CARACTÉRISATION DES DISTRIBUTIONS À QUEUE LOURDE
POUR L’ANALYSE DES CRUES

Rapport de recherche No R-929
Mars 2007
CARACTÉRISATION DES DISTRIBUTIONS À QUEUE LOURDE
POUR L’ANALYSE DES CRUES

Par

Salaheddine El-Adlouni
Bernard Bobée
Taha B.M.J. Ouarda

Chaire en hydrologie statistique (Hydro-Québec / CRSNG)
Chaire du Canada en estimation des variables hydrologiques
INRS-ETE, Université du Québec
490, rue de la Couronne (Québec), Canada G1K 9A9

Rapport de Recherche N° R-929

Mars 2007

© Tous droits réservés, S. El Adlouni et al. 2007
TABLE DES MATIÈRES

TABLE DES MATIÈRES ........................................................................................................ V

LISTE DES FIGURES ........................................................................................................ VII

LISTE DES TABLEAUX ...................................................................................................... IX

1. INTRODUCTION ........................................................................................................ 1

2. PROCESSUS GÉNÉRATEURS DE DISTRIBUTIONS À QUEUE LOURDE .................................................................................................................. 5

3. CLASSES DE DISTRIBUTIONS À QUEUE PLUS LOURDE QUE LA LOI NORMALE .......................................................................................................... 9
   3.1 Distributions avec des moments exponentiels inexistants ........................................ 11
   3.2 Distributions Subexponentielles .............................................................................. 11
   3.3 Distributions à variations régulières d’indice $\alpha > 0$ ........................................... 12
   3.4 Comportement de Pareto avec $\alpha > 0$ .................................................................. 14
   3.5 Distributions $\alpha$-Stables .................................................................................. 15
   3.6 Classification basée sur le comportement asymptotique ...................................... 16

4. ESTIMATION DE L’INDICE DES VALEURS EXTRÊMES ............................................. 21

5. TECHNIQUES DE CHOIX DE LA CLASSE DES EXTRÊMES ....................................... 25
   Méthode 1 : Graphique Log-log .................................................................................. 27
   Méthode 2 : La moyenne empirique des excès ...................................................... 30
   Méthode 3 : Graphique des rapports du maximum et de la somme .................... 31
   Méthode 4 : Graphique du rapport de Hill ............................................................ 33
   Méthode 5 : Statistique de Jackson ........................................................................ 35

6. APPLICATIONS À DES CAS RÉELS ......................................................................... 39
   Comparaison des ajustements pour la station CSS ................................................. 45
   Comparaison des ajustements pour la station CSG ................................................. 46
   Comparaison des ajustements pour la station RDO ................................................. 48
   Comparaison des ajustements pour la station USH ................................................. 49
   Discussion des résultats ............................................................................................ 51

7. CONCLUSION ............................................................................................................ 53

RÉFÉRENCES .................................................................................................................... 55
Annexe A Quelques Distributions Statistiques........................................................ 59
  A.1 Loi Lognormale..................................................................................................... 59
    A.1.1 Loi Lognormale à deux paramètres.......................................................... 59
    A.1.2 La loi Lognormale à trois paramètres (LN3) ........................................... 61
  A.2 Distribution Généralisée des Valeurs Extrêmes................................................... 62
    A.2.1 Loi des Valeurs Extrêmes type I (EV1 ou Gumbel) ........................................... 63
    A.2.2 Loi des Valeurs Extrêmes type II (EV2 ou Fréchet).......................................... 63
    A.2.3 Loi des Valeurs Extrêmes type III (EV3 ou Weibull) ........................................ 64
  A.3 Famille des lois de Halphen ................................................................................... 65

Annexe B Exemples de distributions appartenant à différents Domaines d’Attraction maximum .......................................................... 67

Annexe C Stations obtenues du site de l’UNESCO ...................................................... 71
LISTE DES FIGURES

Figure 1 : Illustration de la différence entre la loi normale et une loi à queue lourde (HIB). ................................................................. 10
Figure 2 : Différentes classes de distributions de queue très légère (E) à très lourde (A). ............................................................... 11
Figure 3 : Distributions ordonnées par rapport à leurs queues droites. .......... 19
Figure 4 : Illustration de la différence entre les différentes classes de distributions. .............................................................. 20
Figure 5 : Illustration du graphique Log-Log pour la discrimination entre la classe C et D. .............................................................. 28
Figure 6 : Graphique Log-Log pour les distributions du tableau 2. ................. 29
Figure 7 : Graphique de la moyenne empirique des excès pour quelques distributions du tableau 1. .............................................. 30
Figure 8 : Graphique du rapport de la somme et du maximum pour les lois Lognormale, GEV et HIB (Cas 2). ................................. 32
Figure 9 : Graphique du rapport de Hill pour les lois Lognormale, GEV, HA et HIB. ................................................................. 34
Figure 10 : Test basé sur la statistique de Jackson pour les lois Lognormale, GEV, HA et HIB (Cas 2 du Tableau 2). ................................. 38
Figure 11 : Coefficients de variation et d’asymétrie de 32 séries tirées du site de UNESCO. ................................................................. 39
Figure 12 : Graphique Log-Log pour les quatre séries étudiées .................... 40
Figure 13 : Test d’indépendance des observations des quatre séries étudiées (t-test). ................................................................. 41
Figure 14 : Fonction de la moyenne des excès pour les quatre séries ............. 42
Figure 15 : Graphique basé sur le rapport du maximum et de la somme ......... 43
Figure 16 : Graphique basé sur le rapport de Hill ......................................................... 44
Figure 17 : Graphique basé sur la statistique de Jackson .............................. 44
LISTE DES TABLEAUX

Tableau 1 : Classement des lois selon les caractéristiques de l’extrémité droite
(OUARDA et al. 1994) ........................................................................................................17

Tableau 2 : Cas considérés pour illustrer les différentes méthodes graphiques. ........26
L’analyse fréquentielle des événements rares représente un intérêt particulier pour la gestion et la prévention des événements extrêmes en hydrologie. La majorité de ces événements extrêmes causent des dégâts humains et matériels importants. L’objectif principal de l’analyse fréquentielle est d’étudier la probabilité qu’un certain événement soit dépassé. Cette procédure est liée à la théorie des valeurs extrêmes, qui est souvent introduite à partir de propriétés asymptotiques. Dans la majorité des cas on se base sur l’ajustement d’une certaine distribution qui a été considérée comme adéquate pour ajuster les crues dans une région donnée, par exemple la loi des valeurs extrêmes généralisée (GEV) en Grande Bretagne, Lognormale (LN) en Chine, Log-Pearson type III (LPIII) aux États Unis… (Bobée 1999). Cependant, deux problèmes se posent en pratique :

- Le premier est lié à la taille de l’échantillon qui est souvent faible, ce qui met en question l’application des résultats asymptotiques (La taille minimale $n = 50$ a été recommandée pour avoir des estimations robustes (Stedinger 2000), mais souvent une telle taille n’est pas suffisante pour faire le choix de l’ajustement le plus adéquat, surtout quand-t-on s’intéresse aux périodes de retour supérieures ou égales à 100 ans).

- Le second est dû au fait qu’une loi de probabilité ne donne pas toujours un bon ajustement dans toutes les applications (Bobée et Rassmussen 1995). D’où l’intérêt : (1) d’effectuer un classement des distributions en fonction du comportement de leurs queues droite (cas des crues) et (2) d’établir des critères de discrimination entre les différentes classes dans le cas d’un échantillon de faible taille, à partir de considérations physiques ou statistiques.
Dans ce travail, on distingue deux principales parties. On présente d’abord deux approches pour classer les différentes lois utilisées en hydrologie, en fonction de leurs queues droites. La première approche, tirée de la théorie des valeurs extrêmes, présente six classes de lois en fonction de différentes expressions qui caractérisent chacune d’elles (Werner and Upper (2002)). La deuxième approche, donnée par Ouarda et al. (1994), est basée sur le comportement asymptotique de la fonction de densité de probabilité. Une des classes qu’on peut caractériser par les deux approches est la classe des distributions à variations régulières (de type puissance) qui représente un intérêt particulier dans la théorie des valeurs extrêmes pour deux raisons : (1) cette classe peut être définie à partir de la théorie classique des extrêmes, (2) elle peut être caractérisée par un seul paramètre appelé « indice des valeurs extrêmes ». Une estimation efficace de cet indice permet de classer la distribution dans la classe à queue lourde ou dans la classe à queue très lourde. Plusieurs méthodes, paramétriques et non-paramétriques, permettent l’estimation de cet indice des valeurs extrêmes. Ces méthodes et leur comparaison seront discutées dans ce document en raison de l’importance de cette étape dans le processus d’ajustement. Dans la deuxième partie de ce travail on présente certaines techniques, les plus connues dans la littérature, permettant une discrimination entre les distributions ayant une queue légère ou lourde. On s’intéressera particulièrement aux techniques graphiques, développées à partir des propriétés des classes de distributions dites subexponentielles (section 3).

Dans toute la suite de ce document on adoptera les notations suivantes. Soit $X_1, \ldots, X_n$ un $n$-échantillon d’une variable aléatoire $X$ non-dégénérée. Il s’agit donc de $n$ variables aléatoires indépendantes et identiquement distribuées (i.i.d.) suivant une distribution continue $F$ et de

1 Une variable aléatoire est dégénérée si son support est réduit à un seul point.
fonction de densité de probabilité (fdp) \( f \). On notera par \( \overline{F} \) la fonction de probabilité au dépassement (connue aussi sous le nom de la fonction de survie en fiabilité) :

\[
\overline{F}(x) = \Pr(X \geq x) = 1 - F(x).
\]

Comme nous l’avons déjà mentionné, l’objectif principal de la théorie des valeurs extrêmes est le développement de techniques qui permettent d’avoir une estimation adéquate des quantiles. En hydrologie on s’intéresse souvent à la notion de période de retour. On notera \( x_r \) l’événement de période de retour \( T \) et qui correspond au quantile de probabilité au dépassement égale à \( \frac{1}{T} \) :

\[
\Pr(X \geq x_r) = \overline{F}(x_r) = \frac{1}{T}.
\]
2. PROCESSUS GÉNÉRATEURS DE DISTRIBUTIONS À QUEUE LOURDE

Les fondements théoriques pour la plupart des distributions, en termes de processus hydrologique qui génère de tels extrêmes, sont rarement considérés pour le choix du modèle (Kidson et al. 2005 et Singh et Strupczewski 2002). La distribution Lognormale est générée par des phénomènes à effet proportionnel (connus sous le nom de processus multiplicatifs). Cependant un petit changement dans ces processus peut les transformer en processus générateurs des lois de type puissance ou de comportement fractal (Champernowne 1953, Mandelbrot 1997, 2003 et Turcotte 1997). Benoît Mandelbrot est un des premiers scientifiques à insister sur l’emploi des fractales basés sur l’invariance d’échelle : Les propriétés sont les mêmes, quelle que soit l’échelle à laquelle on les regarde. Ses travaux depuis les années 50 portaient sur la répartition des revenus, l’évolution des cours boursiers, les crues du Nil et la fréquence des mots. Il distingue trois types de comportement : le premier lié au théorème de la limite centrale et la loi des grands nombres, où la distribution limite est une loi de Gauss (Normale). Le deuxième est le cas lent, représenté par la loi Lognormale. Le troisième est le cas dit « Sauvage », représenté par la loi de Cauchy, dont la moyenne et la variance, sont infinies. Mandelbrot indique que pour ce dernier cas les moyennes ont exactement la même distribution que chacun des tirages individuels : « Aussi loin qu’on aille, on ne verra jamais s’instaurer de compensation entre les différents tirages : La moyenne est aussi incertaine que chacun des épreuves ». Ces trois cas sont des cas extrêmes, et entre eux il y a toute une gamme de cas intermédiaires qu’il résume en « effet Noé et effet Joseph » (Mandelbrot et Wallis 1968). Une comparaison des processus générateurs de lois de type
puissance et la loi Lognormale est donnée par Mitzenmacher (2004), dans le cas du trafic internet. Il conclu que les processus générateurs de ces distributions sont très semblables.

L’utilisation des distributions de type puissance en hydrologie est restreinte à l’emploi de certaines distributions qui ont ce type de comportement et se justifie souvent par des études empiriques. Malamud et Turcotte (2005) ont montré, sur la base de plusieurs séries hydrologiques, que les distributions de type puissance sont les plus adaptées pour représenter les crues. Ces auteurs constatent que les distributions de type puissance, en particulier la loi Log-Gumbel (connue mieux sous le nom de loi de Fréchet ou de loi des valeurs extrêmes de type 2 EV2), sont en agrément avec les séries de crue et donnent des prédictions prudentes au niveau des crues extrêmes futures. Koutsoyiannis (2004) montre par une étude empirique que la loi de Fréchet, qui est de type puissance (Section 3), est plus adéquate pour représenter les précipitations extrêmes que la distribution Gumbel. Bernier (1959) est un des premiers à suggérer l’emploi de la loi Fréchet en hydrologie. Il montre sur la base de quelques séries de débit, que la loi Fréchet et meilleure que la loi Gumbel pour l’ajustement des extrêmes. Strupzewski et al. (2005) et Kochanek et al. (2005), ont comparé plusieurs combinaisons (Distribution/Méthode d’Estimation) pour l’ajustement de séries générées à partir d’une distribution Wakeby à 5 paramètres. Ils constatent que les distributions à deux paramètres donnent de meilleurs résultats en ce qui concerne le biais des quantiles, que celles à trois paramètres, et rejoignent les conclusions déjà connues dans la littérature concernant les méthodes d’estimation : (1) les estimateurs obtenus par la méthode des moments sont moins biaisés que les autres et (2) les estimateurs du maximum de vraisemblance ont souvent une erreur quadratique moyenne relative faible. Mais comme conclusion générale, les auteurs concluent qu’aucune distribution n’est
valable dans tout les cas et soulignent l’importance de développer des techniques pratiques pour le choix de l’ajustement le plus adéquat à une série de données.

D’un point de vue statistique, l’emploi de la loi Lognormale est justifié par le théorème de la limite centrale. En effet pour les processus multiplicatifs, qui représente la succession des effets de plusieurs phénomènes physiques, se transforme en processus additifs dans une échelle logarithmique. L’application du théorème de la limite centrale montre que pour le modèle additif la distribution asymptotique est normale, ce qui est équivalent à une loi Lognormale dans le cas multiplicatif. La justification statistique des distributions de type puissance est basée sur le théorème de la limite centrale généralisé (Section 3.5) qui s’applique dans le cas de variables ayant des moments infinis (Mandelbrot 2003).

En pratique, et comme la taille des séries est souvent très faible, l’application de certains résultats asymptotiques, n’est pas très pertinente. Une alternative est de considérer plusieurs classes de distributions capables de représenter les différentes formes des séries observées, surtout au niveau des queues des distributions. En effet, en pratique la plupart des distributions donnent le même ajustement pour la partie centrale (pour les périodes de retour qui ne dépassent pas la taille de l’échantillon) de la série des observations. Cependant, le problème d’extrapolation concerne surtout la partie des extrêmes. On abordera, dans ce travail, le lien entre les différents types de queues de distributions sous forme de classes. Ceci sera présenté d’une manière assez générale pour montrer la relation qui existe entre les différentes classes et ce à partir de critères qui permettent de choisir le type de distributions le plus adéquat pour représenter une série de données.
3. CLASSES DE DISTRIBUTIONS À QUEUE PLUS LOURDE QUE LA LOI NORMALE

Dans cette partie on présente la notion de distribution à queue lourde et les différentes classes de ce type de distributions. Les distributions à queues lourdes sont liées à la théorie des valeurs extrêmes et permettent de modéliser plusieurs phénomènes rencontrés dans différentes disciplines : finances, hydrologie, télécommunication, géologie… et plus récemment en climatologie. Plusieurs définitions ont été associées à ces distributions en fonction de critère de classification. La caractérisation la plus simple et celle basée sur la comparaison avec la loi normale : On dit, alors, qu’une distribution a la queue lourde si :

\[ C_k = E \left[ \left( \frac{X - \mu_X}{\sigma_X} \right)^4 \right] > 3 \]  

(1)

Ce qui est équivalent à dire qu’une distribution a une queue lourde si et seulement si son coefficient d’aplatissement, \( C_k \), est supérieur à celui de la loi Normale (pour laquelle \( C_k = 3 \)). La différence entre la loi Normale et une loi avec une queue plus lourde a été illustrée par Hubert et Bendjoudi (1996) comme dans la Figure 1. Dans cette figure on présente les fonctions de densité de probabilité de la loi normale et d’une distribution à queue plus lourde (la loi Halphen type B⁻¹ (HIB), voir pour plus de détails l’annexe A). On remarque que (voir agrandissement 1-b) la fdp de la loi normale est presque nulle au niveau des extrêmes (queue droite), alors qu’elle ne l’est pas pour la loi HIB.
Figure 1 : Illustration de la différence entre la loi normale et une loi à queue lourde (HIB)

La caractérisation, donnée par l’équation (1), est très générale et ne peut être appliquée que si le moment d’ordre 4 existe. Par conséquent aucune discrimination, pour les distributions ayant un moment d’ordre 4 infini, ne peut être faite si on ne considère que ce critère. Malheureusement, il n’y a pas de critère pour classer toutes les distributions selon la queue droite. Cependant, on obtient un tel classement pour certaines classes de distributions :

- les distributions avec des moments exponentielles inexistants (E),
- les distributions Subexponentielles (D),
- les distributions à variations régulières (C),
- les distributions avec un comportement de Pareto (B) et
- les distributions α-Stables avec α<2 (A).
Chapitre 3, Classes de distributions à queue plus lourde que la loi Normale

Toutes ces classes sont emboîtées : $A \subset B \subset C \subset D \subset E$ et peuvent être représentées selon la figure 2 (Werner et Upper, 2002).

![Diagramme des classes de distributions]

**Figure 2 : Différentes classes de distributions de queue très légère (E) à très lourde (A).**

### 3.1 Distributions avec des moments exponentiels inexistants

La classe E contient toutes les distributions telles que $E[e^X] = \infty$. On note que la loi Normale n’appartient pas à cette classe parce que la probabilité au dépassement, $F$, pour les extrèmes de cette classe, décroît moins rapidement que celle de la loi normale. Dans ce sens, la classe E est celle des distributions ayant une queue plus lourde que celle de la loi normale.

### 3.2 Distributions Subexponentielles

La classe D des distributions subexponentielles est caractérisée par la définition suivante (Beirlant et al. 2004): On dit qu’une distribution est subexponentielle si :

$$\lim_{x \to \infty} \frac{P(X_1 + \cdots + X_n > x)}{P(\max(X_1, \ldots, X_n) > x)} = 1$$

(2)
C’est équivalent à dire que la somme de n distributions subexponentielles (indépendantes et identiquement distribuées (iid)) est extrême si et seulement si leur maximum est extrême. On peut démontrer que l’équation (2) implique que:

\[
\lim_{x \to \infty} \frac{\bar{F}(x)}{e^{-\varepsilon x}} = \infty \quad \forall \varepsilon > 0
\]  

(3)

On rappelle que \(e^{-tx}\) est la forme de la queue de la loi exponentielle. Comme son nom l’indique, la classe D contient les distributions telles que \(\bar{F}\) décroît plus lentement que n’importe quelle loi exponentielle.

### 3.3 Distributions à variations régulières d’indice \(\alpha > 0\)


\[
\lim_{t \to \infty} \frac{\bar{F}(tx)}{F(t)} = x^{-\alpha}
\]  

(4)

Ce qui est équivalent à dire que, en ce qui concerne les valeurs extrêmes (t tend vers l’infini) la distribution a le même comportement que celui de la loi de Pareto. Par conséquent, la probabilité au dépassement des extrêmes décroît suivant une fonction puissance (appelée aussi décroissance géométrique). Le paramètre \(\alpha\) est appelé « l’indice des valeurs extrêmes » (Tail index) et peut être utilisé comme critère pour classer les distributions par rapport au comportement de la fonction de probabilité au dépassement \(\bar{F}\) au niveau des valeurs extrêmes.
Lien avec la théorie des valeurs extrêmes

La classe des distributions à variations régulières est liée à la théorie des valeurs extrêmes. En effet, la théorie des valeurs extrêmes est basée sur le théorème de Fisher-Tippett (Fisher et Tippet 1928) qui permet de déterminer la distribution du maximum d’un échantillon de $n$ variables indépendantes et identiquement distribuées (iid) :

Si $X_1,\ldots,X_n$ une séquence de variables aléatoires iid, et $M_n$ le maximum défini par :

$$M_n = \max\{X_1,X_2,\ldots,X_n\}$$

alors, le théorème de Fisher-Tippett montre que si la distribution du maximum de l’échantillon converge vers une distribution non dégénérée alors elle suit une des trois fdp suivantes (de fonctions de densité) :

- Gumbel (EV1): $\Lambda(x) = \exp\left(-e^{-x}\right)$, $x \in \mathbb{R}$.
- Fréchet (EV2): $\Phi_\alpha(x) = \begin{cases} 0 & x \leq 0 \\ \exp\left(-x^{-\alpha}\right) & x > 0 \end{cases}$, $\alpha > 0$.
- Weibull (EV3): $\Psi_\alpha(x) = \begin{cases} \exp\left(-(-x^{-\alpha})\right) & x \leq 0 \\ 0 & x > 0 \end{cases}$, $\alpha > 0$.

Les trois distributions peuvent être représentées sous une seule forme, appelée distribution des valeurs extrêmes généralisée (GEV, Annexe A). Lorsque la distribution du maximum converge vers l’une de ces trois distributions, on dit qu’elle appartient au domaine d’attraction maximum de Fréchet, Weibull ou Gumbel. Les distributions appartenant au domaine d’attraction maximum
Caractérisation des Distributions à Queue Lourde pour l’analyse des crues

de Gumbel ont une queue légère (Moderately heavy tails). Alors que, les distributions appartenant
au domaine d’attraction maximum de Weibull ont un support borné supérieurement. D’autres
caractérisations des distributions appartenant à chacun de ces domaines d’attraction maximum
sont données dans la littérature (Resnick 1987 et Embrechts et al. 2003). On peut déterminer sur
la base de ces caractérisations le domaine d’attraction maximum de la majorité des distributions
usuelles. Comme par exemple pour le domaine d’attraction maximum de Gumbel on trouve les
lois : Normale, Exponentielle, Log-normale, Gamma. Pour le domaine d’attraction maximum de
Fréchet : les lois Cauchy, Pareto, Chi-Deux et Student et pour le domaine d’attraction maximum
de Weibull : les lois Uniforme et Beta. Certains exemples sont donnés en Annexe B.

Les distributions du domaine d’attraction maximum de Fréchet présentent un intérêt particulier
pour la modélisation des extrêmes. En effet, une distribution appartient au domaine d’attraction
maximum de Fréchet si et seulement si elle est à variations régulières, c.à.d si elle appartient à la
classe C (Embrechts et al. 2003).

3.4 Comportement de Pareto avec $\alpha > 0$

La classe B est celle des distributions ayant le comportement de Pareto. La distribution de Pareto
est définie à partir de sa fonction de distribution :

$$F(x) = 1 - u^\alpha x^{-\alpha} = 1 - \left(\frac{u}{x}\right)^\alpha \quad x \geq u \text{ et } u > 0$$

Les fonctions de probabilité au dépassement $\overline{F}$, des distributions de la classe B, sont de la forme
$u^\alpha x^{-\alpha}$. L’indice des valeurs extrêmes peut être caractérisé à partir des moments des distributions
de type Pareto. En effet, la fonction de densité de probabilité d’une loi de Pareto est

\[ f_{\text{Pareto}}(x) = \alpha u^\alpha x^{-\alpha - 1} \]

et les moments d’ordre \( k \) sont donnés par :

\[
E[X^k] = \alpha u^\alpha \int_u^\infty x^{k-\alpha - 1} \, dx
\]  

(9)

Par conséquent seuls les moments d’ordre \( k \), tels que \( k < \alpha \), sont finis pour ce type de distributions.

### 3.5 Distributions \( \alpha \)-Stables

La propriété précédente est importante pour définir la classe A, la classe des distributions \( \alpha \)-Stables (appelées aussi distribution stables). Les distributions stables constituent une classe très riche de lois de probabilité capables de représenter différentes asymétries et des queues très lourdes. Cette classe a été caractérisée par Lévy (1925) dans ces travaux sur la somme de variables indépendantes et identiquement distribuées (iid). L’absence de formules explicites des densités de ces distributions a limité leur utilisation. Les distributions de cette classe ont un comportement asymptotique de Pareto avec \( 0 < \alpha \leq 2 \). Lorsque \( \alpha = 2 \) on retrouve la loi Normale, cependant pour \( \alpha < 2 \) le moment d’ordre \( r \geq \alpha \) n’est pas fini, ces distributions ont donc une variance infinie, et par conséquent une queue très lourde. Pour \( \alpha = 1 \) on retrouve la loi de Cauchy (avec une moyenne et une variance infinies). Cette classe a une grande importance dans la théorie des valeurs extrêmes, puisque les distributions stables peuvent être caractérisées à partir du théorème de la Limite Centrale Généralisé. En effet, le théorème de la limite centrale indique que la somme de \( n \) variables aléatoires (iid) de variance finie tend vers une distribution normale lorsque \( n \) tend vers l’infini. Une généralisation de ce théorème par Gnedenko et Kolmogorov
Caractérisation des Distributions à Queue Lourde pour l’analyse des crues

(1954) indique que si la condition de variance finie n’est pas respectée, la seule loi limite possible de la somme de \( n \) variables aléatoire (iid), est une loi stable. L’utilisation de cette famille de distributions était très limitée à cause des problèmes de calcul : leurs fonctions de densité de probabilité et de distributions n’étant pas explicites. Récemment plusieurs logiciels ont été proposés pour permettre la résolution de ces problèmes (Nolan 2001, 2006), et on trouve des applications des distributions Stables dans plusieurs domaines tels que finances, physique et le trafic Internet.

3.6 Classification basée sur le comportement asymptotique

Ouarda et al. (1994) ont donné une classification basée sur le comportement asymptotique de la fonction densité de probabilité, \( f \) : Pour une période de retour, \( T \), assez grande on a l’équivalence suivante (Gumbel 1957, Ouarda et al. 1994):

\[
T \xrightarrow{x_T \to \infty} \left( \frac{1}{f(x_T)} \right)' = -\frac{f'(x_T)}{f^2(x_T)}
\]

(10)

En effet, pour une période de retour \( T \) assez grande \( \bar{F}(x_T) = 1 - F(x_T) = \frac{1}{T} \) et \( f(x_T) \) tendent vers 0. Au voisinage de l’infini le rapport de ces deux fonctions est équivalent au rapport de leurs dérivées (Règle de l’Hopital) :

\[
\lim_{x_T \to \infty} \frac{f(x_T)}{1 - F(x_T)} \quad \lim_{x_T \to \infty} \frac{f'(x_T)}{-f(x_T)} \quad \Rightarrow \quad T = \frac{1}{1 - F(x_T)} \quad \frac{f'(x_T)}{f^2(x_T)} = \left( \frac{1}{f} \right)'(x_T)
\]
On en déduit le comportement asymptotique, des quantiles en fonction de la période de retour $T$, pour plusieurs distributions utilisées en hydrologie (Tableau ci-dessous).

### Tableau 1 : Classement des lois selon les caractéristiques de l’extrémité droite (Ouarda et al. 1994)

<table>
<thead>
<tr>
<th>Classe</th>
<th>Caractéristiques</th>
<th>Loi</th>
<th>Valeurs des paramètres</th>
</tr>
</thead>
<tbody>
<tr>
<td>I : $x \approx T^p$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classe C</td>
<td>$P = 1/\alpha$</td>
<td>Log-Pearson 3 ($\alpha, \lambda, m$)</td>
<td>$\alpha&gt;0, \lambda&gt;0, m \in \mathbb{R}$</td>
</tr>
<tr>
<td>Classe C</td>
<td>$P = 1/\alpha$</td>
<td>Log-Logistic ($\alpha, \lambda$)</td>
<td>$\alpha&gt;0, \lambda&gt;0$</td>
</tr>
<tr>
<td>Classe C</td>
<td>$P = -1/s\lambda$</td>
<td>Gamma Généralisée ($s, \alpha, \lambda$)</td>
<td>$\lambda&gt;0, \alpha&gt;0, \lambda&gt;0$</td>
</tr>
<tr>
<td>Classe C</td>
<td>$P = 1/\lambda$</td>
<td>Gamma Inverse ($\alpha, \lambda$)</td>
<td>$\alpha&gt;0, \lambda&gt;0$</td>
</tr>
<tr>
<td>Classe C</td>
<td>$P = -1/k$</td>
<td>Fréchet ($\alpha, k, u$)</td>
<td>$\alpha&gt;0, k&lt;0$</td>
</tr>
<tr>
<td>Classe C</td>
<td>$P = 1/c$</td>
<td>Weibull ($\alpha, c$)</td>
<td>$\alpha&gt;0, c&lt;0$</td>
</tr>
<tr>
<td>Classe C</td>
<td>$P = 1/2\nu$</td>
<td>Halphen type B ($\alpha, \nu, m$)</td>
<td>$\alpha \in \mathbb{R}, \nu&gt;0, m&gt;0$</td>
</tr>
</tbody>
</table>

| II : $x \approx \exp(\ln(T)^{1/2})$ | | | |
| Classe D | Lognormale 2 ($\mu, \sigma$) | $\mu \in \mathbb{R}, \sigma>0$ |
| Classe D | Lognormale 3 ($\mu, \sigma, m$) | $\mu \in \mathbb{R}, \sigma>0, m \in \mathbb{R}$ |

| III : $x \approx \ln(T)^p$ | | | |
| Classe D | $P = 1$ | Pearson type 3 ($\alpha, \lambda, m$) | $\alpha>0, \lambda>0, m \in \mathbb{R}$ |
| Classe D | | Gamma ($\alpha, \lambda$) | $\alpha>0, \lambda>0$ |
| Classe E | | Exponentielle ($\alpha, m$) | $\alpha>0, m \in \mathbb{R}$ |
| Classe D | | Halphen type A ($\alpha, \nu, m$) | $\alpha>0, \nu \in \mathbb{R}, m>0$ |
| Classe D | | LF ($\lambda, \beta$) | $\lambda>0, \beta>0$ |
| Classe D | | Gumbel ($\alpha, u$) | $\alpha>0, u \in \mathbb{R}$ |
| Classe D | $P = 1/2$ | Halphen type B ($\alpha, \nu, m$) | $\alpha \in \mathbb{R}, \nu>0, m>0$ |
| --- | | N ($\mu, \sigma$) | $\mu \in \mathbb{R}, \sigma>0$ |
| Classe D | $P = 1/s$ | Generalized Gamma ($s, \alpha, \lambda$) | $s>0, \alpha>0, \lambda>0$ |
| Classe D | $P = 1/c$ | Weibull ($\alpha, c$) | $\alpha>0, c<0$ |

| IV : $x \leq \bar{P}$ | | | |
| (borne à droite) | $P = m$ | Pearson 3 ($\alpha, \lambda, m$) | $\alpha<0, \lambda>0, m \in \mathbb{R}$ |
| | $P = 0$ | Gamma ($\alpha, \lambda$) | $\alpha<0, \lambda>0$ |
| | $P = \exp(m/\ln_a(e))$ | Log-Pearson 3 ($\alpha, \lambda, m$) | $\alpha<0, \lambda>0, m \in \mathbb{R}$ |
| | $P = 1$ | Log-Logistic ($\alpha, \lambda$) | $\alpha<0, \lambda>0$ |
| | $P = u+\alpha/k$ | Weibull ($\alpha, k, u$) | $\alpha>0, k>0, u \in \mathbb{R}$ |
| | $P = \alpha/k$ | Pareto Généralisée ($\alpha, k$) | $\alpha>0, k>0$ |
La classification donnée par Ouarda et al. (1994) permet de déterminer les distributions appartenant à chacun des domaines d’attraction maximum des extrêmes. On peut établir des liens entre cette classification et celle présentée précédemment (sections 3.1 à 3.5) qui ont trait aux distributions couramment utilisées en hydrologie pour l’étude des crues. A partir du Tableau 1 on remarque que les distributions appartenant à la première classe (classe I) sont les distributions de type puissance, présentée précédemment sous le nom de la classe des distributions à variations régulières (classe C). Ces distributions appartiennent aussi à la classe des distributions subexponentielles (classe D), c’est pour cette raison qu’on les retrouve dans d’autres classes du même tableau. Dans la même classe, des distributions subexponentielles, qui contient les distributions du domaine d’attraction maximum de Gumbel, on retrouve des distributions à queue légère. On retrouve aussi, à partir de cette classification, le problème de discrimination entre la classe des distributions à variations régulière (Classe I) et la loi Lognormale (Classe II). En effet, comme il a été mentionné à la section 2, un petit changement dans les processus à effet proportionnel (qui génère la loi Lognormale) peut les transformer en processus générateurs des lois de type puissance (Champernowne 1953, Mandelbrot 1997, 2003 et Turcotte 1997).

A partir des deux classifications présentées ci-dessus, les différentes distributions (les plus utilisées en hydrologie) peuvent être ordonnées par rapport à leurs queues droites (Figure 3).
Figure 3 : Distributions ordonnées par rapport à leurs queues droites.

Le fait que ces classes soient emboîtées rend le choix, de la distribution la plus adéquate, difficile. Car les distributions avec des queues lourdes, distributions à variations régulières (C) ou du type Pareto (B) appartiennent aussi aux autres classes (D par exemple). Dans ce cas on voit clairement l’importance d’utiliser la bonne méthode d’estimation des paramètres. Une mauvaise estimation des paramètres, de la distribution choisie pour représenter les données, peut entraîner le passage d’une loi à queue lourde vers une loi à queue légère et vice versa. Il faut donc disposer (1) de techniques de choix de la classe de distributions la plus adaptée pour représenter la série observée et (2) de méthodes efficaces pour l’estimation des paramètres.
Figure 4 : Illustration de la différence entre les différentes classes de distributions.

La figure 4 illustre la différence entre des distributions tirées de différentes classes. On présente la loi de Halphen type B⁻¹, la loi de Fréchet (GEV), la loi Lognormale et la loi Gamma. Toutes ces lois ont les mêmes caractéristiques statistiques (moyenne, variance, coefficient d’asymétrie). Le graphique à droite, illustre bien la notion de queue lourde. En effet, on voit bien que les lois GEV et HIB ont une queue plus lourde que celle de la loi Lognormale, qui est plus lourde que la loi Gamma. Par conséquent, les quantiles de grande période de retour $T$ (correspondant à la probabilité au dépassement $\bar{F} = \frac{1}{T}$), estimés à partir d’une loi D1, à queue plus lourde qu’une loi D2, sont plus élevés que ceux estimés à partir de D2.
4. ESTIMATION DE L’INDICE DES VALEURS EXTRÊMES

L’autre approche, semi-paramétrique, n’est pas très utilisée en hydrologie. Cette approche est liée à la notion de domaine d’attraction maximum discutée dans la section 3.3. En effet, les méthodes d’estimation basées sur cette approche ont pour but d’estimer seulement l’indice des valeurs extrêmes puisque c’est ce paramètre qui détermine la forme de la queue de la distribution. Ces techniques permettent donc de déterminer la distribution asymptotique du maximum par le biais du théorème de Fisher-Tippet. En pratique, on ne peut pas considérer la plus grande observation, de l’échantillon, toute seule ; on considère donc les \( k \) plus grandes observations qui caractérisent la distribution du maximum, où \( k < n \) est grand. Le choix de \( k \) n’est pas une tâche facile. Plusieurs méthodes ont été développées, pour le choix de \( k \), mais aucune n’est adoptée d’une manière générale.


5. TECHNIQUES DE CHOIX DE LA CLASSE DES EXTRÊMES

En parallèle des études de classification des distributions en fonction de leur queue droite, on a besoin de critères et de tests pour identifier la classe qui représente le mieux la forme de la distribution des extrêmes en pratique. Il est donc nécessaire de disposer de méthodes qui permettent, sur la base d’un échantillon de données, de déterminer la classe des distributions à laquelle appartient la distribution d’origine des observations. L’objectif de cette partie, qui est très importante pour toute inférence statistique sur les extrêmes, est de présenter certaines méthodes qui permettent de caractériser la distribution de la variable étudiée. Ces méthodes peuvent être classées en trois groupes : méthodes graphiques, critères et tests statistiques. La majorité des critères et tests statistiques (Kolmogorov-Smirnov, Khi-Deux, Anderson-Darling, Cramer-Von Mises, Critère d’Information d’Akaike (AIC), Critère d’information Bayesien (BIC)...) permettent de tester l’ajustement des distributions à l’échantillon au niveau de la partie centrale. Dans ce travail on présente quelques méthodes graphiques qui ont été développées pour caractériser les différentes classes des distributions à queue lourde. Ces techniques sont tirées des récents travaux sur la théorie des valeurs extrêmes (Beirlant et al. 2004, Heyde et Kou 2004, Beirlant et al. 2006 et Goegebeur et al. 2006).

Afin d’illustrer l’utilisation pratique de chacune des méthodes présentées, on considère quatre distributions appartenant à la classe des distributions subexponentielles : la loi Halphen type A (HA, la classe III du tableau 1) ; la loi Lognormale (classe II) ; La loi de Fréchet (GEV) et Halphen type B⁻¹ (HIB), (la classe I ou la classe C des distributions à variations régulières qui est un sous-ensemble de la classe D des distributions subexponentielles). Nous avons considéré trois
asymétries $C_s = 1.20 ; 1.84$ et 3.26, pour représenter différentes asymétries souvent observées en hydrologie (Tableau 2). Nous avons fixé les paramètres de la loi HIB et ensuite on a calculé les paramètres des lois LN et GEV qui ont les mêmes caractéristiques statistiques (moyenne, variance et asymétrie). Nous n’avons pas trouvé de triplet de paramètres de la loi HA pour représenter le troisième cas (Cas 3) avec les mêmes caractéristiques statistiques que les autres lois. En effet, la famille des lois de Halphen représente un système complet de lois (à chaque échantillon il correspond une seule distribution de cette famille, Perreault et al. 1999a) et il est difficile de trouver deux lois de la famille Halphen qui ont exactement les mêmes caractéristiques statistiques.

**Tableau 2 : Cas considérés pour illustrer les différentes méthodes graphiques.**

<table>
<thead>
<tr>
<th>$C_s$</th>
<th>HIB</th>
<th>Fréchet</th>
<th>LN</th>
<th>HA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cas 1</td>
<td>$\nu = 5$</td>
<td>$u = 31$</td>
<td>$m = 14$</td>
<td>$\nu = 5$</td>
</tr>
<tr>
<td>$C_s = 1.20$</td>
<td>$\alpha = 3$</td>
<td>$\alpha = 5.5$</td>
<td>$\mu = 2.9$</td>
<td>$\alpha = 8$</td>
</tr>
<tr>
<td></td>
<td>$m = 100$</td>
<td>$\kappa = -0.01$</td>
<td>$\sigma = 0.34$</td>
<td>$m = 100$</td>
</tr>
<tr>
<td>Cas 2</td>
<td>$\nu = 3$</td>
<td>$u = 37$</td>
<td>$m = 18$</td>
<td>$\nu = 1$</td>
</tr>
<tr>
<td>$C_s = 1.84$</td>
<td>$\alpha = 3$</td>
<td>$\alpha = 7.8$</td>
<td>$\mu = 3$</td>
<td>$\alpha = 0.5$</td>
</tr>
<tr>
<td></td>
<td>$m = 100$</td>
<td>$\kappa = -0.10$</td>
<td>$\sigma = 0.46$</td>
<td>$m = 100$</td>
</tr>
<tr>
<td>Cas 3</td>
<td>$\nu = 2$</td>
<td>$u = 41$</td>
<td>$m = 19$</td>
<td>-----</td>
</tr>
<tr>
<td>$C_s = 3.26$</td>
<td>$\alpha = 3$</td>
<td>$\alpha = 10.3$</td>
<td>$\mu = 3.2$</td>
<td>-----</td>
</tr>
<tr>
<td></td>
<td>$m = 100$</td>
<td>$\kappa = -0.17$</td>
<td>$\sigma = 0.53$</td>
<td>-----</td>
</tr>
</tbody>
</table>
Méthode 1 : Graphique Log-log

Pour les distributions de type exponentiel de moyenne $\theta$, la fonction de survie

$$\bar{F}(u) = P(X > u) = e^{-u/\theta}$$

et pour les distributions de type puissance :

$$\bar{F}(u) = P(X > u) \approx C \int_u^{\infty} \frac{1}{x^\alpha} \, dx = C \frac{1}{\alpha - 1} u^{-\alpha+1}$$

En considérant $\log(P(X > u))$, on obtient respectivement pour les deux types de distributions $-\frac{u}{\theta}$ et $\log(C) - (\alpha - 1) \log(u)$. Ainsi, en portant sur un graphique les valeurs de $\log(P(X > u))$ en fonction de $\log(u)$, on devrait obtenir une courbe linéaire pour les distributions de type puissance, et concave pour les distributions de type exponentiel.

En pratique, on calcule $\log[P(X > u)]$ à l'aide de la fonction de répartition empirique pour certaines valeurs de $u$. On trace ensuite le graphique Log-Log de la probabilité de la queue de la distribution et la décision est prise à partir de l'allure de la courbe (Figure 4).
Figure 5 : Illustration du graphique Log-Log pour la discrimination entre la classe C et D.

Cette technique utilise le même principe que le graphique quantile-quantile appelée QQ-plot. La différence entre les deux techniques est que pour le graphique Log-Log on n’a pas besoin d’estimer les paramètres de la distribution à ajuster, étape indispensable pour le QQ-plot où on doit calculer les estimateurs des quantiles à partir de la distribution considérée. Étant donné que l’objectif est de déterminer une classe qui représente le mieux la série de données sans préciser une distribution particulière, la méthode Log-Log est plus adéquate pour répondre à ce but que le QQ-plot.

La figure 5 présente le graphique Log-Log pour des échantillons de taille $n = 50$ générés à partir des distributions du tableau 2. On remarque que pour la loi Halphen type A, ce graphique permet, pour les deux cas considérés, de conclure qu’il ne s’agit pas de distribution à variations régulières. Cependant, ce graphique ne permet pas de discriminer entre la loi Lognormale et les
Chapitre 5, Techniques de choix de la classe des extrêmes

lois à variations régulières (HIB et Fréchet). En effet, on remarque que pour les trois cas, la courbe correspondant à la loi LN est linéaire et ressemble aux deux autres courbes des lois HIB et Fréchet.

**Figure 6 : Graphique Log-Log pour les distributions du tableau 2.**

Les résultats obtenus dans cette partie rejoignent la discussion faite à la section 2, sur les processus générateurs des lois de type puissance (variations régulières) et de la loi Lognormale. Les deux processus sont très semblables et dans certains cas il est difficile de discriminer entre les deux, surtout pour les asymétries plus grandes que 1.
Méthode 2 : La moyenne empirique des excès

Cette méthode est basée sur la moyenne des excès $e(u) = E[X - u | X > u]$. Cette fonction est constante pour les distributions de type exponentiel et égale à $\frac{u}{(\alpha - 2)}$ ($\alpha > 2$) pour les distributions de type puissance.

En traçant la valeur de cette fonction $\hat{e}(u) = \frac{1}{k} \sum_{i=1}^{k} (x[i] - u)$ où $x[i]$ sont les valeurs de $X$ plus grandes que $u$, on devrait avoir une courbe linéaire pour des distributions de type exponentielle et puissance, sinon on est en présence d’un autre type de comportement. Si la pente de cette droite est nulle, la distribution est de type exponentiel. Si la distribution est de type subexponentielle, cette droite a une pente positive et passe par l’origine. Lorsque la pente est négative, il s’agit de queue plus légère que celle de la loi exponentielle.

![Figure 7 : Graphique de la moyenne empirique des excès pour quelques distributions du tableau 1.](image)

*Figure 7 : Graphique de la moyenne empirique des excès pour quelques distributions du tableau 1.*
La figure 6 illustre l’utilisation de cette méthode graphique pour l’identification de la classe des queues. On remarque que pour les quatre lois définies dans le tableau 2 (Cas1) le graphique basé sur la moyenne des excès, permet de caractériser les distributions subexponentielles, c.à.d. permet de discriminer entre la classe D et E. Cependant, elle ne permet pas de dire s’il s’agit de lois à variations régulières (classe C) ou non.

**Méthode 3 : Graphique des rapports du maximum et de la somme**

Cette méthode est basée sur le rapport du maximum et de la somme (max-sum ratio) défini par :

\[
R_n(p) = \frac{\max(X_1^p, \ldots, X_n^p)}{\sum_{i=1}^{n} X_i^p}
\]

où \( p \) est une puissance. Ce rapport tend vers 0 lorsque \( n \) tend vers l’infini si et seulement si \( E[X^p] < \infty \). En traçant le graphique de \( R_n(p) \) en fonction de \( n \) pour différentes valeurs de \( p \), on peut déterminer la forme de la distribution. En effet, si ce rapport tend vers l’infini pour les grandes valeurs de \( n \), on peut croire que \( E[X^p] = \infty \) et on est donc en présence d’une distribution de type puissance (Cf. Section 3.4). Sinon, la distribution est de type exponentiel. En pratique, on présente la fonction \( R_n \) en fonction de \( n \) pour différentes valeurs de \( p \). Si la courbe ne tend pas vers zéro à partir d’une certaine valeur \( p_0 \), alors la distribution appartient à la classe C et \( p_0 \) correspond à l’indice des valeurs extrêmes. La figure 7 illustre l’utilisation de ce graphique pour le cas 2 pour les lois du tableau 2 (pour les deux autres cas on obtient des résultats similaires). Chaque petit graphique, correspond à une valeur particulière \( p = 1, 2, \ldots, 9 \).
Figure 8 : Graphique du rapport de la somme et du maximum pour les lois Lognormale, GEV et HIB (Cas 2).

On remarque que pour la loi Lognormale, le rapport de la somme et du maximum a une forme régulière et tend vers zéro quand \( n \) devient assez grand. On rappelle que la loi LN appartient à la classe II (distributions subexponentielles (D) mais n’appartenant pas à la classe des distributions à variations régulières (C)). Pour la loi de Fréchet (GEV avec \( \kappa < 0 \)) et HIB, les courbes ne gardent plus la même forme (convergence vers zéro) à partir d’un certain ordre \( p_0 \). Pour la loi GEV \( p_0 \) est autour de 3 alors que pour la loi HIB \( p_0 \) est presque égale à 2. Cet ordre correspond au plus grand degré pour lequel le moment est fini et peut donc être considéré pour l’estimation de l’indice des valeurs extrêmes.
Méthode 4 : Graphique du rapport de Hill

Soit le rapport de Hill :

\[ a_n(x_k) = \frac{\sum_{i=1}^{n} I(X_i > x_k)}{\sum_{i=1}^{n} \log(X_i/x_k) * I(X_i > x_k)} \]  \hspace{1cm} (14)

où \( I(X > x) = \begin{cases} 1 & \text{si } X > x \\ 0 & \text{sinon} \end{cases} \),

et les \( X_i \) sont les valeurs de la variable \( X \) et \( x_k \) est la \( k \)ième plus grande valeur de \( X \). Sur le graphique de \( a_n(x_k) \) en fonction de \( x_k \), on cherche une région stable pour déterminer l’estimateur de la queue de la distribution. Cette approche est illustrée dans la figure 8 pour les distributions définies dans le tableau 2.
Caractérisation des Distributions à Queue Lourde pour l’analyse des crues

Figure 9 : Graphique du rapport de Hill pour les lois Lognormale, GEV, HA et HIB.

On remarque que la partie extrême droite du graphique (Figure 8) est stable dans le cas des distributions GEV et HIB, alors que pour les lois Lognormale et Halphen type A, la courbe décroît vers zéro surtout pour les cas 2 et 3 (cas de fortes asymétries).

Notons que l’estimateur de Hill, comme plusieurs autres estimateurs, a été développé à partir de la linéarité du graphique Log-Log ou celui de la moyenne empirique des excès. La pente de la partie linéaire de ces graphiques peut être considérée pour l’estimation de l’indice des valeurs extrêmes pour les distributions de la classe de Pareto (classe B) ou la classe C des distributions à

Méthode 5 : Statistique de Jackson

Beirlant et al. (2006) ont présenté un nouveau test pour l’identification des distributions de type Pareto (classe B) ou appartenant à la classe C d’une manière générale. Ce test, basé sur la statistique de Jackson, a été développé comme test d’ajustement pour la loi exponentielle (Jackson 1967). Le fait qu’une transformation logarithmique d’une variable distribuée suivant une loi de Pareto suit une loi exponentielle a permis l’extension de ce test pour l’identification des variables qui ont un comportement de Pareto (Classe B). Après avoir modifié la statistique de Jackson et l’avoir appliqué aux observations de l’extrémité droite de l’échantillon, Beirlant et al. (2006) ont déterminé la loi limite de la statistique du test ainsi qu’une correction du biais dans le cas des échantillons de taille finie. On rappelle ici la version avec correction du biais de la statistique de Jackson adaptée pour l’identification d’un comportement de Pareto. On illustre son utilisation pour des séries simulées à partir des lois LN, Fréchet, HA et HIB (Tableau 2).

La statistique du test est :

\[
T_k^* = \frac{1}{k} \sum_{j=1}^{k} C_{k-j+1,k} Z_j \quad \frac{H_{k,n}}{H_{k,n}}
\]

où \( Z_j = j \left( \log X_{n-j+1,n} - \log X_{n-j,n} \right) \), \( j = 1, \ldots k \) ; \( H_{k,n} = \frac{1}{k} \sum_{j=1}^{k} Z_j \) et \( C_{k-j+1,k} = 1 - \log \left( \frac{j+1}{k+1} \right) \).
D’un point de vue pratique, ce test est basé sur le même principe que celui du rapport de Hill. En effet, on présente les valeurs de $T_k^*$ en fonction de $k$ et on cherche si au niveau des grandes valeurs de $k$ la courbe reste constante (convergence vers la moyenne de la statistique du test).

Le biais observé lors de l’utilisation de la statistique $T_k^*$ pour des séries de taille fini, rend la discrimination, entre les lois de type Pareto et les autres lois subexponentielles, difficile. Beirlant et al. (2006) ont donc développé une version de ce test avec correction du biais. Ils ont remarqué que les deux quantités au numérateur et dénominateur de $T_k^*$ (équation 15) correspondent à des moyennes pondérées des écarts logarithmiques des statistiques d’ordre $Z_j$, $j=1,\ldots,k$. Or, sous certaines conditions de régularité une approximation de $Z_j$ est donnée sous la forme suivante (Beirlant et al. 1999):

$$Z_j = \gamma + b_{n,k} \left( \frac{j}{k+1} \right)^{-\rho} + \varepsilon_j, \quad j=1,\ldots,k$$

(16)

où $b_{n,k}$ est fonction de $\frac{n}{k}$ et $\varepsilon_j$, $j=1,\ldots,k$ représentent des erreurs centrées en zéro. Cette approximation mène à la statistique $\tilde{T}_k$ qui correspond à la statistique de Jackson avec correction du Biais :

$$\tilde{T}_k = \frac{1}{k} \sum_{j=1}^{k} C_{k-j+1,k} \left( Z_j - \hat{b}_{LS,k} \left( \hat{\rho} \right) \left( \frac{j}{k+1} \right)^{-\hat{\rho}} \right) \frac{\hat{\gamma}_{LS,k} \left( \hat{\rho} \right)}{\hat{\gamma}_{LS,k} \left( \hat{\rho} \right)}$$

(17)
Avec \( \hat{\rho} \) un estimateur consistant de \( \rho \) et \( \hat{b}_{LS,k}(\hat{\rho}) \) et \( \hat{\gamma}_{LS,k}(\hat{\rho}) \) sont les estimateurs par moindres carrées de \( b_{n,k} \) et \( \gamma \) respectivement, obtenus à partir de l’équation (16). Pour les distributions de type Pareto, la statistique \( \sqrt{k} \left( \hat{\gamma}_k(\rho) - 2 \right) \) \( \overset{\text{Loi}}{\rightarrow} N \left( 0, \left( \frac{\rho}{1-\rho} \right)^2 \right) \) et converge pour \( k \) assez grand vers sa moyenne 2.

Pour \( \rho \) fixé,

\[
\hat{\gamma}_{LS,k}(\rho) = \frac{1}{k} \sum_{j=1}^{k} Z_j - \frac{\hat{b}_{LS,k}(\hat{\rho})}{1-\rho}
\]

\[
\hat{b}_{LS,k}(\hat{\rho}) = \frac{(1-\rho)^2(1-2\rho)}{\rho^2} \frac{1}{k} \sum_{j=1}^{k} \left( \frac{j}{k+1} \right)^{-\rho} - \frac{1}{1-\rho} Z_j
\]

Et comme estimateur de \( \rho \), on peut considérer par exemple l’estimateur suivant (Beirlant et al. 2006) :

\[
\hat{\rho}_k = -\frac{1}{\log \lambda} \log \frac{H_{[\hat{\lambda k}]n} - H_{[\hat{\lambda k}]n}}{H_{[\hat{k}]n} - H_{[\hat{k}]n}}, \quad \text{(18)}
\]

pour un \( \lambda \in (0,1) \) et \( \hat{k} \) tel que \( \sqrt{k} b \left( \frac{n}{\hat{k}} \right) \rightarrow \infty \) comme proposé par Drees et Kaufmann (1998).

Pour l’implémentation pratique de ce test, Beirlant et al. (2006) ont considéré différentes valeurs pour estimer \( \rho \) (par exemple \( \rho = -1 \) et \(-2 \)). Après plusieurs simulations nous avons remarqué que pour \( \rho = -1 \) on obtient de bons résultats.
La figure 9 illustre l’utilisation de cette méthode pour l’identification des distributions de type Pareto. En plus de la statistique $\tilde{T}_k$, on représente les intervalles de confiance (au niveau 10%), de $\tilde{T}_k$, obtenus par simulation de 1000 échantillons de taille 100, pour le Cas2 du tableau 2 pour les distributions Lognormale, GEV (Fréchet), HA et HIB. On remarque que pour les lois de Fréchet et HIB, la statistique du test est stable et converge vers la valeur 2 (qui est la moyenne de la statistique du test de Jackson), alors qu’elle ne l’est pas pour la loi HA. Cependant, on remarque que pour la loi LN la statistique de Jackson converge vers sa moyenne et cette convergence est stable. On signale donc la même difficulté pour discriminer entre les distributions de la classe C (distributions à variations régulières) et la loi Lognormale. Notons que les intervalles de confiance, obtenus par Bootstrap paramétrique, sont indispensables pour évaluer la robustesse de la convergence de la statistique du test de Jackson vers sa moyenne 2.

Figure 10 : Test basé sur la statistique de Jackson pour les lois Lognormale, GEV, HA et HIB (Cas 2 du Tableau 2).
6. APPLICATIONS À DES CAS RÉELS


![Figure 10: Coefficients de variation et d’asymétrie de 32 séries tirées du site de UNESCO](image)

**Figure 11 : Coefficients de variation et d’asymétrie de 32 séries tirées du site de UNESCO**

La période d’observation des débits pour ces rivières dépasse 70 ans. Certaines séries ont plus de 150 ans d’observations. Cette taille est beaucoup plus élevée que celle des séries souvent rencontrées en hydrologie.
Quatre stations ont été choisies pour cette étude. Nous avons considéré les séries pour lesquelles le graphique Log-Log semble être linéaire (Figure 11).

![Graphique Log-Log pour les quatre séries étudiées](image)

**Figure 12 : Graphique Log-Log pour les quatre séries étudiées**

Ce choix permettra de comparer les résultats obtenus à partir des différentes méthodes pour discriminer entre la classe des distributions subexponentielles et celles à variations régulières. Les séries considérées sont :

1. Canada_SouthSaskatchewan_Saskatoon (CSS)
2. Canada_StMaurice_GrandMere (CSG)
3. Roumanie_Danube_Orsova (RDO)
4. USA_Susquehanna_harrisburg (USH).
Plusieurs tests ont été considérés pour vérifier les hypothèses de base de l’analyse fréquentielle. La figure 12 présente les autocorrélations d’ordre 1 à 10 des quatre séries considérées dans cette section. On remarque que pour les séries de débit maximum annuel, des stations RDO et USH, les autocorrélations d’ordre supérieur ou égal à 1 sont toutes non-significatives au niveau de confiance 5% (les autocorrélations ne dépassent pas la valeur critique de la statistique du test au niveau 5%). Alors que pour les deux autres, CSS et CSG, pour certains ordres, le seuil correspondant au niveau de signification 5% a été dépassé. Cependant au niveau 1%, l’hypothèse d’indépendance est acceptée pour les quatre séries. La stationnarité des séries étudiées, a été vérifiée par le test de Mann-Kendall. Les résultats montrent, qu’au seuil de confiance 1% toutes les séries sont stationnaires sauf celle de la station CSG. Étant donné que l’objectif principal de cette partie est d’illustrer l’utilisation des différentes méthodes sur des cas réels, on considère l’application de ces techniques sur ces quatre séries.
On remarque qu’à partir de l’ensemble de ces graphiques (Figure 11 et les Figures 13 à 16) on peut avoir une idée claire sur la classe de la distribution la plus adéquate pour représenter chacune des quatre séries étudiées.

![Figure 14 : Fonction de la moyenne des excès pour les quatre séries](image)

Ce constat est clair pour les séries CSG et USH. Pour ces deux séries on remarque que toutes les méthodes montrent qu’il s’agit de lois à variations régulières. En effet pour ces deux séries on remarque qu’il s’agit bien d’une distribution de la classe C avec un indice de valeurs extrêmes faible.
Ceci peut être tiré de la figure 14 du rapport du maximum et la de somme, qui ne tend pas vers zéro à partir de l’ordre $p = 3$ pour la série CSG et $p = 4$ pour USH. Les figures 15 (rapport de Hill) et 16 (test de Jackson) confirment aussi qu’il s’agit de distributions à variations régulières.
Caractérisation des Distributions à Queue Lourde pour l’analyse des crues

Figure 16 : Graphique basé sur le rapport de Hill

Figure 17 : Graphique basé sur la statistique de Jackson
Pour évaluer l’erreur commise lors d’un choix d’une loi Lognormale (qui n’appartient pas à la classe C) à la place d’une loi de type puissance, comme la loi de Fréchet (classe C), ou l’inverse, on présente dans cette section les ajustements des quatre séries avec les lois Lognormale et Fréchet.

**Comparaison des ajustements pour la station CSS**

Nombre d'observations: 73

Quantiles obtenus à partir de la distribution Fréchet (Maximum de vraisemblance)

<table>
<thead>
<tr>
<th>Paramètres</th>
<th>alpha = 258.00</th>
<th>k = -0.17</th>
<th>u = 555.55</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>q=1-1/T</td>
<td>XT</td>
<td>Ecart-type</td>
</tr>
<tr>
<td>10000.0</td>
<td>0.9999</td>
<td>6260</td>
<td>3010</td>
</tr>
<tr>
<td>2000.0</td>
<td>0.9995</td>
<td>4540</td>
<td>1690</td>
</tr>
<tr>
<td>1000.0</td>
<td>0.9990</td>
<td>3930</td>
<td>1280</td>
</tr>
<tr>
<td>200.0</td>
<td>0.9950</td>
<td>2760</td>
<td>634</td>
</tr>
<tr>
<td>100.0</td>
<td>0.9900</td>
<td>2350</td>
<td>449</td>
</tr>
<tr>
<td>10.0</td>
<td>0.9000</td>
<td>1260</td>
<td>112</td>
</tr>
<tr>
<td>2.0</td>
<td>0.5000</td>
<td>653</td>
<td>40.0</td>
</tr>
</tbody>
</table>
Quantiles obtenus à partir d’une loi LN3 (Maximum de vraisemblance) :

Paramètres : m = 62.30 ,  mu = 6.38 et  sigma = 0.56.

<table>
<thead>
<tr>
<th>T</th>
<th>Q</th>
<th>XT</th>
<th>Ecart-type</th>
<th>Intervalle de confiance (95%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10000.0</td>
<td>0.9999</td>
<td>4800</td>
<td>1250</td>
<td>2350</td>
</tr>
<tr>
<td>2000.0</td>
<td>0.9995</td>
<td>3790</td>
<td>848</td>
<td>2130</td>
</tr>
<tr>
<td>1000.0</td>
<td>0.9990</td>
<td>3400</td>
<td>703</td>
<td>2020</td>
</tr>
<tr>
<td>200.0</td>
<td>0.9950</td>
<td>2560</td>
<td>426</td>
<td>1730</td>
</tr>
<tr>
<td>100.0</td>
<td>0.9900</td>
<td>2240</td>
<td>331</td>
<td>1590</td>
</tr>
<tr>
<td>10.0</td>
<td>0.9000</td>
<td>1270</td>
<td>110</td>
<td>1060</td>
</tr>
<tr>
<td>2.0</td>
<td>0.5000</td>
<td>654</td>
<td>41.1</td>
<td>574</td>
</tr>
</tbody>
</table>

Les résultats obtenus pour la station CSS, montre que les quantiles estimés à partir de la loi LN sont inférieurs à ceux obtenus à partir de la loi de Fréchet. Ce qui est normal, vu que la deuxième a une queue plus lourde relativement à la première. La différence est de 14% pour une période de retour 1000 ans et atteint 30 % pour une période de 10000 ans.

**Comparaison des ajustements pour la station CSG**

Nombre d'observations: 83

Quantiles obtenus à partir de la distribution Fréchet (Maximum de vraisemblance)

Paramètres alpha = 611.13  k = -0.20 et  u = 1766.07
Chapitre 6, Applications à des cas réels

<table>
<thead>
<tr>
<th>T</th>
<th>Q</th>
<th>XT</th>
<th>Ecart-type</th>
<th>Intervalle de confiance (95%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10000.0</td>
<td>0.9999</td>
<td>1.85E+004</td>
<td>8900</td>
<td>-</td>
</tr>
<tr>
<td>2000.0</td>
<td>0.9995</td>
<td>1.30E+004</td>
<td>4760</td>
<td>-</td>
</tr>
<tr>
<td>1000.0</td>
<td>0.9990</td>
<td>1.11E+004</td>
<td>3550</td>
<td>-</td>
</tr>
<tr>
<td>200.0</td>
<td>0.9950</td>
<td>7620</td>
<td>1680</td>
<td>4330</td>
</tr>
<tr>
<td>100.0</td>
<td>0.9900</td>
<td>6440</td>
<td>1170</td>
<td>4160</td>
</tr>
<tr>
<td>10.0</td>
<td>0.9000</td>
<td>3510</td>
<td>271</td>
<td>2980</td>
</tr>
<tr>
<td>2.0</td>
<td>0.5000</td>
<td>2000</td>
<td>90.0</td>
<td>1820</td>
</tr>
</tbody>
</table>

Quantiles obtenus à partir d’une loi LN3 (Maximum de vraisemblance) :

Paramètres $m = 489.44$, $\mu = 7.33$ et $\sigma = 0.53$

<table>
<thead>
<tr>
<th>T</th>
<th>Q</th>
<th>XT</th>
<th>Ecart-type</th>
<th>Intervalle de confiance (95%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10000.0</td>
<td>0.9999</td>
<td>1.17E+004</td>
<td>2680</td>
<td>6430</td>
</tr>
<tr>
<td>2000.0</td>
<td>0.9995</td>
<td>9400</td>
<td>1830</td>
<td>5810</td>
</tr>
<tr>
<td>1000.0</td>
<td>0.9990</td>
<td>8500</td>
<td>1530</td>
<td>5510</td>
</tr>
<tr>
<td>200.0</td>
<td>0.9950</td>
<td>6580</td>
<td>934</td>
<td>4750</td>
</tr>
<tr>
<td>100.0</td>
<td>0.9900</td>
<td>5820</td>
<td>729</td>
<td>4390</td>
</tr>
<tr>
<td>10.0</td>
<td>0.9000</td>
<td>3540</td>
<td>248</td>
<td>3050</td>
</tr>
<tr>
<td>2.0</td>
<td>0.5000</td>
<td>2030</td>
<td>95.8</td>
<td>1840</td>
</tr>
</tbody>
</table>

La différence entre les deux lois est beaucoup plus importante pour cette série, surtout pour les grandes périodes de retour. Elle peut atteindre 60% pour une période de retour 10000 ans. Même pour une période de retour 100 ans (la période de retour la plus utilisée), la différence est de 10%.
Comparaison des ajustements pour la station RDO

Nombre d'observations: 150

Quantiles obtenus à partir de la distribution Fréchet (Maximum de vraisemblance)

Paramètres  alpha = 1710.74  k = -0.05  et  u = 7929.29

<table>
<thead>
<tr>
<th>T</th>
<th>Q</th>
<th>XT</th>
<th>Ecart-type</th>
<th>Intervalle de confiance (95%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10000.0</td>
<td>0.9999</td>
<td>2.03E+004</td>
<td>2840</td>
<td>1.47E+004</td>
</tr>
<tr>
<td>2000.0</td>
<td>0.9995</td>
<td>1.86E+004</td>
<td>2030</td>
<td>1.46E+004</td>
</tr>
<tr>
<td>1000.0</td>
<td>0.9990</td>
<td>1.78E+004</td>
<td>1710</td>
<td>1.44E+004</td>
</tr>
<tr>
<td>200.0</td>
<td>0.9950</td>
<td>1.58E+004</td>
<td>1060</td>
<td>1.37E+004</td>
</tr>
<tr>
<td>100.0</td>
<td>0.9900</td>
<td>1.49E+004</td>
<td>830</td>
<td>1.33E+004</td>
</tr>
<tr>
<td>10.0</td>
<td>0.9000</td>
<td>1.16E+004</td>
<td>307</td>
<td>1.10E+004</td>
</tr>
<tr>
<td>2.0</td>
<td>0.5000</td>
<td>8550</td>
<td>172</td>
<td>8210</td>
</tr>
</tbody>
</table>

Quantiles obtenus à partir d’une loi LN3 (Maximum de vraisemblance) :

Paramètres  m = 438.44  mu = 9.00  et  sigma = 0.23.

<table>
<thead>
<tr>
<th>T</th>
<th>Q</th>
<th>XT</th>
<th>Ecart-type</th>
<th>Intervalle de confiance (95%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10000.0</td>
<td>0.9999</td>
<td>2.02E+004</td>
<td>1900</td>
<td>1.65E+004</td>
</tr>
<tr>
<td>2000.0</td>
<td>0.9995</td>
<td>1.83E+004</td>
<td>1420</td>
<td>1.55E+004</td>
</tr>
<tr>
<td>1000.0</td>
<td>0.9990</td>
<td>1.74E+004</td>
<td>1230</td>
<td>1.50E+004</td>
</tr>
<tr>
<td>200.0</td>
<td>0.9950</td>
<td>1.55E+004</td>
<td>840</td>
<td>1.38E+004</td>
</tr>
<tr>
<td>100.0</td>
<td>0.9900</td>
<td>1.46E+004</td>
<td>689</td>
<td>1.33E+004</td>
</tr>
<tr>
<td>10.0</td>
<td>0.9000</td>
<td>1.15E+004</td>
<td>297</td>
<td>1.09E+004</td>
</tr>
<tr>
<td>2.0</td>
<td>0.5000</td>
<td>8600</td>
<td>170</td>
<td>8260</td>
</tr>
</tbody>
</table>
Pour ce troisième exemple la différence n’est pas très importante. Les deux distributions donnent les mêmes résultats et ce pour toutes les périodes de retour. On rappelle que pour cette série, les critères de choix de la classe des extrêmes (rapport de Hill et statistique de Jackson, Figures 14 et 15) montrent qu’il ne s’agit pas d’une distribution de la classe C (des distributions à variations régulières). Cependant l’utilisation du graphique Log-Log, seul, montre qu’elle l’est.

**Comparaison des ajustements pour la station USH**

Nombre d'observations: 93

Quantiles obtenus à partir de la distribution Fréchet (Maximum de vraisemblance)

Paramètres alpha = 543.29 k = -0.05 et u = 2399.99

<table>
<thead>
<tr>
<th>T</th>
<th>Q</th>
<th>XT</th>
<th>Ecart-type</th>
<th>Intervalle de confiance (95%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10000.0</td>
<td>0.9999</td>
<td>8870</td>
<td>2340</td>
<td>N/D N/D</td>
</tr>
<tr>
<td>2000.0</td>
<td>0.9995</td>
<td>7500</td>
<td>1500</td>
<td>N/D N/D</td>
</tr>
<tr>
<td>1000.0</td>
<td>0.9990</td>
<td>6940</td>
<td>1210</td>
<td>4560 9320</td>
</tr>
<tr>
<td>200.0</td>
<td>0.9950</td>
<td>5730</td>
<td>682</td>
<td>4390 7060</td>
</tr>
<tr>
<td>100.0</td>
<td>0.9900</td>
<td>5230</td>
<td>511</td>
<td>4230 6240</td>
</tr>
<tr>
<td>10.0</td>
<td>0.9000</td>
<td>3700</td>
<td>159</td>
<td>3390 4010</td>
</tr>
<tr>
<td>2.0</td>
<td>0.5000</td>
<td>2600</td>
<td>71.6</td>
<td>2460 2740</td>
</tr>
</tbody>
</table>
Caractérisation des Distributions à Queue Lourde pour l’analyse des crues

Quantiles obtenus à partir d’une loi LN3 (Maximum de vraisemblance) :

Paramètres $m = 913.61$   $\mu = 7.43$ et $\sigma = 0.38$

<table>
<thead>
<tr>
<th>T</th>
<th>Q</th>
<th>XT</th>
<th>Ecart-type</th>
<th>Intervalle de confiance (95%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10000.0</td>
<td>0.9999</td>
<td>8100</td>
<td>1290</td>
<td>5560</td>
</tr>
<tr>
<td>2000.0</td>
<td>0.9995</td>
<td>7000</td>
<td>926</td>
<td>5180</td>
</tr>
<tr>
<td>1000.0</td>
<td>0.9990</td>
<td>6540</td>
<td>787</td>
<td>5000</td>
</tr>
<tr>
<td>200.0</td>
<td>0.9950</td>
<td>5520</td>
<td>509</td>
<td>4530</td>
</tr>
<tr>
<td>100.0</td>
<td>0.9900</td>
<td>5100</td>
<td>407</td>
<td>4300</td>
</tr>
<tr>
<td>10.0</td>
<td>0.9000</td>
<td>3700</td>
<td>156</td>
<td>3400</td>
</tr>
<tr>
<td>2.0</td>
<td>0.5000</td>
<td>2610</td>
<td>73.0</td>
<td>2470</td>
</tr>
</tbody>
</table>

Halphen de type B inverse (Maximum de vraisemblance)

Paramètres $\alpha = 0.57$ $m = 5446$ $\nu = 4.13$

<table>
<thead>
<tr>
<th>T</th>
<th>Q</th>
<th>XT</th>
<th>Ecart-type</th>
<th>Intervalle de confiance (95%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10000.0</td>
<td>0.9999</td>
<td>9920</td>
<td>1290</td>
<td>5560</td>
</tr>
<tr>
<td>2000.0</td>
<td>0.9995</td>
<td>7990</td>
<td>926</td>
<td>5180</td>
</tr>
<tr>
<td>1000.0</td>
<td>0.9990</td>
<td>7260</td>
<td>787</td>
<td>5000</td>
</tr>
<tr>
<td>200.0</td>
<td>0.9950</td>
<td>5800</td>
<td>509</td>
<td>4530</td>
</tr>
<tr>
<td>100.0</td>
<td>0.9900</td>
<td>5270</td>
<td>407</td>
<td>4300</td>
</tr>
<tr>
<td>10.0</td>
<td>0.9000</td>
<td>3690</td>
<td>156</td>
<td>3400</td>
</tr>
<tr>
<td>2.0</td>
<td>0.5000</td>
<td>2600</td>
<td>73.0</td>
<td>2470</td>
</tr>
</tbody>
</table>
Pour cette quatrième série, nous avons étudié l’ajustement à partir de trois distributions : Les lois LN, Fréchet et HIB. Rappelons que les deux dernières lois appartiennent à la même classe (classe C). L’objectif est de comparer les résultats obtenus pour deux distributions de la même classe. On remarque que, la différence est significative entre les quantiles obtenus à partir de la loi Fréchet et LN, surtout pour les périodes de retour supérieures ou égales à 1000 ans. Notons aussi, que les quantiles estimés à partir de la loi HIB sont plus élevés que ceux obtenus par la loi Fréchet. La différence est de 4% pour $T=1000$ et de 11% pour $T=10000$. Il faut donc utiliser d’autres critères pour la comparaison des distributions appartenant à la même classe.

**Discussion des résultats**

On remarque que pour les quatre séries, les quantiles obtenus à partir de la loi Lognormale sont inférieurs à ceux obtenus par la loi Fréchet. La différence est importante, surtout pour les événements de période de retour assez élevée. Nous avons ajusté à la série USH la loi de Halphen type $B^1$. L’objectif est de comparer les quantiles estimés à partir de deux lois appartenant à la même classe. On remarque que la différence entre ces deux lois est assez importante au niveau des grandes périodes de retour.

Cet exemple montre que la détermination de la classe de lois qui peut représenter le mieux une série de données, est une étape importante pour le choix du meilleur ajustement mais n’est pas suffisante pour avoir une bonne estimation des événements extrêmes. En effet, et comme il a été déjà mentionné, la méthode d’estimation des paramètres est une étape importante pour le processus d’ajustement. Comme conclusion, on constate que les techniques de sélection de la
classe constituent un outil important pour le choix de la classe la plus adéquate. Cependant, d’autres critères doivent être considéré pour le choix à l’intérieur de chaque classe. Parmi ces critères, il y a la méthode d’estimation des paramètres. Si par exemple, pour une des lois de la même classe on dispose d’une technique d’estimation des paramètres qui est optimale (la variance des estimateurs est minimale), c’est cette distribution qu’il est préférable dans le processus d’ajustement.
7. CONCLUSION

Le problème de choix de la distribution des séries hydrologiques constitue un des sujets les plus étudiés dans la littérature. La majorité des études mentionnent la nécessité de développer des outils pratiques pour caractériser les queues des distributions. En effet, au niveau de la partie centrale des distributions, plusieurs tests statistiques permettent de mesurer la qualité de l’ajustement. Cependant au niveau de la queue, on ne dispose pas de techniques efficaces de choix de la distribution la plus adéquate.

Le problème des queues de distributions est lié à la théorie des valeurs extrêmes. Cependant d’autres distributions autres que celles des valeurs extrêmes doivent être considérées étant donné que l’application des résultats asymptotiques est limitée par les faibles tailles des séries hydrologiques.

En tenant compte des processus physiques qui génère les extrêmes, nous avons ramené le problème à la discrimination entre les distributions de type puissance (présentées aussi sous le nom de distributions à variations régulières) et les autres distributions de la classe subexponentielle. Dans ces classes on retrouve les différentes distributions utilisées en hydrologie, telles que présentées dans la classification donnée par Ouarda et al. (1994).

Dans ce rapport, nous avons présenté des techniques graphiques qui permettent d’identifier les séries issues d’une distribution à variation régulières. On s’intéresse particulièrement à cette classe, car un mauvais choix mène à une sous estimation significative surtout pour les grandes périodes de retour. Les graphiques présentés, sont les plus utilisés dans la théorie des extrêmes.
Les illustrations faites sur des séries générées montrent que ces outils permettent de discriminer entre la classe des lois à variations régulières (classe C) et des autres lois subexponentielle (classe D), ce qui constitue un premier pas important. En effet ces procédures, permettent d’éliminer plusieurs distributions et limite l’ensemble des lois à une classe donnée. Cependant, nous avons remarqué que dans la majorité des cas étudiés, on n’arrive pas à séparer la loi Lognormale des autres lois de la classe C. Ceci est dû (1) au fait que les processus physiques générateurs de ces lois, sont très semblables et (2) à la taille des échantillons qui est souvent faible (\( n = 50 \) pour les études par simulation).

Un autre problème de l’analyse fréquentielle est l’estimation des paramètres. En effet, pour chaque distribution, plusieurs méthodes d’estimation des paramètres sont disponibles, mais leur efficacité dépend de la taille de l’échantillon et de la distribution elle même. Certains travaux ont étudié les propriétés de certaines méthodes d’estimation pour la majorité des distributions utilisées en hydrologie, mais ils arrivent à la conclusion suivante : aucune méthode n’est la meilleure dans tous les cas. Cependant, pour certaines classes de lois, les estimateurs du maximum de vraisemblance sont efficaces. On recommande l’utilisation des deux étapes suivantes : (1) Déterminer la classe des distributions qui semblent représenter le mieux la série de données, (2) parmi les lois appartenant à la même classe choisir celle qui possède une méthode efficace pour l’estimation des paramètres.
RÉFÉRENCES


A.1 Loi Lognormale

La distribution obtenue par une transformation logarithmique, la plus connue en hydrologie est la loi lognormale. Si X est une variable aléatoire distribuée suivant une loi Lognormale distribution, alors, $Z = \ln(X)$ est distribuée suivant une loi normale. Le terme « loi Lognormale » est souvent utilisé pour désigner la loi Lognormale à deux paramètres. Il y a aussi la loi Lognormale à trois paramètres.

A.1.1 Loi Lognormale à deux paramètres

Si X est distribuée suivant une loi Lognormale $LN(\mu, \sigma)$ alors la variable :

$$Z = \ln X$$  \hspace{1cm} (A1)

est distribuée suivant une loi Normale de moyenne $\mu$ et de variance $\sigma^2$.

La fonction de densité de probabilité de $Z$ est :

$$\Phi(z) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{z-\mu}{\sigma} \right)^2}$$  \hspace{1cm} (A2)
On en déduit la fonction de densité de probabilité \( f \) de \( X \) à partir de la relation suivante :

\[
f(x) = \Phi \left( \frac{z(x)}{dx} \right) \frac{dz}{dx}
\]  
(A3)

On obtient donc l’expression de \( f \) :

\[
f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2} \left[ \frac{(\ln(x) - \mu)}{\sigma} \right]^2}
\]  
(A4)

La moyenne et la variance sont données par les expressions suivantes:

Moyenne \( E[X] = \exp \left( \mu + \frac{1}{2} \sigma^2 \right) \)  
(A5)

Variance \( E \left[ (X - E(X))^2 \right] = \exp \left( 2\mu + \sigma^2 \right) (e^{\sigma^2} - 1) \)  
(A6)

Le coefficient de variation, qui est fonction uniquement de \( \sigma^2 \), est donné par :

\[
CV = \left( e^{\sigma^2} - 1 \right)^{\frac{1}{2}}
\]  
(A7)

Le coefficient d’asymétrie est :

\[
C_v = \left( e^{\sigma^2} - 1 \right)^{\frac{1}{2}} \left( e^{\sigma^2} + 2 \right)
\]  
(A8)
A.1.2 La loi Lognormale à trois paramètres (LN3)

La loi Lognormale à trois paramètres (LN3) diffère de LN2 par l’introduction d’une borne inférieure $x_0$ telle que :

$$z = \ln(x - x_0) \quad (A9)$$

Suit alors une distribution normale de paramètres $\mu$ et $\sigma^2$. Il s’agit donc d’une distribution à trois paramètres, où $x_0$ est un paramètre d’origine et les paramètres $\mu$ et $\sigma^2$ contrôlent l’échelle et la forme. On peut déduire de l’équation (A9) que la fonction de densité de probabilité d’une variable aléatoire $X$ distribuée suivant une loi $LN3(x_0, \mu, \sigma^2)$ est donnée par :

$$f(x) = \frac{1}{(x-x_0)\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\ln(x-x_0)-\mu)/\sigma^2} \quad ; \quad \text{avec } x \geq x_0 \quad (A10)$$

Les moments de $X$ peuvent être obtenus à partir de ceux de la loi LN2 parce que les deux distributions ne diffèrent que d’un paramètre de location on a alors :

Moyenne $= E[X] = x_0 + \exp\left(\mu + \frac{1}{2}\sigma^2\right) \quad (A11)$

Variance $= E\left[(X-E(X))^2\right] = \exp\left(2\mu + \sigma^2\right)(e^{\sigma^2} - 1) \quad (A12)$
A.2 Distribution Généralisée des Valeurs Extrêmes

La distribution généralisée des valeurs extrêmes généralisée (GEV) est une des distributions à trois paramètres les plus utilisées en hydrologie. La fonction de répartition de la loi GEV est donnée par (Jenkinson, 1955):

\[ F(x) = \exp \left[ -\left(1 - \kappa \frac{(x-u)}{\alpha} \right)^\frac{1}{\kappa} \right] \quad (A15) \]

Une autre reparamétrisation a été proposée par Von Mises (1954) :

\[ F(x) = \exp \left[ -\left(1 + \xi \frac{(x-u)}{\alpha} \right)^{-\frac{1}{\xi}} \right]. \]

En hydrologie on utilise surtout la paramétrisation de Jenkinson, pour laquelle les asymétries positives correspondent aux valeurs négative du paramètre de forme \( \kappa \). La famille des distributions des extrêmes (présentée dans la section 2), contient les trois distributions présentées dans ce travail : la loi Gumbel, la loi de Fréchet et la loi de Weibull appelées aussi EV1, EV2 et EV3, respectivement. Chacune de ces trois lois est caractérisée par les valeurs du paramètre de forme \( \kappa \). Les valeurs négatives correspondent à la loi EV2, le cas nul à la loi EV1 et le cas des valeurs positives correspondent à la loi EV3.
A.2.1 Loi des Valeurs Extrêmes type I (EV1 ou Gumbel)

La fonction de répartition d’une loi Gumbel est donnée par :

\[ F(x) = \exp\left(-e^{-(x-\mu)/\alpha}\right) \]  \hspace{1cm} (A16)

\( \mu \) est un paramètre de position et \( \alpha \) un paramètre d’échelle. La moyenne et la variance sont données par :

Moyenne = \( \mu + 0.5772\alpha \)  \hspace{1cm} (A17)

Variance = \( \frac{\pi^2\alpha^2}{6} \)  \hspace{1cm} (A18)

Le coefficient d’asymétrie est constant pour une loi Gumbel et est égale à 1.14. Si on considère la transformation réduite d’une loi Gumbel : \( Y = (X - \mu) / \alpha \) on retrouve la fonction de répartition donnée dans l’équation (7).

A.2.2 Loi des Valeurs Extrêmes type II (EV2 ou Fréchet)

La fonction de répartition de la loi des valeurs extrêmes type 2 (EV2), connue aussi sous le nom de loi de Fréchet ou encore de loi Log-Gumbel, est donnée par :

\[ F(x) = \exp\left\{-\left[1 - \kappa(x - \mu)/\alpha\right]^{\frac{1}{\kappa}}\right\} \]  \hspace{1cm} (A19)
où $k < 0, \alpha > 0$ et $\mu + \frac{\alpha}{k} \leq x \leq \infty$

Cette distribution a trois paramètres: $\mu$ le paramètre de position, $\alpha$ le paramètre d’échelle et $k$ le paramètre de forme qui est négatif. La variable a une borne inférieure $\mu + \frac{\alpha}{k}$. La fonction de densité de probabilité est

$$f(x) = \frac{1}{\alpha} \left(1 - \frac{x - \mu}{\alpha}\right) \exp\left\{-\left[1 - \frac{\kappa(x - \mu)}{\alpha}\right]^{\frac{1}{\kappa}}\right\}$$

(A20)

A.2.3 Loi des Valeurs Extrêmes type III (EV3 ou Weibull)

La fonction de répartition d’une loi Weibull est donnée par :

$$F(x) = \exp\left\{-\left[1 - \frac{\kappa(x - \mu)}{\alpha}\right]^{\frac{1}{\kappa}}\right\}$$

(A21)

où $k > 0, \alpha > 0, -\infty \leq x \leq \mu + \frac{\alpha}{\kappa}, \kappa > 0, \alpha > 0$ et $\mu \in \mathbb{R}$

Une variable distribuée suivant une loi de Weibull a une borne supérieure : $-\infty \leq x \leq \mu + \frac{\alpha}{\kappa}$.

La fonction de densité de probabilité est :

$$f(x) = \frac{1}{\alpha} \left(1 - \frac{x - \mu}{\kappa}\right)^{-1} \exp\left\{-\left[1 - \frac{\kappa(x - \mu)}{\alpha}\right]^{\frac{1}{\kappa}}\right\}$$

(A22)
A.3 Famille des lois de Halphen

La loi Halphen type A (HA)

\[ f_A(x) = \frac{1}{2m^{\nu} k_\nu(2\alpha)} x^{\nu-1} \exp \left[ -\alpha \left( \frac{x + m}{x} \right) \right], \quad x > 0 \]  

(A23)

où \( m > 0 \) est un paramètre d’échelle, \( \alpha > 0 \) et \( \nu \in \mathbb{R} \) sont des paramètres de forme. La constante \( k_\nu(2\alpha) \) correspond à la fonction de Bessel modifiée de deuxième espèce (Watson 1996). Une revue complète sur les propriétés de cette distribution est faite dans Perreault et al. (1999a et 1999b).

La loi Inverse Gaussienne Généralisée (GIG) (Jorgensen 1982, Seshadri 1993) correspond, à une paramétrisation près, à la loi Halphen type A. En effet, la f.d.p. de la loi GIG est :

\[ g(z) = \frac{1}{2k_\nu(\sqrt{\chi \psi})} z^{\nu-1} \exp \left\{ -\frac{1}{2} \left( \frac{\chi + \psi z}{z} \right) \right\}. \]

Si nous posons \( \alpha = \frac{1}{2} \sqrt{\chi \psi} \) et \( m = \sqrt{\frac{\chi}{\psi}} \), on retrouve la f.d.p. de la loi Halphen type A (équation (2)).

La loi Halphen type B (HB)

\[ f_B(x) = \frac{2}{m^{2\nu} e f_\nu(\alpha)} x^{2\nu-1} \exp \left[ -\left( \frac{x}{m} \right)^2 + \alpha \left( \frac{x}{m} \right) \right], \quad x > 0 \]  

(A24)
où $m > 0$ est un paramètre d’échelle et $\alpha \in \mathbb{R}^+$ et $\nu > 0$ sont des paramètres de forme. La fonction de normalisation $e_{\nu}(.)$ a été définie par Halphen (1955) et appelée fonction exponentielle factorielle.

**La loi Halphen type B inverse (HB$^{-1}$)**

$$f_{B^{-1}}(x) = \frac{2}{m^{2\nu} e_{\nu}(\alpha)} x^{-2\nu-1} \exp \left[ -\left( \frac{m}{x} \right)^2 + \alpha \left( \frac{m}{x} \right) \right], \quad x > 0$$

(A25)

où $m > 0$ est le paramètre d’échelle et $\alpha \in \mathbb{R}^+$ et $\nu > 0$ sont des paramètres de forme.

Remarque : Si $X$ suit une distribution $HB(x; m, \alpha, \nu)$ alors $Y=1/X$ suit $HIB(y; m^{-1}, \alpha, \nu)$.

**Références (Annexe A) :**


Dans cette annexe on présente quelques théorèmes qui donnent des conditions nécessaires et suffisantes d’appartenance au domaine d’attraction maximum d’une des distributions des valeurs extrêmes (Embrechts et al. 2003).

Les notations utilisées dans cette annexe sont les suivants :

Le point terminal d’une loi de probabilité $F$ : $S(F) = \sup \{ x : F(x) < 1 \}$

Le point initial : $I(F) = \inf \{ x : F(x) > 0 \}$.

**Théorème 1 :** (Domaine d’attraction maximum de Fréchet) $F$ appartient au domaine d’attraction maximum de Fréchet si et seulement si : $S(F) = +\infty$, et pour tout $x > 0$

$$\lim_{t \to +\infty} \frac{1 - F(tx)}{1 - F(t)} = x^{-\frac{1}{\tau}} \quad (B1)$$

Comme nous l’avons déjà mentionné à la section 3.3, les distributions appartenant à ce domaine d’attraction maximum correspondent à la classe des distributions à variations régulières.

**Exemple :** (Loi de Cauchy)
La fonction de répartition d’une variable aléatoire $X$ distribuée suivant une loi de Cauchy est donnée par :

$$F(x) = \frac{1}{2} + \frac{1}{\pi} \arctan(x)$$

on a donc $0 < F(x) < 1$ et $S(F) = +\infty$.

Au voisinage de l’infini, on a l’approximation suivante : $\arctan(x) \approx \frac{\pi}{2} + \frac{1}{3x^3} - \frac{1}{5x^5}$, ce qui donne :

$$1 - F(x) = -\frac{1}{3\pi x^3} + \frac{1}{5\pi x^5}, \quad \text{alors :} \quad \frac{1 - F(tx)}{1 - F(t)} = -\frac{1}{3(tx)^3} + \frac{1}{5(tx)^5} \quad \text{et} \quad \frac{1 - F(tx)}{1 - F(t)} \underset{+\infty}{\longrightarrow} \frac{-1}{3(x)^3} + \frac{1}{5(x)^5} \underset{+\infty}{\longrightarrow} x^{-3}$$

L’équation (B1) est donc vérifiée pour $\tau = \frac{1}{3}$. Ce qui montre que la loi de Cauchy appartient au domaine d’attraction maximum de Fréchet.

**Théorème 2 :** (Domaine d’attraction maximum de Weibull) $F$ appartient au domaine d’attraction maximum de Weibull si et seulement si : $S(F) < +\infty$, et la fonction définie pour tout $x > 0$ par :

$$F^*(x) = F\left(S(F) - \frac{1}{x}\right),$$

vérifie la condition pour tout $x > 0$ :

$$\lim_{t \to +\infty} \frac{1 - F^*(tx)}{1 - F^*(t)} = x^\tau \quad (B2)$$
**Exemple :** (Loi Uniforme)

La fonction de répartition d’une variable aléatoire $X$ distribuée suivant une loi Uniforme sur l’intervalle $[0,1]$ est :

$$F(x) = \int_{-\infty}^{x} I_{[0,1]}(t) \, dt ,$$

et on a : $F(1) = 1$, donc $S(F) = 1 < +\infty$ .

Pour $t$ suffisamment grand, $S(F) - \frac{1}{t}$ et $S(F) - \frac{1}{tx}$ appartiennent à l’intervalle $[0,1]$ ce qui donne, pour tout $x > 0$ :

$$\frac{1 - F^*(tx)}{1 - F^*(t)} = \frac{1}{x} = x^{-1}$$

L’équation (B2) est donc vérifiée pour $\tau = -1$. On déduit donc que la loi Uniforme appartient au domaine d’attraction maximum de Weibull.

**Théorème 3 :** (Domaine d’attraction maximum de Gumbel) $F$ appartient au domaine d’attraction maximum de Gumbel si et seulement si :

- il existe un réel $r$ fini tel que :

$$\int_{r}^{S(F)} \left[ 1 - F(t) \right] \, dt < +\infty \quad (B3)$$

- la fonction $R$ définie sur $]I(F) , S(F)[$ par :

$$R(t) = \frac{1}{1 - F(t)} \int_{t}^{S(F)} \left[ 1 - F(u) \right] \, du ,$$

vérifie pour tout réel $x$ :

$$\lim_{t \to S(F)} \frac{1 - F(t + xR(t))}{1 - F(t)} = \exp(-x) \quad (B4)$$
**Exemple :** (Loi Exponentielle)

La fonction de répartition d’une variable aléatoire $X$ distribuée suivant une loi Exponentielle est :

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

La loi exponentielle appartient au domaine d’attraction maximum de Gumbel. En effet, $I(F) = 0$, $S(F) = +\infty$ et pour $r = 0$ on a :

$$\int_0^{+\infty} \exp(-\lambda t) dt = \frac{1}{\lambda} < +\infty,$$

L’équation (B3) est donc vérifiée.

Pour $t \in ]0, +\infty[$ on a $R(t) = \frac{1}{\lambda}$, alors pour tout réel $x$ :

$$\frac{1 - F\left(\frac{t + \frac{x}{\lambda}}{\lambda}\right)}{1 - F(t)} = \frac{\exp(-\lambda t - x)}{\exp(-\lambda t)} = \exp(-x),$$

L’équation (B4) est vérifiée et la loi exponentielle appartient au domaine d’attraction maximum de Gumbel.

**Références (Annexe B):**

ANNEXE C     STATIONS OBTENUES DU SITE DE L’UNESCO

(http://www.unesco.org/water/ihp/db/index_fr.shtml)

Différentes stations tirées du site de l’UNESCO des plus importantes rivières du monde

<table>
<thead>
<tr>
<th>Abr.</th>
<th>Station</th>
<th>Taille de la Série</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADH</td>
<td>Allemagne_Donau (Danube)_Hofkirchen_Tr</td>
<td>86</td>
</tr>
<tr>
<td>CAB</td>
<td>Canada_Assiniboine_Brandon_Tr</td>
<td>80</td>
</tr>
<tr>
<td>CAH</td>
<td>Canada_Assiniboine_Headingley_Tr</td>
<td>73</td>
</tr>
<tr>
<td>CFH</td>
<td>Canada_Fraser_Hope_Tr</td>
<td>74</td>
</tr>
<tr>
<td>CNP</td>
<td>Canada_NorthSaskatchewan_PrinceAlbert_Tr</td>
<td>73</td>
</tr>
<tr>
<td>CRE</td>
<td>Canada_Red_Emerson_Tr</td>
<td>72</td>
</tr>
<tr>
<td>CSI</td>
<td>Canada_Saguenay_IsleMaligne_Tr</td>
<td>68</td>
</tr>
<tr>
<td>CSM</td>
<td>Canada_SouthSaskatchewan_MedecineHat_Tr</td>
<td>70</td>
</tr>
<tr>
<td>CSS</td>
<td>Canada_SouthSaskatchewan_Saskatoon_Tr</td>
<td>73</td>
</tr>
<tr>
<td>CSP</td>
<td>Canada_StJohn_Pokiok_Tr</td>
<td>66</td>
</tr>
<tr>
<td>CSG</td>
<td>Canada_StMaurice_GrandMere_Tr</td>
<td>83</td>
</tr>
<tr>
<td>CWS</td>
<td>Canada_Winnipeg_SlaveFalls_Tr</td>
<td>75</td>
</tr>
<tr>
<td>FKP</td>
<td>Finelande_Kymijoki_Pernoo_Tr</td>
<td>85</td>
</tr>
<tr>
<td>FVI</td>
<td>Finelande_Vuoksi_Imatra_Tr</td>
<td>138</td>
</tr>
<tr>
<td>FVP</td>
<td>Finelande_Vuoksi_Pernoo_Tr</td>
<td>138</td>
</tr>
<tr>
<td>HDN</td>
<td>Hongrie_Duna Danube_Nagymaros_Tr</td>
<td>96</td>
</tr>
<tr>
<td>IKV</td>
<td>INDE_Krishna_Vijayawada_Tr</td>
<td>74</td>
</tr>
<tr>
<td>LNS</td>
<td>Lithuania_Neman_Smalininkai_Tr</td>
<td>170</td>
</tr>
<tr>
<td>MND</td>
<td>Mali_niger_dire_Tr</td>
<td>55</td>
</tr>
<tr>
<td>MNK</td>
<td>Mali_niger_koulkoro_Tr</td>
<td>81</td>
</tr>
<tr>
<td>POG</td>
<td>Pologne_Odra_Gozdowiece_Tr</td>
<td>87</td>
</tr>
<tr>
<td>RDO</td>
<td>Roumanie_Danube_Orsova_Tr</td>
<td>150</td>
</tr>
<tr>
<td>RNN</td>
<td>Russie_Neva_Novosaratovkla_Tr</td>
<td>125</td>
</tr>
<tr>
<td>RSU</td>
<td>Russie_SevernayaDvina_Ust-Pinega_Tr</td>
<td>112</td>
</tr>
<tr>
<td>SSB</td>
<td>Senegal_senegal_bakel_Tr</td>
<td>81</td>
</tr>
<tr>
<td>SVV</td>
<td>Suede_Vanerngota_Vanesborg_Tr</td>
<td>178</td>
</tr>
<tr>
<td>TLD</td>
<td>Tchechoslovaqie_Labe_Decin_Tr</td>
<td>134</td>
</tr>
<tr>
<td>ULFG</td>
<td>UK_Lee_FeildesWeir_gauged</td>
<td>118</td>
</tr>
<tr>
<td>ULFN</td>
<td>UK_Lee_FeildesWeir_naturalised</td>
<td>116</td>
</tr>
<tr>
<td>UTKG</td>
<td>UK_Thames_kingston_gauged</td>
<td>120</td>
</tr>
<tr>
<td>UTKN</td>
<td>UK_Thames_kingston_naturalised</td>
<td>120</td>
</tr>
<tr>
<td>USH</td>
<td>USA_Susquehanna_harrisburg_Tr</td>
<td>82</td>
</tr>
</tbody>
</table>
Position géographique des stations obtenues à partir du site de l’UNESCO.