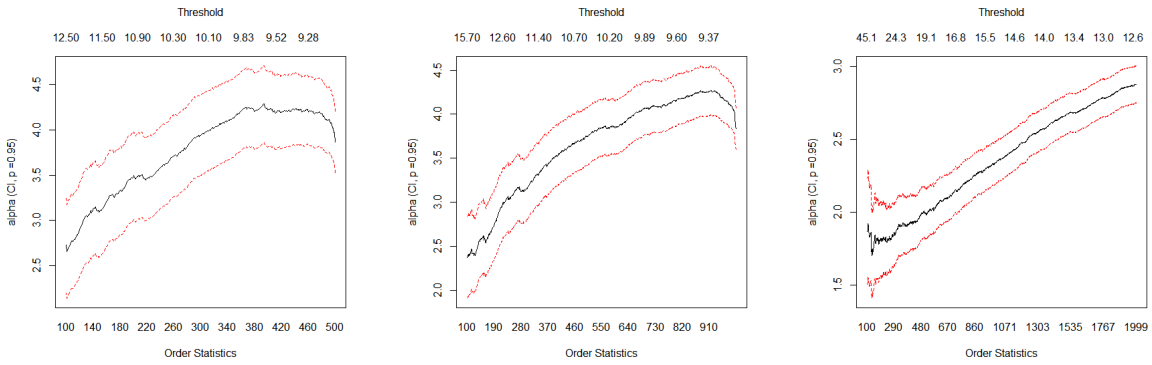


FIGURE 3.12: *Estimateur de Hill en fonction du seuil et du nombre d'excès pour (**GPD**) considérés (n=(500,1000,10000), $\xi = 0.7$)*

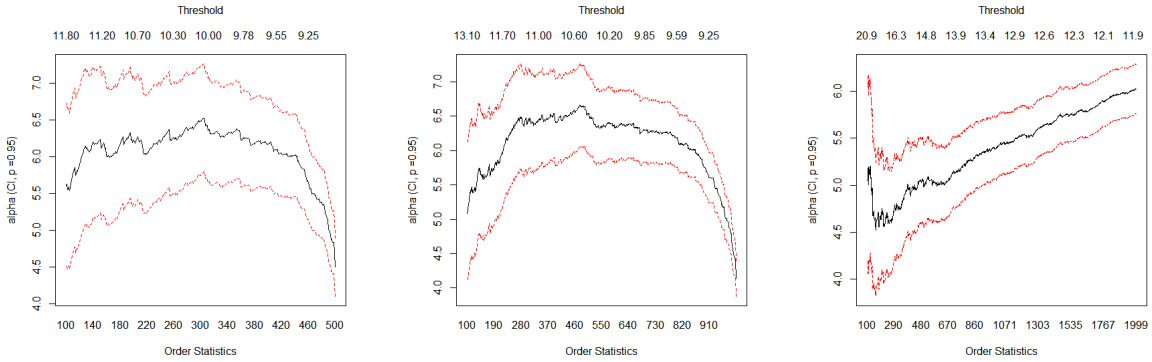


FIGURE 3.13: *Estimateur de Hill en fonction du seuil et du nombre d'excès pour (**GEV**) considérés (n=(500,1000,10000), $\xi = 0.3$)*

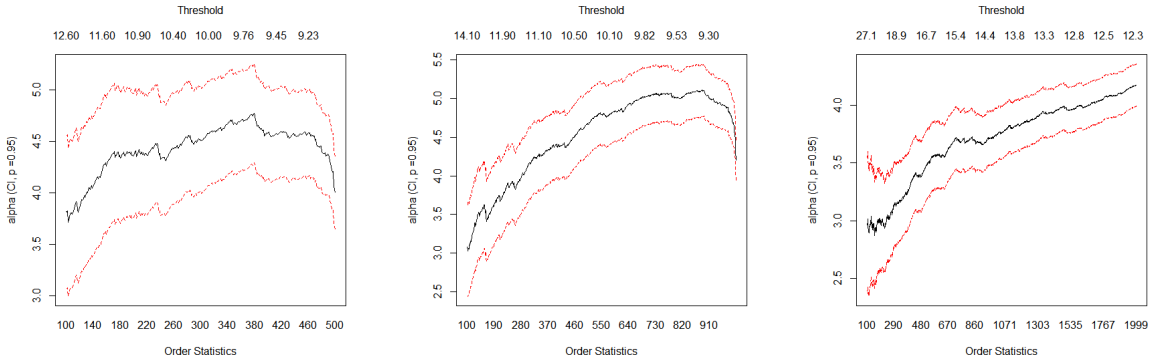


FIGURE 3.14: *Estimateur de Hill en fonction du seuil et du nombre d'excès pour (**GEV**) considérés (n=(500,1000,10000), $\xi = 0.5$)*

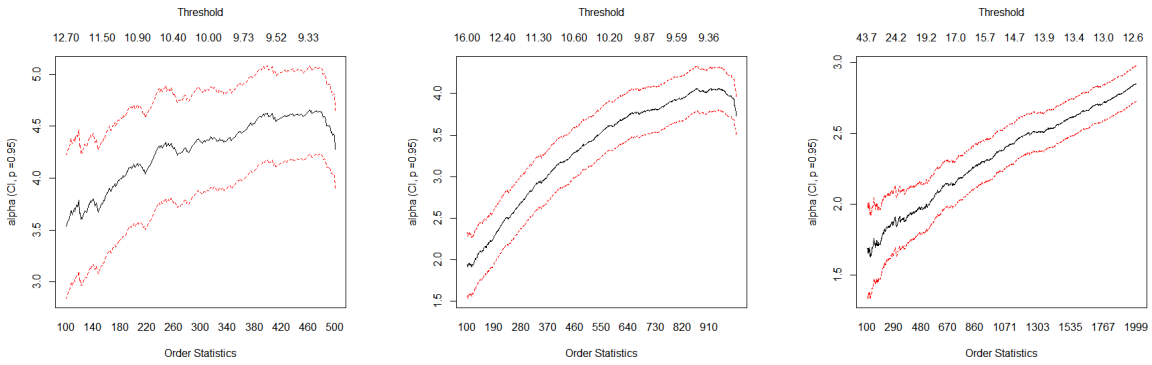


FIGURE 3.15: *Estimateur de Hill en fonction du seuil et du nombre d'excès pour (**GEV**) considérés (n=(500,1000,10000), $\xi = 0.7$)*

Après simulation, nous pouvons confirmer que cette méthode graphique détecte bien le seuil de la loi.

3.1.4 Méthodes Analytiques

Les méthodes numériques à savoir La distance de Kolmongorov et double bootstap sont étudiées par simulation pour la loi GPD pour différents tailles et paramètres. Les résultats sont données dans les tableaux suivants :

a. Metrique de Kolmongorov

TABLE 3.1: *K.S Metric pour u=100*

	n=500	n=1000	n=5000	n=10000
$\xi = 0.3$	100.3794	103.6869	100.8472	102.0041
$\xi = 0.5$	100.9990	100.3415	100.3507	100.1178
$\xi = 0.7$	101.3970	100.7246	100.4318	100.8451

TABLE 3.2: *K.S Metric pour u=10*

	n=500	n=1000	n=5000
$\xi = 0.3$	12.6704	11.3685	11.9999
$\xi = 0.5$	8.7111	9.6403	9.6305
$\xi = 0.7$	9.5897	9.5329	9.5444

b. Double Bootstrap

TABLE 3.3: *double bootstrap pour u=100*

	n=500	n=1000	n=5000
$\xi = 0.3$	100.0272	100.0372	100.0051
$\xi = 0.5$	100.3004	100.0733	100.0001
$\xi = 0.7$	100.0051	100.0001	100.0003

TABLE 3.4: *double bootstrap pour u=10*

	n=500	n=1000	n=5000
$\xi = 0.3$	10.0428	10.0118	10.0016
$\xi = 0.5$	10.1204	10.0110	10.0017
$\xi = 0.7$	10.4842	10.0766	10.0054

l'étude de simulation nous a montré que les méthodes estiment bien le seuil, mais la méthode de double boostrap nécessite plus de temps d'exécution.

3.2 Applications aux données Météorologiques

3.2.1 Zone d'étude et données utilisées

Les données utilisées sont fournies par l'office national météorologique (ONM) d'Alger. Nous nous sommes intéressés à la station météorologique de Dellys .
Les données climatiques utilisées dans cette étude sont la température, l'humidité, la précipitation et la vitesse du vent. il s'agit des moyennes mensuelles sur une durée de 13 ans (156 mois).
Lorsque nous disposons d'un échantillon de données à analyser, notre première préoccupation est celle de savoir si ce dernier est fiable, c'est-à-dire s'il ne contient pas de fausses mesures. La détection de données aberrantes est d'une première importance pour l'analyse des valeurs extrêmes (maxima et minima d'un échantillon).
Il est donc important de détecter les données aberrantes pour ne pas les confondre avec les valeurs extrêmes. Il est également important de montrer que les données sont indépendantes du temps.

3.2.2 Analyse des données

Le tableau suivant donnée un résumé d'analyse descriptive :

TABLE 3.5: *Résultats de l'analyse descriptive de la Températures, de l'Humidité,de la Précipitation et de la vitesse du vent moyennes mensuelles de durée 156 mois*

	Température	**Humidité**	**Précipitation**	**Vitesse du vent**
Max	27.6	86.00	308.70	5.900
Min	10.10	9	0.00	1.600
Moyenne	18.22	69.47	58.23	3.097
Variance	18.76356	258.2506	3905.701	0.6417316
Ecart-type	4.331692	16.07018	62.70711	0.8010815
Mediane	18.20	73.00	38.45	3.00
Skewness	0.2349155	-3.019081	1.674318	0.7970614
Kurtosis	-1.018806	8.646657	2.792674	0.9609168

Les données de Températures, de Précipitations, l'Humidité et de la Vitesse du Vent ont été traitées sous **matlab** (double bootstrap et KS.metric) et sous **R** (méthode mrlplot et tcplot et Hill-plot)

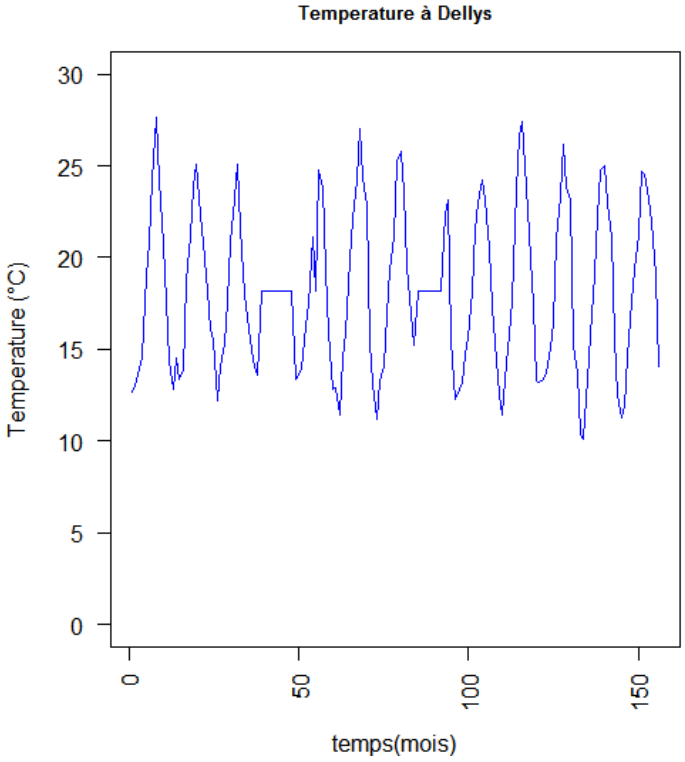


FIGURE 3.16: Distribution des moyennes mensuelles de la **Température** pour une durée de 13 ans

3.2.3 Estimation du seuil de la Température moyenne mensuelle

Le graphe de la durée de vie moyenne résiduelle a été utilisée dans cette étude pour déterminer l'intervalle dans lequel pourrait se situer le seuil.
Le but de l'étude des variations de Températures moyennes mensuelles, de 156 mois est de caractériser l'évolution générale du climat durant cette période dans la région de notre étude.
La figure (3.17) présente la Durée de vie moyenne résiduelle des températures mensuelles.

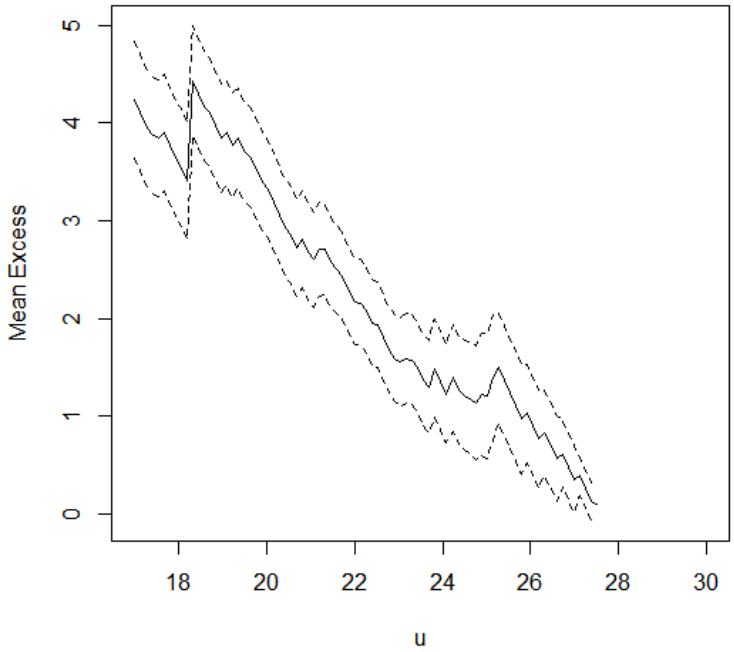


FIGURE 3.17: Fonction moyenne des excès de **Température** moyennes mensuelles de "Dellys" avec un intervalle de confiance à 95 %.

On observe une linéarité sur le graphe de la Durée de vie moyenne résiduelle entre 18 et 19

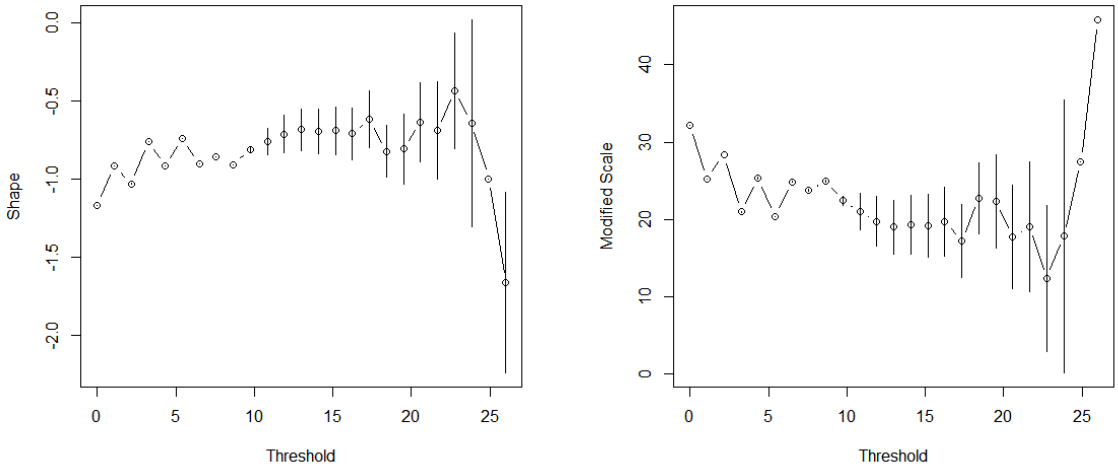


FIGURE 3.18: *Évolution du paramètre d'échelle modifié (courbe a droite) et du paramètre de forme (courbe a gauche) en fonction du seuil des* ***Températures*** *moyennes mensuelles à Dellys avec un intervalle de confiance à 95 %.*

Sur La figure (3.18) on remarque que la stabilité est plus forte sur l'intervalle [18-20] .

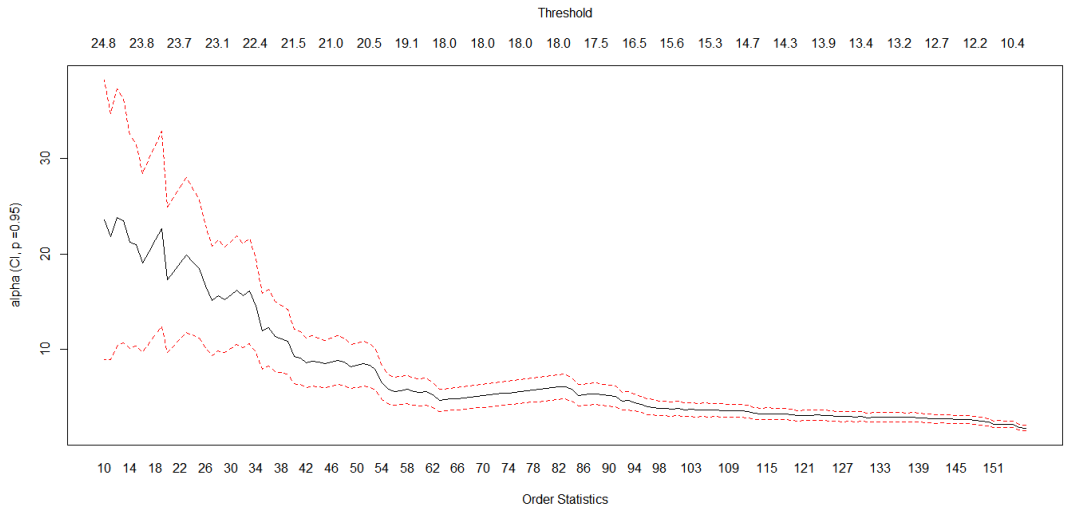


FIGURE 3.19: Hill-plot de Température moyenne mensuelle

Nous remarquons une **zone de stabilité** entre 60 et 82 excès,correspond à un seuil de 18. Au delà de 82 excès, l'estimateur n'est plus du tout stable.

L'ajustement des données au modèle a été réalisé en prenant 18.0C comme seuil, ce qui permet donc de détecter le nombre de données extrêmes (nombre des excès) de l'échantillon à partir duquel l'ajustement sera fait suivant une loi de GPD (Generalized Pareto Distribution). Ce seuil a été choisi en utilisant MRL-plot , Hill-plot et les graphes des paramètres d'échelle et de forme en fonction des différents seuils.

Le tableau suivant résume la valeur du seuil estimé par les cinq méthodes.

TABLE 3.6: *Seuil obtenue par différent méthodes*

	mrlplot	**tcplot**	Hill-plot	**Double bootstrap**	**Ks.metric**
μ	[18-19]	[18-20]	[18-19.1]	18.7	18.2000

3.2.4 Estimation du seuil de la Précipitation moyenne mensuelle

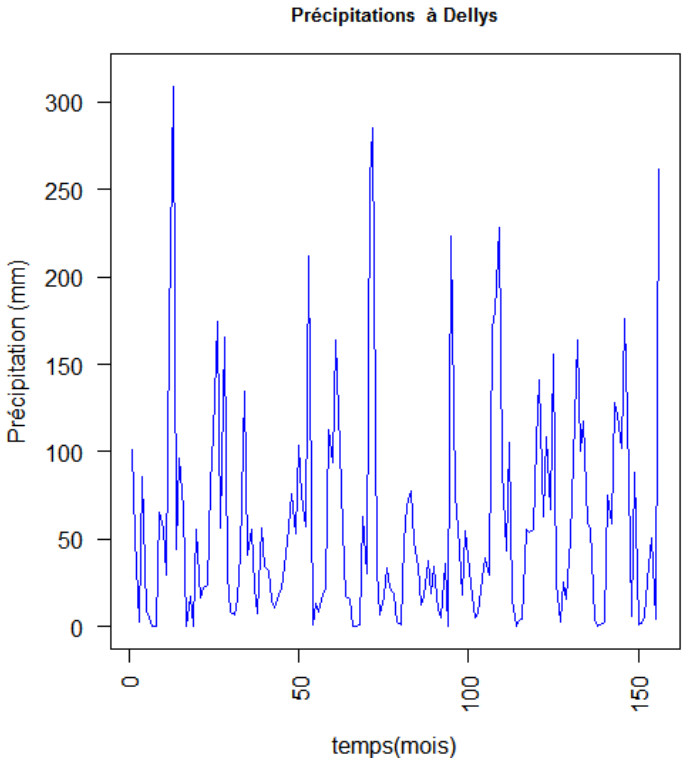


FIGURE 3.20: Distribution des moyennes mensuelles de **Précipitation** pour une duree de 156 mois

La figure (3.21) présente la Durée de vie résiduelle moyenne des Précipitation mensuelles.

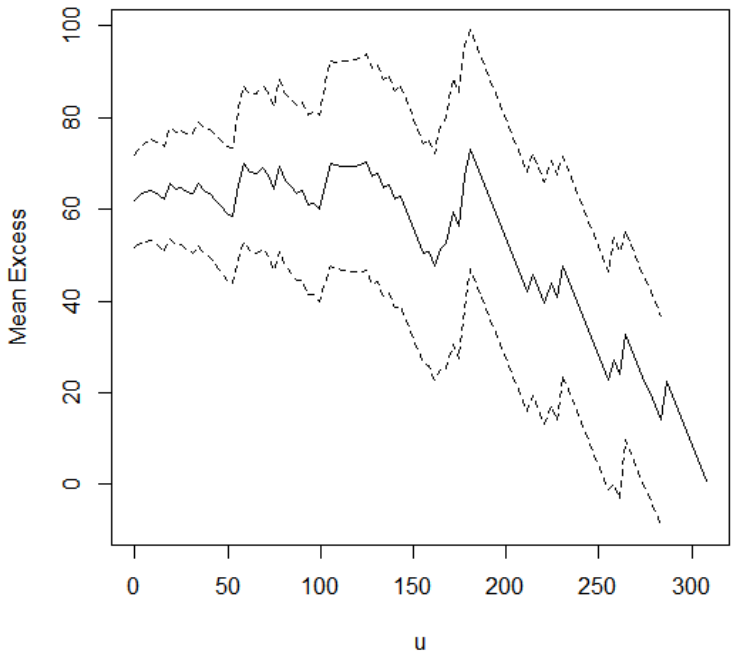


FIGURE 3.21: Fonction moyenne des excès de **Précipitation** mensuelles de Dellys avec un intervalle de confiance à 95 %.

On observe une linéarité sur le graphe de la Durée de vie moyenne résiduelle entre 95 et 105

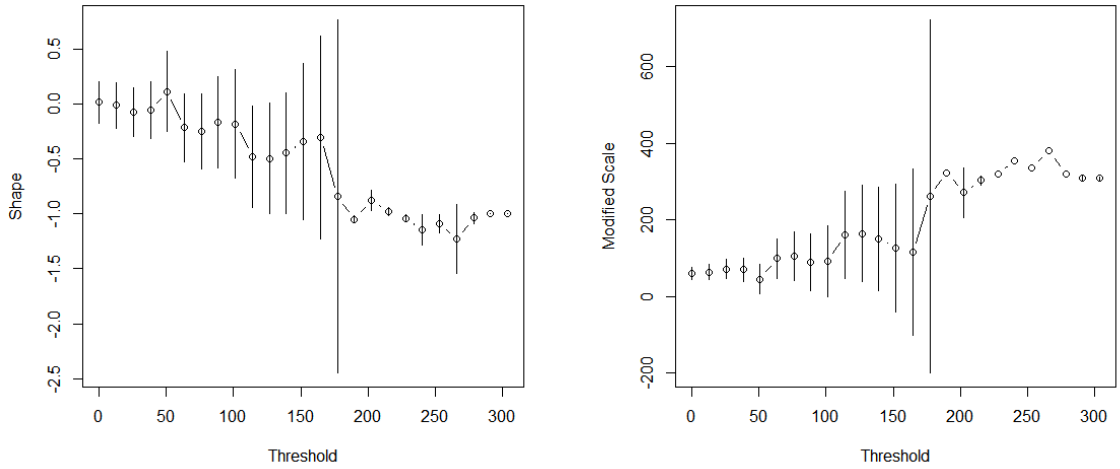


FIGURE 3.22: *Évolution du paramètre d'échelle modifié (courbe a droite) et du paramètre de forme (courbe a gauche) en fonction du seuil des* ***Précipitation*** *moyennes mensuelles avec un intervalle de confiance à 95 %.*

Sur La figure (3.22) on remarque que la stabilité est plus forte sur l'intervalle [100-110]

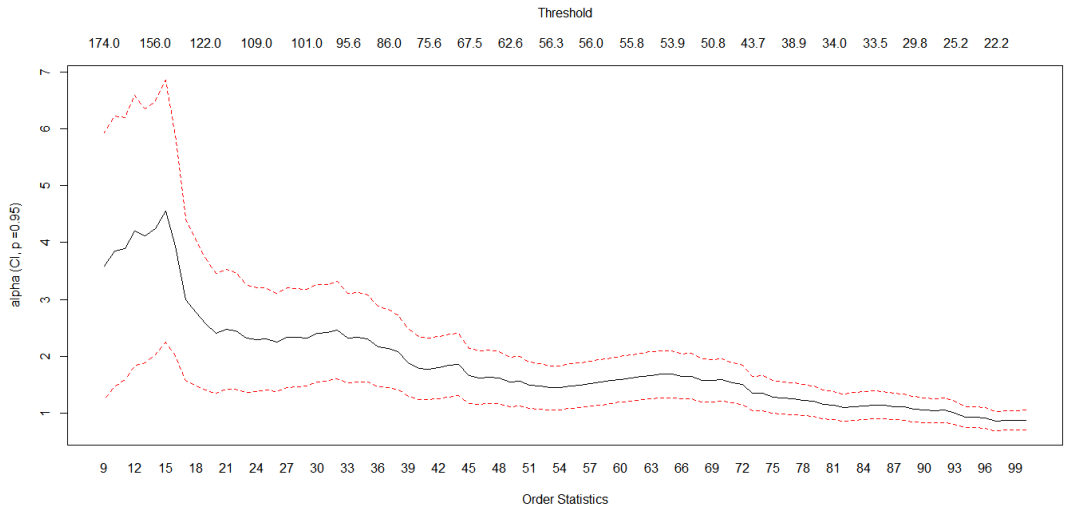


FIGURE 3.23: Hill-plot de **Précipitation** mensuelle

Nous remarquons une **zone de stabilité** entre 22 et 36 excès,correspond à un seuil entre [100-110] . Au delà de 36 excès, l'estimateur n'est plus du tout stable.

L'ajustement des données au modèle a été réalisé en prenant 100 comme seuil, ce qui permet donc de détecter le nombre de données extrêmes (nombre des excès) de l'échantillon à partir duquel l'ajustement sera fait suivant une loi de GPD (Generalized Pareto Distribution). Ce seuil a été choisi en utilisant MRLplot, Hill-plot et graphes des paramètres d'échelle et de forme en fonction des différents seuils.
Le tableau suivant résume la valeur du seuil estimé par les cinq méthodes.

TABLE 3.7: *Seuil obtenue par différent méthode*

	mrl-plot	**tc-plot**	**Hill-plot**	**Double bootstrap**	**Ks.métrique**
μ	[95-105]	[100-110]	[100-110]	58.5	100.9000

3.2.5 Estimation du seuil de l'Humidité moyennes mensuelles

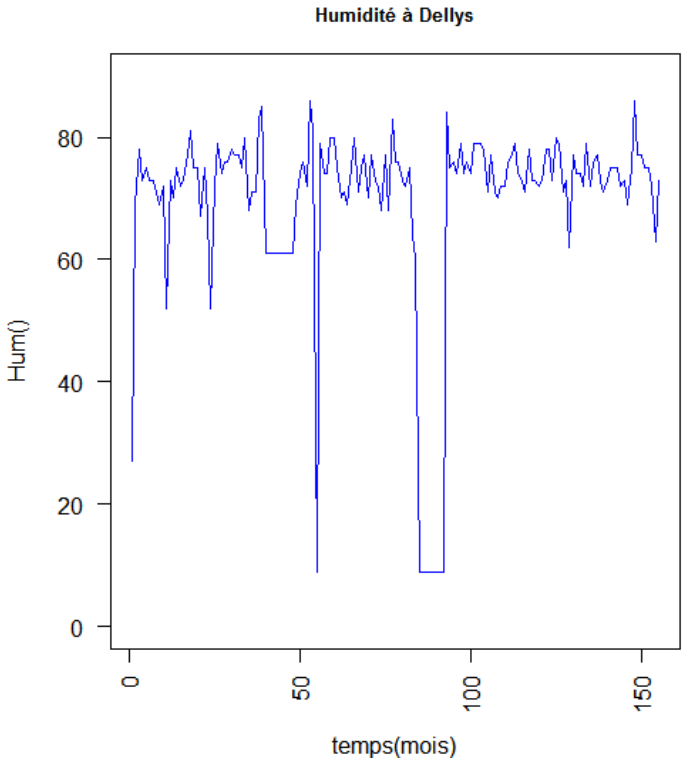


FIGURE 3.24: Distribution des moyennes mensuelles de **l'Humidité** pour une durée de 156 mois)

La figure (3.25) présente la Durée de vie résiduelle moyenne de l'Humidité mensuelles.

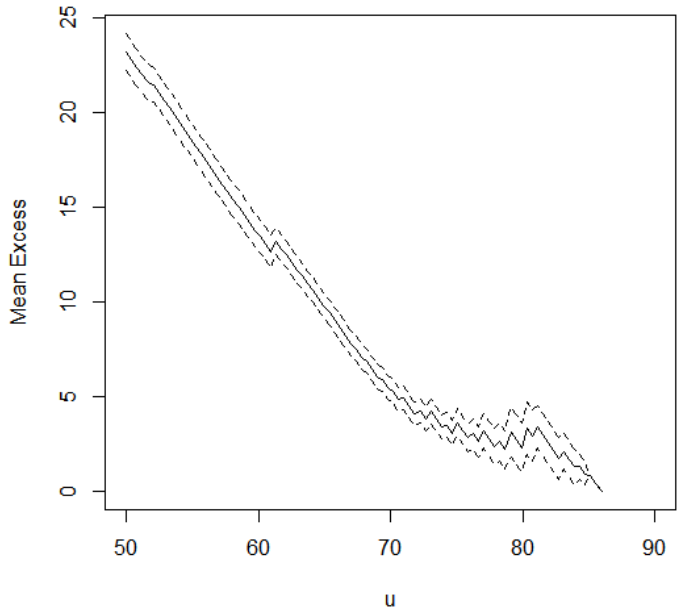


FIGURE 3.25: Fonction moyenne des excès de **l'Humidité** mensuelles de Dellys avec un intervalle de confiance à 95 %

On observe une linéarité sur le graphe de la Durée de vie moyenne résiduelle entre 73 et 85 et entre 60 et 65

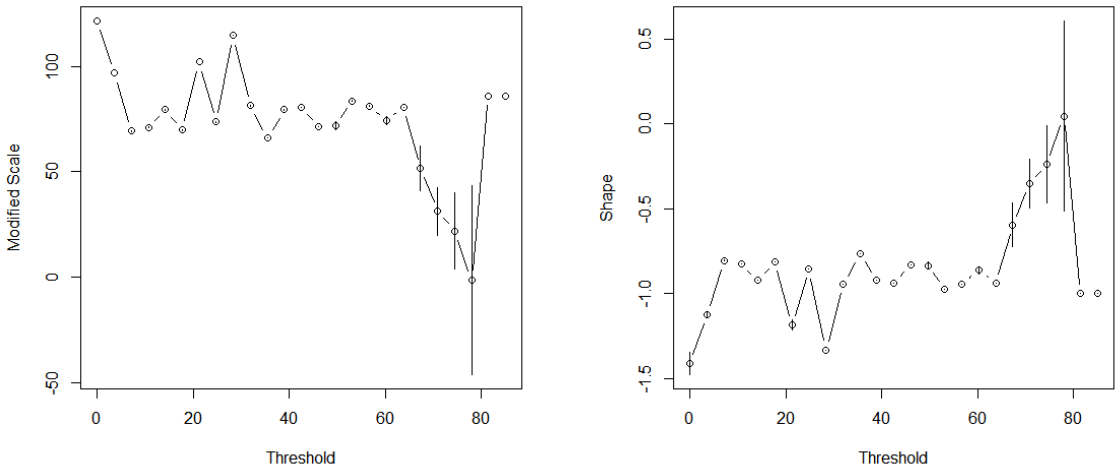


FIGURE 3.26: *Évolution du paramètre d'échelle modifié (courbe a gauche) et du paramètre de forme (courbe a droite) en fonction du seuil des* ***Humidité*** *moyennes mensuelles avec un intervalle de confiance à 95 %.*

Sur La figure (3.26) on remarque que la stabilité est plus forte sur l'intervalle [75-83]

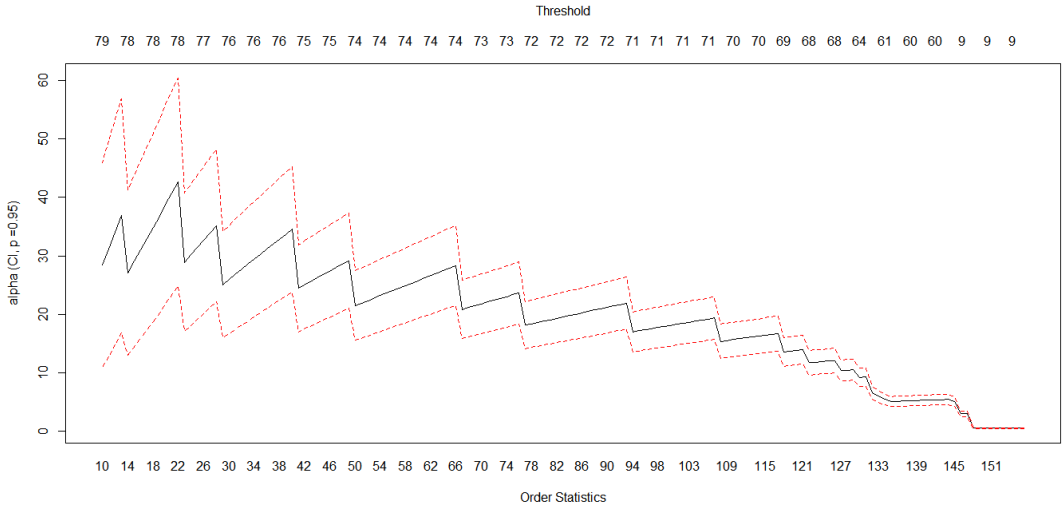


FIGURE 3.27: Hill-plot de l'humidité

Nous remarquons une **zone de stabilité** entre 134 et 145 excès,correspond à un seuil entre [60-68] . Au delà de 145 excès, l'estimateur n'est plus du tout stable.

L'ajustement des données au modèle a été réalisé en prenant 73 comme seuil, ce qui permet donc de détecter le nombre de données extrêmes (nombre des excès) de l'échantillon à partir duquel l'ajustement sera fait suivant une loi de GPD (Generalized Pareto Distribution). Ce seuil a été choisi en utilisant MRLplot, Hill-plot et des graphes des paramètres d'échelle et de forme en fonction des différents seuils.

Le tableau suivant résume la valeur du seuil estimé par les cinq méthodes.

TABLE 3.8: *Seuil obtenue par différent méthode*

	mrlplot	**tc-plot**	**Hill-lot**	**Double bootstrap**	**Ks.metric**
μ	[60-65] et [73-85]	[75-83]	[60-68]	70	72

3.2.6 Estimation du seuil de la Vitesse du Vent moyenne mensuelle

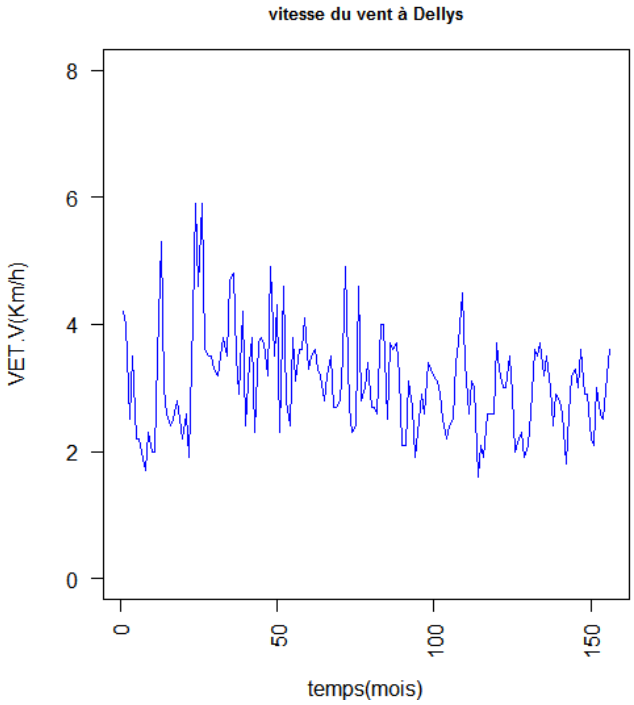


FIGURE 3.28: Distribution des moyennes mensuelles de **vitesse du vent** pour une durée de 156mois

La figure (3.29) présente la Durée de vie moyenne résiduelle de Vitesse du Vent mensuelles.

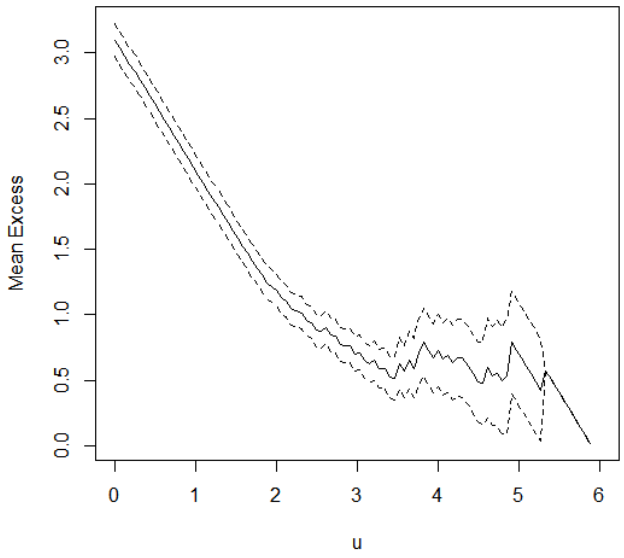


FIGURE 3.29: Fonction moyenne des excès de **Vitesse de vent** mensuelles avec un intervalle de confiance à 95 %

On observe une linéarité sur le graphe de Durée de vie moyenne résiduelle entre 3.4 et 3.9.

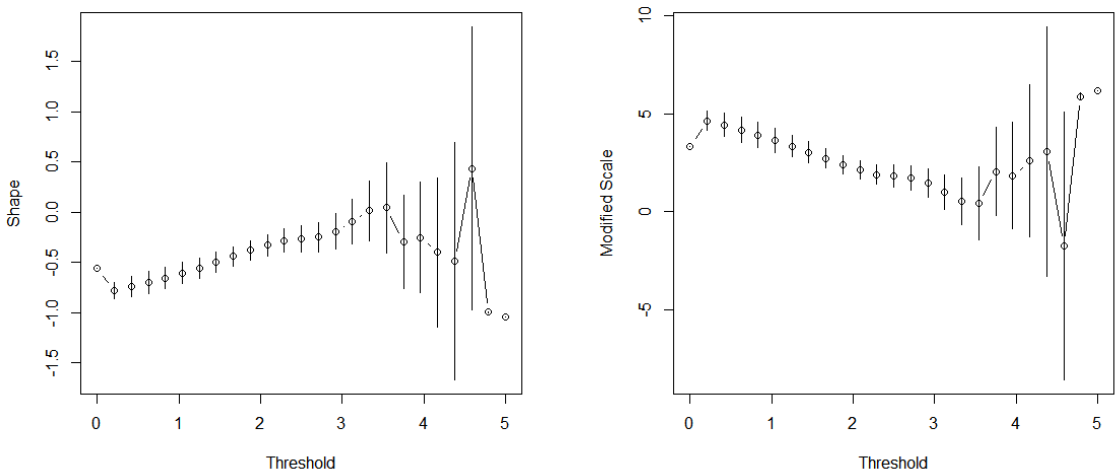


FIGURE 3.30: *Évolution du paramètre d'échelle modifié (courbe a droite) et du paramètre de forme (courbe a gauche) en fonction du seuil de* ***vitesse du vent*** *moyenne mensuelle avec un intervalle de confiance à 95 %.*

Sur La figure (3.30) on remarque que la stabilité est plus forte sur l'intervalle [3.8-4]

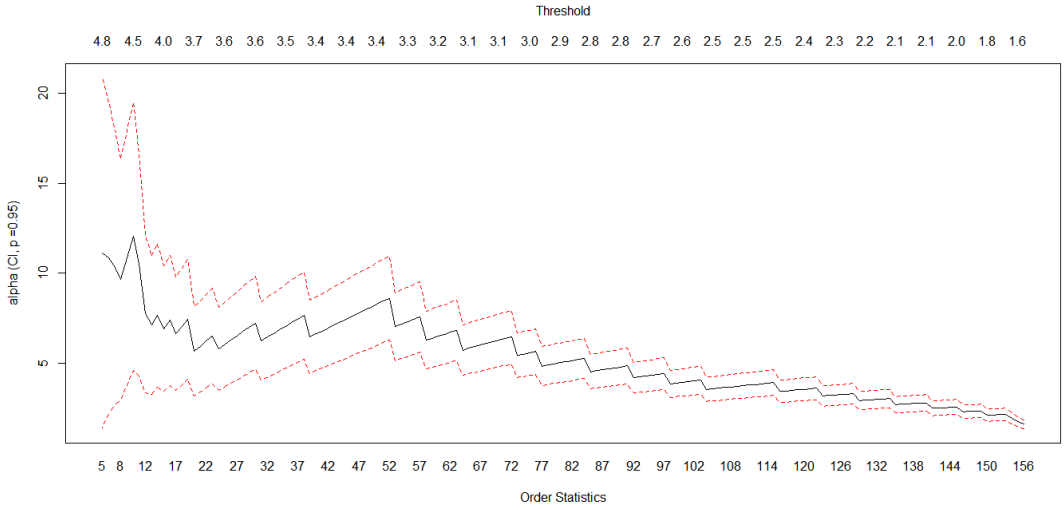


FIGURE 3.31: Hill-plot de vitesse du vent

Nous remarquons une **zone de stabilité** entre 20 et 27 excès,correspond à un seuil entre [3.6-3.8] . Au delà de 27 excès, l'estimateur n'est plus du tout stable.

L'ajustement des données au modèle a été réalisé en prenant 3 comme seuil, ce qui permet donc de détecter le nombre de données extrêmes (nombre des excès) de l'échantillon à partir duquel l'ajustement sera fait suivant une loi de GPD (Generalized Pareto Distribution). Ce seuil a été choisi en utilisant MRLplot, Hill-plot et graphes des paramètres d'échelle et de forme en fonction des différents seuils.

Le tableau suivant résume la valeur du seuil estimé par les cinq méthodes.

TABLE 3.9: *Seuil obtenue par différent méthode*

	mrlplot	**tc-plot**	**Hill-plot**	**Double bootstrap**	**Ks.metric**
μ	[3.4-3.9]	[3.8,4]	[3.6-3.8]	3.1	2.5000

Conclusion Générale

Les phénomènes rares qui causent des catastrophes suscitent l'étude des événements extrêmes tels que les sécheresses, les vagues de chaleur, les inondations, etc., est le fait de leur caractère imprévisible et des préjudices causés sur la société. Notre compréhension du comportement moyen du climat et de sa variabilité s'est beaucoup amélioré ces dernières décennies. Par contre il est toujours difficile de comprendre les événements climatiques extrêmes et encore plus dur de les prévoir, puisqu'ils sont rares et suivent des lois statistiques différentes de celles des moyennes.

La sélection du seuil optimal est une question non négligeable, car le seuil choisi doit être suffisamment élevé pour assurer la validité de l'hypothèse de GPD ,et en même temps, maintenir une quantité suffisante de données pour éviter une augmentation inutile de la variance d'estimation des paramètres de distribution des GPD.

En littérature de la théorie des valeurs extrêmes nous rencontrons beaucoup de méthodes utilisées pour analyser les dépassements sur un seuil élevé qui sont en forte demande en économie, en sciences de l'environnement et dans d'autres domaines. La distribution généralisée de Pareto(GPD) a été largement utilisé pour adapter les observations dépassant un seuil.

Nous avons fait une étude de comparaison entre ces méthodes par le biais de simulation.
L'étude de simulation nous a montré que ces méthodes estiment bien le seuil.
Nous avons appliqué les méthodes d'estimation du seuil aux données météorologiques à savoir la Température,l'Humidité, la Précipitation et le vitesse du vent de la région de Dellys sur une durée de 13 ans.

L'approche POT est largement utilisée mais elle est très sensible au choix du seuil pour obtenir une bonne approximation des excès au-dessus du seuil considéré

Parmi les perspectives de ce travail est d'appliquer ces méthodes pour les valeurs extrêmes censurées et dans le cas multivariées.

Bibliographie

[1] A.Amar, *Apport de la théorie des valeurs extrêmes à la modélisation et la gestion des risques bourisiers, financiers et hydro-météorologiques*, Thèse doctorat. Université Mohammed V-Agdal, 2014 (19-44).

[2] A.Borchani, *Statistiques des valeurs extrêmes dans le cas de lois discrètes*, Document de recherche ESSEC/Centre de recherche de l'ESSEC ISSN. École Supérieure de la Statistique et de l'Analyse de l'Information, Tunis, December 2010. [3-14]

[3] A.Langousis,A.Mamalakis,M.Pulig, *detection for the generalized Pareto distribution : Review of representative methods and application to the NOAA NCDC daily rainfall database*Water Resources, 2016 - Wiley Online Library.

[4] A.Langousis,A.Mamalakis,M.Puliga, and R.Deidda (2016)., *Threshold detection for the generalized Pareto distribution : Review of representative methods and application to the NOAA NCDC daily rainfall database.*

[5] A.C Davison, and R.L.Smith,(1990), *Models for Exceedances Over High Thresholds," Journal of the Royal Statistical Society.* Series B (Methodological), 52, 393-442.

[6] Arnold, C.Barry,N.Balakrishnan, et H.N.Nagaraja, (1992),*A rst course in order statistics. Wiley Series in Probability and Mathematical Statistics : Probability and Mathematical Statistics.* A Wiley-Interscience Publication. John Wiley and Sons, Inc., New York, 1992.

[7] A.L.M.Dekkers,J.H.Einmahl, de L.Haan, *A moment estimator for the index of an extrem value distribution*, The Annals of Statistics, 1989.

[8] A.F.Jenkinson, 1955,*The frequency distribution of the maximum or minimum of meteorological elements Q.J.R.* Meteorol. Soc., 81. 158-171.

[9] Bernard Kamsu-Foguem et ,*Contribution de la théorie des valeurs extrêmes à la gestion et a la santé des système* thèse Doctorat de l'université de Toulouse 2018.

[10] B.Gnedenko(1943), *Sur la distribution limite du terme maximum d'une série aléatoire*, The Annals of Mathematics, 44, 423-453.

[11] Bernard Rapacchi, *Une introduction au Bootstrap*, Centre Interuniversitaire de Calcul de Grenoble 15 décembre 1994.

[12] B.M.Hill, 1975, *A simple approach to inference about the tail of a distribution.*Ann. Statist. 3, 1136-1174.

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

[14] D.M.Mason, 1982, *Laws of large numbers for sums of extreme values.*, Ann.Probab. 10, 756-764.

[15] D.Cox and N.Reid,1987,*Parameter orthogonality and approximate conditional inference(with discussion)*,J.R.statist.soc.B,49,1-39

[16] F.Caeiro, and M.I.Gomes, (2016), *Threshold Selection in Extreme Value Analysis," Extreme Value Modeling and Risk Analysis : Methods and Applications.* 69-82.

[17] Frédéric PLANCHET *Utilisation de la théorie des valeurs extrêmes dans le cadre de Solvabilité* modèles financiers en assurance et analyses dynamiques.

[18] H.Drees, De L.Haan, and S.Resnick, (2000), *How to Make a Hill Plot,"The Annals of Statistics*, 28, 254-274.

[19] H.D.David, 1970,*Order statistics.*John Wiley & Sons, Inc., New York-London-Sydney.

[20] J.El Methni, *Contributions à l'estimation de quantiles extrêmes. Applications à des données environnementales. Thèse de doctorat*, Université de Grenoble, 2013.

[21] J.Beirlant, Y.Goegebeur,J.Segers and J.Teugels, (2004), *Statistics of Extremes : Theory and Applications*, Wiley Series in Probability and Statistic.

[22] J.R.M. Hosking, J.R. Wallis, E.F. Wood,*Estimation of the Generalized Extreme Value distribution by the method of Probability-Weighted Moments*, Technometrics, 1985.

[23] J.Brian,D.Reich and Michael. Porter,*discussion of "estimating the historical and future probabilities of large terrorist events" by AARON CLAUSET AND RYAN WOODARD* North Carolina State University and University of Alabama 2014

[18] Max Rydman,*Application of the Peaks-Over-Threshold Method on Insurance Data*2018.

[24] J.Carpenter, et J.Bithel, (2000), *Bootstrap confidence intervals when, which, what ? a pratical guide for medical statisticians* ,Statistics in medicine[1141– 1164].

[25] J.Pikands,1975, *Statistical inference using extreme order statistics* ,Annals of Statistics. 3, (119-131)

[26] J.Caers, J.Beirlant, and M.A.Maes (1999), *Statistics for modeling heavy tailed distributions in geology : Part I.* Methodology," Mathematical geology, 31, 391-410.

[27] J.Danielsson,L.M.Ergun, De L.Haan and de Vries.C. G, 2016, *TailIndex Estimation : Quantile Driven Threshold Selection,Available at SSRN.* https ://ssrn.com/abstract=2717478.

[28] J.Beirlant, P.Vynckier and J.L.Teugels (1996), *Tail Index Estimation, Pareto Quantile Plots, and Regression Diagnostics*,"Journal of the American Statistical Association, 91,1659-1667.

[29] J.Galambos, 1978, *The asymptotic theory of extreme order statistics.Wiley Series in Probability and Mathematical Statistics.* John Wiley & Sons, New York-Chichester-Brisbane.

[47] Jinyuan Chang & Peter Hall , *Double-bootstrap methods that use a sing double-bootstrap simulation* Department of Mathematics and Statistics The University of Melbourne, VIC, 3010, Australia.

[30] J.Pickands,*Statistical inference using extreme order statistics 1975.*

[31] L.Gardes and S.Girard, 2013, *Estimation de quantiles extrêmes pour les lois à queue de type Weibull* , Journal de la Société Française de Statistique.

[32] Laurent Gardes, *Théorie des valeurs extrêmes*, Université de Strasbourg.

[33] M.Garrido,*Modélisation des événements rares et estimation des quantiles extrêmes, Méthodes de sélection de modèles pour les queues de distribution. Modélisation et simulation* Université Joseph-Fourier - Grenoble, 2002.

[34] M.Fréchet, 1927,*Sur la loi de probabilité de l'écart maximum*, Ann. Soc. Polon.Math., vol. 6(3). 92-116.

[35] N.Temame,*Estimation du quantile extreme et de la VAR. Thèse de Magistère.* Université de Mouloud Mammeri,Tizi-Ouzou, 2011.

[36] N.Bingham,C.Goldie, and J.Teugels, 1987. *Regular Variation. Cambridge*,University Press.

[37] N.Balakrishnan,et A.C Cohen, 1991.,*Order statistics and inference. Estima-tion methods.* Statistical Modeling and Decision Science. Academic Press, Inc., Boston, MA.

[38] N.Smirnov,1948, *The Annals of Mathematical Stastistics [19-279].*

[39] N.Balakrishnan, et A.C.Cohen, 1991,*Order statistics and inference. Estimation methods. Statistical Modeling and Decision Science.* Academic Press, Inc.,Boston, MA

[40] P.Embrechts,C.Klüppelberg, and T.Mikosch, 1997, *Modelling extremal events For insurance* , Applications of Mathematics (New York),33. Springer-Verlag, Berlin.

[41] R.Bechir, *Fondements de la théorie des valeurs extrêmes, ses principales applications et son apport à la gestion des risques du marché pétrolier Math*,Mathematics and Social Sciences (2009, p.29-63)

[42] R.Fisher,L.Tippett, 1928., *Limiting forms of the frequency loi of the largest or smallest member of a sample*, Proceedings of the Cambridge Philosophical Society 24, 1928, p. 180-190.

[43] R.L.Smith, 1987. Estimating tails of probability lois. The Annals of Statistics.3, 1174-1207.

[44] R.Palm, *Utilisation du bootstrap pour les problèmes statistiques liés à l'estimation des paramètres.*Biotechnol. Agron. Soc. Environ. 2002, 6(3), [143–153]

[45] S.Coles, 2001, *An introduction to statistical modeling of extreme values. Springer Series in Statistics*, Springer-Verlag London, Ltd., London.

[46] SG.Coles and MJ.Dixon, *Likelihood-Based Inference for Extreme Value Models Extremes 5-23, 1999.*

[47] S.L Resnick, *extrême values,regular variation and point processes* SpringerVerlag,1987.[cité en page 11].

Annexe

Méthodes d'estimation du seuil :

— **Mean Excess Plot (MRLplot)** :

library{evir+ismev}
x<-rgpd(n,shape=xi,location=mu,scale=sigma)
mrl.plot(x,umin=.., umax=..)

x<-rgev(100,0.5,10,1)
x<-rgpd(3000,0.3,30,1)
mrl.plot(x,umin=10, umax=60)

library{pot}
rgpd(5, loc = 1, scale = 2, shape = -0.2)
mrlplot(x, u.range = c(1, quantile(x, probs = 0.995)), col = c("green", "black", "green"), nt = 2000)

— **tcplot** :

library(POT)
x<-rgpd(3000, 1, 30, -0.2)
tcplot(x, c(2, 60))

rgpd(5, loc = 1, scale = 2, shape = -0.2)
tcplot(x, u.range = c(0.9, 0.995))
shape :paramétré de forme
scale :paramètre d'échelle
location :le seuil

— **Estimateur de Hill**

library(evir+evmix)
x<-rgev(1000,0.5,10,1)

x<-rgpd(1000,0.7,10,1)
hill(x, option = c("alpha","xi","quantile"), start = 100, end = 700, reverse = FALSE)
, p = 0.001, ci = 0.8)
auto.scale = TRUE, labels = TRUE, ...)

— **estimation d'une paramètre par bootstrap**
library(boot)

x<-c(12,15,42,20,32,25,36,13,40)
m<-mean(x) (la moyenne) m sd(x) sd(x)/sqrt(length(x)) moyenne <- function(d,w) n <- length(d) return(sum(d[w[1 :n]])/n) x.boot <- boot(x,moyenne, R=500) x.boot$t0 x.boot$t (vecteur des R répliqats du paramètre obtenus par bootstrap)
(L'estimation bootstrap de la moyenne et l'erreur type correspondante peuvent être obtenues par :) print(mean(x.boot$t)) print(sd(x.boot$t))
print(x.boot) print(boot.array(rem.boot)) plot(x.boot)
var(x)
print(varp(x)) (variance)
varp.boot <- function(d,w) _ n <- length(d)_ m <- sum(d[w[1 :n]])/n_
$return(sum((d[w[1 : n]] - m)^2)/n)$
x.boot <- boot(x,varp.boot, R=500)
x.boot$t0
x.boot$t
print(mean(x.boot$t)) (Moyenne des variances obtenues par bootstrap)
x.boot <- boot(x, statistic=varp.boot, R=500)
plot(x.boot)

— **application aux données météorologiques**
library(e1071)
skewness(data)
kurtosis(data)
length(data)
summary(data)
sd(data)
var(data)
library(evir+evmix)
plot.new()

```
par(mar=c(4,4,3,4))

barplot(data, col="blue", names.arg=Mois,ylab="Hum(%)",
xlab="temps(mois)",
main="Humidité à Dellys",ylim=c(0,100),
las=2,space=0,cex.main=0.8)

plot(data,col="blue",type="l",ylab="Hum(%)"
,xlab="temps(mois)",
main="Humidité à Dellys",ylim=c(0,90),
las=2,cex.main=0.8)

mrl.plot(data,umin=50, umax=90)
library(POT)
tcplot(data,c(0,85))
```