République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université A. MIRA de Béjaïa
Faculté des Sciences Exactes
Département de Recherche Opérationnelle

Mémoire de fin d'étude

En vue de l'obtention d'un Master en Recherche Opérationnelle

Option

Fiabilité et Evaluation des Performances dans les Réseaux

Thème

Estimation non paramétrique des densités Heavy-tailed par la méthode du noyau. Application au trafic web

Présenté par :
AFROUN Fairouz
MERAH Sabrina

Soutenu le 27/06/2013 devant le jury composé de :

Présidente	LAGHA Karima	M.C.B	U. de Béjaïa
Rapporteur	ADJABI Smail	Professeur	U. de Béjaïa
Examinateurs	ZOUGAB Nabil	M.C.B	U. de Tizi-Ouzou
	SEMCHEDINE Fouzi	M.C.B	U. de Béjaïa
	ZIANE Yasmina	Doctorante	U. de Béjaïa
Invité	CHERFAOUI Mouloud	M.A.A	U. de Biskra

Promotion 2012-2013

Remerciements

Nous remercions le Dieu pour le courage, la patience et la volonté qui nous ont été utiles tout au long de notre parcours.

Nous tenons à remercier Mr ADJABI Smail pour la proposition du thème, l'encadrement de ce travail, pour ses précieux conseils et orientations.

Nous remercions également Mr CHERFAOUI Mouloud pour ses remarques pertinentes qui nous ont permis de mieux structurer notre travail et de mieux le décrire.

Nous remercions également les membres du jury pour avoir accepté d'examiner et d'évaluer notre travail.

Nos sincères remerciements s'adressent enfin à tous ceux qui nous ont soutenu de près ou de loin.

Dédicaces

A mes chers parents

A mon cher mari

A mes chers frères et mes chères sœurs

A mes chers beaux parents

A toute ma famille et tous mes proches

A tous mes amis, et ma chère binôme Sabrina

Je dédie ce travail.

AFROUN Fairouz

Dédicaces

A celle qui m'a transmis la vie, l'amour, le courage, à toi chère maman toutes mes joies, mon amour et ma reconnaissance.

A mon père pour l'éducation qu'il m'a prodigué; avec tous les moyens et au prix de toutes les sacrifices qu'il a consentis à mon égard et à mes études depuis mon enfance.

A mes chers frères Abdelghani, Karim et Youcef

A mes chères sœurs Fouzia, Khira, Saliha et Djahida

A mes neveux Nasraddine, Fatima, Amani et Ouassim

A mes belles-sœurs Hakima, Faïza et Ratiba

A mon beau-frère

A mes amies

A mes condisciples et surtout à Fairouz

A toutes ces personnes, sincèrement: Merci!

Sabrina MERAH

Table des matières

Liste des tableaux

Table des figures

Introduction Générale

Jusqu'à nos jours la plupart des modèles statistiques utilisés pour décrire différents phénomènes, sont des modèles statistiques classiques tel que : la distribution de Poisson, la distribution exponentielle, etc. Ces modèles ont l'avantage d'être simple à utiliser, mais ils montrent vite leurs limites. En effet, il est difficile de modéliser avec précision des phénomènes complexes uniquement à l'aide des modèles statistiques classiques. Pour faire face à certaines de ces limites, on fait appel aux lois puissance. Une loi puissance permet de modéliser certains phénomènes qui ne pourraient être approximés avec une précision suffisante au moyen des modèles classiques. En effet, de nombreuses études ont montré que de telles distributions se retrouvaient sur la fréquence des signes en linguistique, les revenus en économie, la taille des villes, la durée de vie en biologie ou encore les fluctuations turbulentes ou la distribution des degrés des nœuds de nombreux graphes.

Les distributions en "lois puissance" fascinent des chercheurs de toutes disciplines. Cependant, dans la pratique, il n'est pas évident d'identifier leurs distributions par les méthodes paramétriques. Pour pallier les insuffisances et les défauts des familles paramétriques, une seconde approche dite non paramétrique propose de "laisser parler les données". Actuellement, il existe plusieurs méthodes non paramétriques pour l'estimation de la densité de probabilité, on peut citer : la méthode de l'histogramme, méthode d'estimation par les séries orthogonales et la méthode du noyau.

Dans la littérature plusieurs auteurs ont abordé le sujet de l'estimation non paramétrique d'une densité à queue lourde par la méthode du noyau. On cite le récent travail de A. SADOUN et Y. ZIANE [52], où les auteurs ont proposé de modéliser un trafic web à l'aide d'une densité à queue lourde afin d'évaluer ses caractéristiques statistiques en utilisant la méthode du noyau. En effet, après avoir sélectionné un sous échantillon à partir des données du serveur de la coupe du monde France 98, dans un premier lieu, l'utilisation de l'estimateur de Hill a permis de conclure que la distribution de l'échantillon en question est à queue lourde. En deuxième lieu, les auteurs ont proposé d'utiliser la méthode du noyau asymétrique pour l'estimation de la densité de l'échantillon sélectionné après l'échec du test d'ajustement K.S. pour l'identification d'une loi usuelle correspondante à la distribution

de l'échantillon.

Dans le présent travail, nous proposons de reprendre les mêmes données (globales), c'est-à-dire celles du trafic web du serveur de la coupe du monde 98, comme exemple d'application pour l'estimation d'une densité à queue lourde par la méthode du noyau. Pour réaliser cette étude, nous avons suggéré de décomposer nos données selon le type afin d'évaluer les caractéristiques statistiques (Indice de queue, densité,...) de chaque type et cela pour deux principales raisons : d'une part pour déterminer l'influence des caractéristiques de chaque type sur les caractéristiques du trafic global et d'autre part pour estimer les distributions de ces données.

Après une analyse statistique préliminaire, on classifie les distributions dans différentes classes (queue lourde, queue fine, sans queue) en utilisant la distribution de survie et QQ-plot. Dans le but de confirmer les précédentes classifications on a utilisé la méthode non paramétrique de Hill et la méthode paramétrique de l'estimateur du maximum de vraisemblance MLE (pour estimer les paramètres de la distribution des valeurs extrêmes généralisée GEV). En deuxième lieu, en choisissant quelques noyaux asymétriques et des paramètres de lissage optimaux, nous avons obtenu différents estimateurs des densités de la taille des différents types de fichiers.

Ce mémoire est composé d'une introduction, de trois chapitres et d'une conclusion. Le premier chapitre présente des généralités sur les lois puissance et heavy-tailed, ensuite les méthodes d'estimation de l'indice de variabilité (méthodes paramétrique, semi-paramétrique et non paramétrique). Le deuxième chapitre présente l'estimation non paramétrique de la densité de probabilité.
Dans le troisième chapitre, nous exposons la démarche d'analyse et de modélisation d'un exemple d'un échantillon du trafic web ainsi que les résultats numériques et les résultats graphiques obtenus. Ce travail se termine par une conclusion générale et quelques perspectives.

Chapitre 1

Outils statistiques pour la modélisation

Contents

1.1 Introduction

Dans ce chapitre, nous allons présenter quelques lois puissance qui ont déjà été utilisées pour modéliser des phénomènes extrêmement divers. Nous allons en particulier nous intéresser aux lois puissance du type heavy-tailed, de la loi de Zipf et de la loi de Pareto qui établissent une relation entre les fréquences d'apparition des différentes occurrences d'un phénomène et le rang de ces occurrences dans une suite ordonnée.

Avant de présenter ces différentes études, il convient de préciser ce que nous entendons par le terme de loi puissance. Nous allons ensuite présenter les principales notions de la théorie des valeurs extrêmes qui nous permettra d'extrapoler le comportement de la queue de distribution des données à partir des plus grandes données observées ainsi que les méthodes de mesure et d'estimation de l'indice de queue. Enfin, nous présenterons les tests d'ajustement.

1.2 Lois puissance

Les distributions en "lois puissance" fascinent des chercheurs de toutes disciplines. De nombreuses études ont montré que de telles distributions se retrouvaient sur la fréquence

des signes en linguistique, les revenus en économie, la taille des villes, la durée de vie en biologie ou encore les fluctuations turbulentes ou la distribution des degrés des nœuds de nombreux graphes, dont Internet.

Définition 1.1. (La loi de puissance [48])

La loi de puissance est une relation entre deux ensembles d'observation x et y, satisfaisant : $y = ax^k$.

✓ a est la constante de proportionnalité ;

✓ k est l'exposant de la loi de puissance.

La représentation d'une loi de puissance dans un repère (x, y) est donnée par la figure 1.1 :

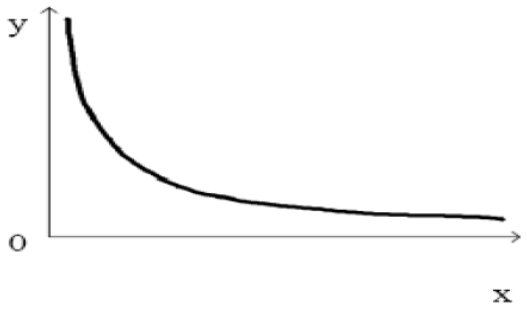

FIGURE 1.1 – Représentation d'une loi de puissance dans un repère (x, y).

Le graphe d'une loi de puissance est une droite sur une échelle log-log :

$$log(y) = klog(x) + log(a) \tag{1.1}$$

comme il est représenté dans la figure 1.2 :

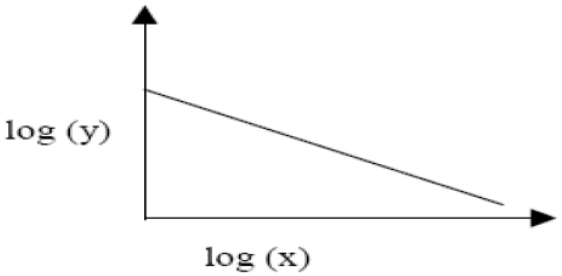

FIGURE 1.2 – Représentation d'une loi de puissance sur une échelle log-log.

1.2.1 La distribution Heavy-Tailed (Queue lourde)

Définition 1.2. Soit X une variable aléatoire réelle définie sur un espace de probabilité (Ω,A,P). On appelle fonction de répartition de X la fonction $F(x)$ de $\mathbb{R}$ dans $[0,1]$ définie par

$$F_X(x) = P(X \leq x), x \in \mathbb{R} \tag{1.2}$$

et on appelle fonction de survie de X la fonction $\overline{F}_x$ de $\mathbb{R}$ dans $[0,1]$ définie par

$$\overline{F}_X(x) = 1 - F_X(x), x \in \mathbb{R} \tag{1.3}$$

Définition 1.3. Soit X une variable aléatoire réelle définie sur un espace (Ω,A,P). On dit que X est à queue lourde si sa fonction de survie est telle que

$$\overline{F}_{(x)} \sim x^{-\alpha}, \tag{1.4}$$

au voisinage de ∞ pour un $\alpha > 0$.

On peut immédiatement remarquer que cette définition implique que la fonction de survie de X est à variation régulière d'indice "$-\alpha$". Dans la pratique, cet indice de variation régulière α sera pris dans l'intervalle $]0,2[$. Il est également à noter qu'on ne s'intéresse pas vraiment à la variable aléatoire elle-même mais plutôt à sa loi, comme le suggère la définition en termes de fonction de survie. Nous dirons donc qu'une loi est à queue lourde si toute variable aléatoire distribuée selon cette loi est à queue lourde. Enfin, cette appellation de "queue lourde" se comprend aisément car l'étude du comportement asymptotique de la fonction de survie fournit bien des informations sur la queue de distribution de la variable aléatoire que l'on observe. De plus, la soi-disant lourdeur de cette queue vient du fait que cette fonction de survie décroît très lentement au voisinage de l'infini[44].

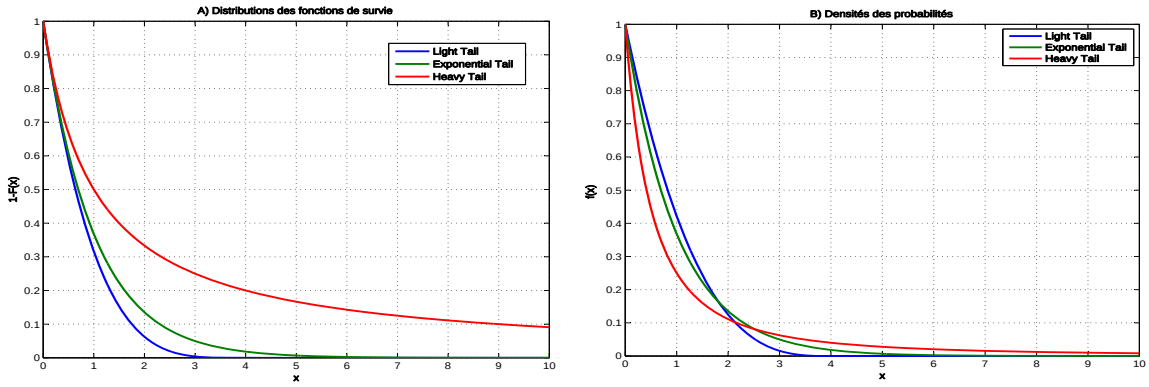

FIGURE 1.3 – Comparaison du comportement de queue : A) distribution de survie B) densité.

1.2.2 La loi de Zipf

La loi de Zipf est une loi empirique énoncée en 1949 par George Kingsley Zipf, elle est basée sur une approche d'utilisation d'une loi puissance. Elle décrit la répartition statistique des fréquences d'apparition des différents éléments d'un ensemble, comme par exemple les mots d'un texte. Selon Zipf, les n-uplets de symboles d'un ensemble organisés topologiquement ne s'organisent pas de manière aléatoire. Si l'on classe les n-uplets de symboles suivant l'ordre décroissant de leurs fréquences d' apparition, on obtient la suite $(F_1, F_2, F_3, ..., F_n)$ des fréquences d'apparition. La fréquence d'un n-uplet de rang i vérifie la formule suivante :

$$F(i) = ki^{\alpha}, \tag{1.5}$$

où k et α sont des constantes.

Cette loi puissance est caractérisée par la valeur α de l'exposant. Elle peut être représentée par un graphe à échelles logarithmiques. Sur cette représentation graphique que l'on appellera courbe de Zipf [18].

Il existe plusieurs phénomènes suivant une loi de puissance à la Zipf (Fréquence vs rang) : Fréquence d'accès des pages web, Population des villes, Trafic Internet par site, Noms dans une population, etc [28].

Exemple 1.1. *La figure 1.4 montre les courbes de Zipf obtenues pour les 1500 mots les plus fréquents dans les deux langues [10].*

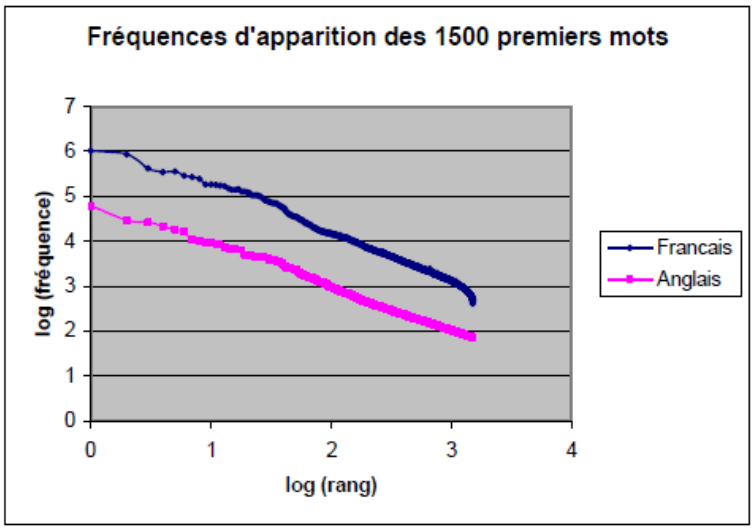

FIGURE 1.4 – Fréquence d'apparition des 1500 mots les plus fréquents en français et en anglais.

1.2.3 La loi de Pareto

Historiquement, la première mise en évidence d'une loi puissance a été faite en 1897 par l'économiste italien Vilfredo Pareto. En étudiant la répartition des revenus personnels des individus dans les principaux pays industrialisés, Pareto a constaté que cette répartition suivait une loi puissance.

La distribution de Pareto a des applications dans des différents domaines. La "fonction de Pareto" (ou "loi de Pareto") est la formalisation du principe des 80-20. Cet outil d'aide à la décision détermine les facteurs (environ 20%) cruciaux qui influencent la plus grande partie (80%) de l'objectif.

Cette loi est un outil fondamental et basique en gestion de la qualité (Génie Industriel et Techniques de Gestion). Elle est aussi utilisée en réassurance. La théorie des files d'attente s'est intéressée à cette distribution, lorsque des recherches des années 90 ont montré que cette loi régissait aussi au nombre de grandeurs observées dans le trafic Internet (et plus généralement sur tous les réseaux de données à grande vitesse).

Une variable aléatoire est dite par définition suit une loi de Pareto si sa fonction de répartition est donnée par :

$$P(X \leqslant x) = 1 - \left(\frac{\alpha}{x}\right)^k, x \geq \alpha, k > 0. \tag{1.6}$$

La fonction de densité (fonction de distribution) de Pareto est alors :

$$f(x) = \frac{d}{dx}\left(1 - \left(\frac{\alpha}{x}\right)^k\right) = k\frac{\alpha^k}{x^{k+1}} = k\alpha^k x^{-k-1}, \tag{1.7}$$

avec $k \in \mathbb{R}_+$ et $x \geq \alpha \geq 0$. La distribution de Pareto est donc définie par deux paramètres, α et k (nommé "index de Pareto").

Exemple 1.2. *Tracé de la fonction de distribution 1.5 et répartition 1.6 pour la fonction de Pareto de paramètres* $(x, \alpha, k) = (x, 1, 2)$.

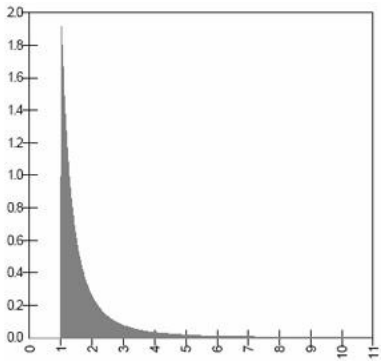

FIGURE 1.5 – Fonction de distribution pour la fonction de Pareto$(x, 1, 2)$

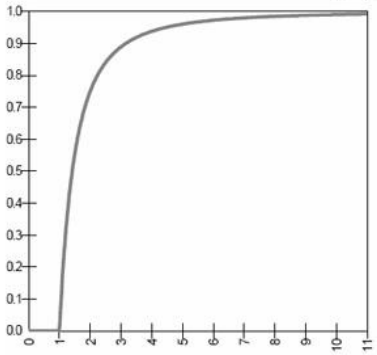

FIGURE 1.6 – Fonction de répartition pour la fonction de Pareto$(x, 1, 2)$

Il s'agit d'une loi à décroissance lente dont la distribution cumulative est proportionnelle à $x^{-\alpha}$, $\alpha > 0$. *Tel que :*

Pour $1 < \alpha < 2$, *la valeur moyenne est finie, mais la variance ne l'est pas.*

Pour $0 < \alpha < 1$, *la valeur moyenne elle-même devient infinie.*

Remarque 1.1. L'étude des valeurs extrêmes revient à l'étude des queues de distributions de fonctions, ou de façon équivalente, à l'analyse de la plus grande observation d'un échantillon. En ce sens, nous pouvons considérer la théorie des valeurs extrêmes comme la contrepartie de la théorie statistique classique, qui est principalement basée sur l'étude de la moyenne d'un échantillon plutôt que des observations extrêmes.

1.3 Théorie des valeurs extrêmes

Nous supposons donner n variables aléatoires $X_1, \ldots, X_n$ indépendantes et identiquement distribuées de fonction de répartition $\mathbf{F}$ définie par :

$$\mathbf{F}(x) = Pr(X \leq x). \tag{1.8}$$

Une manière simple d'étudier le "comportement" des évènements extrêmes est de considérer les variables aléatoires iid : $X_1, X_2, ..., X_n$, et M_n le maximum définit par :

$$M_n = max\{X_1, X_2, ..., X_n\}. \tag{1.9}$$

Nous adopterons la convention que le maximum est un nombre positif. Comme les variables aléatoires sont indépendantes et identiquement distribuées, on obtient :

$$Pr\{M_n \leq x\} = Pr\{X_1 \leq x, X_2 \leq x, ..., X_n \leq x\} = [\mathbf{F}(x)]^n. \tag{1.10}$$

La difficulté provient du fait que l'on ne connaît pas en général la fonction de répartition $\mathbf{F}$. C'est la raison pour laquelle on s'intéresse au comportement asymptotique de la variable aléatoire M_n. Ainsi en identifiant la famille de loi vers laquelle M_n va converger, on pourra remplacer $\mathbf{F}$ par cette dernière pour des grandes valeurs de n. Pour caractériser cette loi de distribution des extrêmes, nous allons recourir au Théorème de Fisher-Tippet.

1.3.1 Théorème limite de Fisher-Tippet

Avant d'énoncer le principal théorème de cette section, nous définissons des classes d'équivalences sur l'ensemble des fonctions de répartition (sur les distributions des probabilités). Les distributions $\mathbf{F}$ et $\mathbf{F}^*$ sont dites de même type si

$$\exists\ a, b \in \mathbb{R}, \forall x \in \mathbb{R}, \qquad \mathbf{F}^*(ax+b) = \mathbf{F}(x). \tag{1.11}$$

Nous pouvons à présent énoncer le théorème de Fisher-Tippet qui permet de caractériser la loi de distribution des extrêmes.

S'il existe des constantes $a_n > 0$ et b_n telles que

$$\lim_{n\to\infty} Pr\left\{\frac{M_n - b_n}{a_n} \leq x\right\} = \mathbf{G}(x). \tag{1.12}$$

Avec $\mathbf{G}$ une fonction de distribution non dégénérée, alors $\mathbf{G}$ appartient à l'un des trois types suivants (I, II, ou III)

Type I : Gumbel

$$\mathbf{G}(x) = \exp(-e^{-x}) \qquad \forall x \in \mathbb{R}$$

.

Type II : Fréchet

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

avec α, un paramètre de forme.

Type III : Weibull

$$\mathbf{G}(x) = \begin{cases} \exp(-(-x)^{-\alpha}), & \text{si } x \leq 0, \alpha > 0\,; \\ 1, & \text{si } x > 0. \end{cases}$$

La loi de Gumbel peut être considérée comme une loi de transition entre les lois de Fréchet et de Weibull. La majorité des lois de probabilité usuelles appartiennent à l'un des trois domaines d'attraction (Maximum Domain of Attraction(MDA)) : Gumbel, Fréchet ou Weibull. Par exemple, les distributions exponentielle, Gamma et Log-normale appartiennent au MDA de Gumbel regroupant la majorité des distributions à queue fine ; les distributions de Pareto, Log-Gamma, et Student appartiennent au MDA de Fréchet regroupant la majorité des distributions à queue lourde et la distribution uniforme appartient au MDA de Weibull regroupant la majorité des distributions sans queue.

En fait, les trois types de distribution précédents peuvent être caractérisés par une distribution unique

$$\mathbf{G}(x) = \exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{\frac{-1}{\xi}}\right\}. \tag{1.13}$$

Cette fonction de distribution correspond à la loi de probabilité des valeurs extrêmes généralisées, **"Generalized Extreme Value distribution" (GEV)**. Nous avons alors les correspondances suivantes :

Fréchet	$\xi = \alpha^{-1} > 0$
Weibull	$\xi = -\alpha^{-1} < 0$
Gumbel	$\xi \to 0$

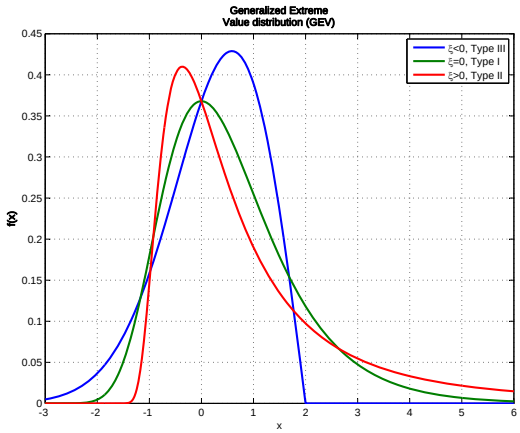

FIGURE 1.7 – Lois des valeurs extrêmes(GEV).

Nous remarquons que les paramètres μ et σ sont les limites de b_n et a_n.

Le paramètre σ joue le rôle d'une variance, c'est pourquoi nous le considérons comme un paramètre de dispersion. Le paramètre μ est un paramètre de localisation.

Le paramètre ξ est lié au caractère leptokurtique de la fonction de distribution **F**, c'est pourquoi on lui donne généralement le nom d'indice de queue ou d'indice de valeur extrême. Plus cet indice est élevé en valeur absolue, plus le poids des extrêmes dans la distribution initiale est important. On parle alors de distribution á " queues épaisses".

Mais notons toutefois qu'en pratique nous ne connaissons pas les paramètres μ, σ et ξ ; il faut donc les estimer à partir des données et les remplacer par leurs estimations. Il existe plusieurs méthodes pour estimer les paramètres de la distribution **GEV**. Par exemple, nous pouvons citer les méthodes d'estimation de l'indice de queue, la méthode des moments et la méthode du maximum de vraisemblance [41].

1.3.2 Méthode des excès et distribution de Pareto généralisée

Fondée sur la théorie des valeurs extrêmes, la méthode des excès également connue sous le nom de Peaks Over Threshold (POT), permet la modélisation des queues de distribution d'une série de données à partir de laquelle il devient possible d'estimer la probabilité d'occurrence d'événements rares au-delà des plus grandes valeurs observées. La fonction de distribution des excès par rapport à un seuil élevé μ est définie par :

$$\mathbf{F}_{\mu}(y) = Pr\{X - \mu \leq y / X > \mu\}, \tag{1.14}$$

pour $0 \leq y \leq x_0 - \mu$, on a généralement $x_0 = +\infty$.

La fonction de distribution des excès représente la probabilité qu'une certaine valeur dépasse le seuil μ d'au plus une quantité y, sachant qu'elle dépasse μ. Cette fonction s'écrit sous la forme :

$$\mathbf{F}_{\mu}(y) = \frac{Pr\{X - \mu \leq y, X > \mu\}}{Pr(X > \mu)} = \frac{\mathbf{F}(y+\mu) - \mathbf{F}(\mu)}{1 - \mathbf{F}(\mu)}. \tag{1.15}$$

Nous allons à présent énoncer le théorème de Pickands-Balkema- de Haan qui va être le résultat théorique central de la théorie des valeurs extrêmes. Le théorème énonce que si $\mathbf{F}$ appartient à l'un des trois domaines d'attraction de la loi limite des extrêmes (Fréchet, Gumbel ou Weibull), alors il existe une fonction de répartition des excès au-delà de μ, note $\mathbf{F}_{\mu}$ qui peut-être approchée par une loi de Pareto généralisée (GPD) telle que :

$$\lim_{\mu \to x_0} \sup_{0 \leq y \leq x_0 - \mu} |\mathbf{F}_{\mu}(y) - \mathbf{G}_{\xi,\beta(\mu)}(y)| = 0. \tag{1.16}$$

Cette considération théorique suggère que, lorsque nous avons des données issues d'une distribution inconnue, il est possible d'approximer la distribution au-delà d'un certain seuil (assez grand) par une distribution de Pareto généralisée. $\mathbf{G}_{\xi,\beta}$ est la fonction de répartition de la distribution de Pareto généralisée (GPD) de paramètres ξ et β définie par :

$$\mathbf{G}_{\xi,\beta}(x) = \begin{cases} 1 - \left(1 + \frac{\xi x}{\beta}\right)^{\frac{-1}{\xi}}, & \text{si } \xi \neq 0\,; \\ 1 - \exp\left(\frac{-x}{\beta}\right), & \text{si } \xi = 0. \end{cases} \tag{1.17}$$

avec $\beta > 0$,$x \geq 0$ pour $\xi \geq 0$ et $0 \leq x \leq \frac{-\beta}{\xi}$ pour $\xi < 0$.

Le paramètre ξ est lié au caractère leptokurtique de la fonction de distribution et β est un paramètre d'échelle.

La valeur prise par le paramètre ξ informe sur le poids des queues dans la distribution parente. En d'autres termes, plus les indices de queue ξ sont élevés plus la distribution considérée possède des queues épaisses [41].

1.3.3 Représentation Quantile-Quantile

Supposons que $X_1, X_2, ..., X_n$ sont des variables aléatoires iid, de fonction de répartition F (et $F^{\leftarrow}$ son inverse généralisé) et introduisons la statistique d'ordre : $X_1 \geq X_2 \geq ... \geq X_n$. Le graphique quantile-quantile est défini par :

$$\left\{ \left(X_k, F^{\leftarrow}\left(\frac{n-k+1}{n+1}\right)\right), \quad k = 1, ...n. \right\}. \tag{1.18}$$

Un graphique QQ-plot est un outil convenable pour voir si la distribution d'une variable dans un échantillon provient d'une distribution théorique spécifique. Le QQ-plot est un graphique qui oppose les quantiles de la distribution empirique aux quantiles de la distribution théorique envisagée. Si l'échantillon provient bien de cette distribution théorique, alors le QQ-plot sera linéaire.

Dans la théorie des valeurs extrêmes, le QQ-plot se base sur la distribution exponentielle. Le QQ-plot sous l'hypothèse d'une distribution exponentielle est la représentation des quantiles de la distribution empirique sur l'axe des X contre les quantiles de la fonction de distribution exponentielle sur l'axe des Y.

L'intérêt de ce graphique est de nous permettre d'obtenir la forme de la queue de la distribution.

Trois cas de figures sont possibles :

1. Les données suivent la loi exponentielle : la distribution présente une *queue très légère*, les points du graphique présentent une forme linéaire.
2. Les données suivent une distribution à queue épaisse *"heavy-tailed distribution"* : le graphique QQ-plot est concave.
3. Les données suivent une distribution à queue légère *"light-tailed distribution"* : le graphique QQ-plot a une forme convexe.

Enfin, les règles des interprétations des graphiques obtenus par les instructions QQ-plot, sont données par le tableau suivant :

Pattern		Interpretation
	All but a few points fall on a line	Outliers in the data
	Left end of the pattern is below the line while the right end of the pattern is above the line	Symmetric, long tails at both ends
	Left end of the pattern is above the line while the right end of the pattern is below the line	Symmetric, short tails at both ends
	Curved pattern with slope increasing from left to right	Skewed to right
	Curved pattern with slope decreasing from left to right	Skewed to left
	Staircase pattern	Data have been rounded or may be discrete

Exemple 1.3. *Soit les résultats graphiques 1.8 :*

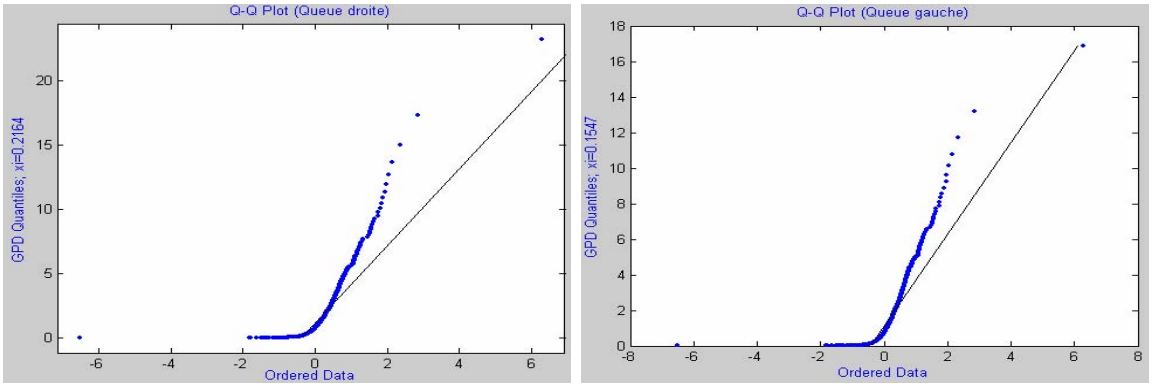

FIGURE 1.8 – Q-Q Plot : A) queue droite B) queue gauche

L'alignement des points du Q-Q plot laisse supposer que la loi GPD déterminée pour un seuil de α, décrit relativement bien le comportement des excès situés dans la queue droite (gauche) de la distribution.

Il est à noter que la concavité des déviations par rapport à la ligne droite dans les deux Q-Q plots précédents révèle une distribution à queue épaisse. En revanche, des déviations convexes témoignent d'une distribution à queue fine (pour plus de détails sur cette exemple voir [41]).

1.3.4 Estimation non paramétrique de l'indice de queue

Plusieurs techniques non paramétriques pour l'estimation de l'indice de queue(l'indice de variabilité ξ), ont été proposées dans la littérature. On cite : l'estimateur de Hill [33], l'estimateur de Pickands [47], l'estimateur de Dekkers, Einmalh et Hann (DEdh) [49].

Estimateur de Hill

L'estimateur de Hill de l'indice de queue de la loi GPD a été étudié par Mason (1982), Goldie et Smith (1987) et Rootzèn (1992). La méthode consiste à ordonner les observations par ordre décroissant $X_1 \geq X_2 \geq ... \geq X_n$, l'indice de queue étant donné par l'équation ci-dessous avec N_μ, le nombre d'observations supérieures au seuil μ :

$$\xi^{Hill} = \frac{1}{N_\mu} \sum_{i=1}^{N_\mu} ln \left(\frac{X_i}{X_{N_\mu+1}} \right). \tag{1.19}$$

Il s'agit de sélectionner graphiquement le nombre d'excès au-delà duquel la valeur de l'indice de queue ξ devient stable. Selon Dress, de Haan et Resnick (1998), cette méthode serait particulièrement bien adaptée aux distributions d'excès convergeant vers une GPD en assurant un bon équilibre entre biais et variance. L'estimateur de Hill n'est valable que pour les distributions de Fréchet [41].

Exemple 1.4. *Le graphique 1.9 présente le résultat de l'estimation de l'indice de queue ξ calculé par la méthode de Hill en fonction du nombre d'excès considéré. Le choix de 50 excès correspondant à un seuil de 0,9932 semble pertinent, dans la mesure où l'indice de queue devient relativement stable au-delà de ce point. D'une manière analogue, pour la queue gauche, l'indice de queue devient relativement stable pour un nombre d'excès égal à 50 (pour plus de détails sur cette exemple voir [41]).*

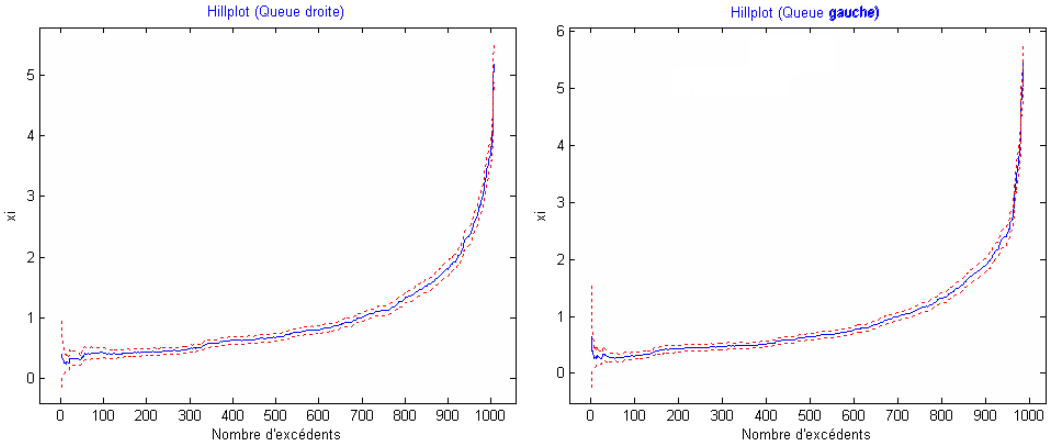

FIGURE 1.9 – Hill plot

L'estimateur de Pickands et l'estimateur de Dekkers, Einmalh et Hann(DEdh)

Supposons que $X_1, X_2..., X_n$ sont des observations indépendantes de fonction de répartition F à support continu sur (x_e, x_f), ces estimateurs sont basés sur la statistique d'ordre : $X_{1,n} \leq X_{2,n} \leq ... \leq X_{n,n}$, ils sont définis par :

$$\xi_{n,N_\mu}^{Pickands} = \frac{\left[\log \frac{X_{(N_\mu,n)} - X_{(2N_\mu,n)}}{X_{(2N_\mu,n)} - X_{(4N_\mu,n)}}\right]}{\log 2},$$

$$\xi_{n,N_\mu}^{DEdh} = \xi_{n,N_\mu}^{H(1)} + 1 - \frac{1}{2}\left[1 - \frac{(\xi_{n,k}^{H(1)})^2}{\xi_{n,N_\mu}^{H(2)}}\right]^{-1}$$

Avec :

$$\xi_{n,N_\mu}^{H(r)} = \frac{1}{N_\mu}\sum_{i=1}^{N_\mu - 1}\left[\log X_{(i,n)} - \log X_{(N_\mu,n)}\right]^r, r = 1, 2, ...$$

Ils ne sont calculés en ne conservant que les N_μ valeurs les plus importantes et leurs valeurs dépendent donc du niveau N_μ retenu. Le choix de N_μ est un problème délicat. Une valeur de N_μ trop grande conduit à prendre en compte des observations qui ne sont plus forcément dans les queues et conduit ainsi à un biais dans l'estimation. Une valeur de N_μ

trop petite signifie que l'on utilise peu d'observations et aboutit à des estimateurs avec des variances importantes. Il y a donc un arbitrage à effectuer. Pour un choix optimal de N_μ, plusieurs méthodes ont été proposées dans la littérature [17, 30, 3, 4, 24, 31].

Propriétés des estimateurs : Pickands, Hill, DEdh

la convergence :
Si $N_\mu \longrightarrow \infty$ et $n \longrightarrow \infty$, de façon que $\frac{N_\mu}{n} \to 0$, alors : les estimateurs

$$\xi_{n,N_\mu}^{Pickands},\ \xi^{Hill} \text{ et} \xi_{n,N_\mu}^{DEdh}$$

convergent.

La normalité asymptotique :

Sous des hypothèses de régularité et de vitesse de divergence de N_μ, les estimateurs, $\xi_{n,N_\mu}^{Pickands}$, ξ_{n,N_μ}^{Hill}, ξ_{n,N_μ}^{DEdh} sont asymptotiquement normaux, tel que :

$$\begin{array}{lll} \sqrt{N_\mu}(\xi_{n,N_\mu}^{Pickands} - \xi) & \curvearrowright & N\left(0, \xi^2 \frac{2^{2\xi+1}+1}{\left(2\left(2^\xi-1\right)\log 2\right)^2}\right), \\ \sqrt{N_\mu}(\xi^{Hill} - \xi) & \curvearrowright & N\left(0, \xi^2\right), \\ \sqrt{N_\mu}(\xi_{n,N_\mu}^{DEdh} - \xi) & \curvearrowright & N\left(0, 1+\xi^2\right), \end{array}$$

pour ξ positif ou nul.

D'un point de vue théorique, toutes ces méthodes partagent les mêmes propriétés de consistance et de normalité asymptotique. Cependant, les études de simulation montrent qu'il y a de grandes différences entre ces estimateurs. En général, il n'y a pas une meilleure méthode dans toutes les situations. La méthode la plus utilisée est celle de Hill. Ceci est dû probablement au fait qu'il est le plus ancien et il fournit un estimateur de l'indice de queue plus efficace que les deux autres estimateurs. En effet, pour des valeurs de ξ positives, l'estimateur de Hill possède une variance plus petite que celle des deux autres estimateurs, ce qui explique qu'il soit le plus souvent utilisé. Les propriétés de consistance de cet estimateur pour la distribution heavy-tailed sont données dans [40]. La normalité asymptotique de cet estimateur a été démontrée par plusieurs auteurs, voir par exemple, de Haan et Resnick [15], Csörgö et Mason[14], Häusher et Teugeles[35], et Beirlant and Teugels[2]. Une bonne explication détaillée sur les propriétés mathématiques de cet estimateur peut être trouvée dans Beirlant[1].

1.3.5 Fonction d'excès en moyenne

On appelle fonction d'excès en moyenne (mean excess function), la fonction $e(\mu)$ définie par :

$$e(\mu) = \mathbf{E}[X - \mu / x > \mu] \quad \text{pour } \mu > 0. \tag{1.20}$$

L'estimateur empirique de la fonction d'excès en moyenne est défini comme étant le rapport entre le nombre total des excès par rapport au seuil μ et le nombre total de points dépassant

le seuil μ. Il est donné par :

$$e_n(\mu) = \frac{\sum_{i=1}^{n}(X_i - \mu)^+}{\sum_{i=1}^{n} 1_{\{X_i > \mu\}}}. \tag{1.21}$$

Ainsi, cette fonction est linéaire en μ. Il s'agit donc de repérer les valeurs de μ à partir desquelles $e(\mu)$ est approximativement linéaire. Graphiquement, cela se traduit par un changement de la pente de la courbe qui ensuite reste stable. Nous observons, pour la queue droite (gauche), que le graphique devient presque linéaire quand le seuil μ est égal à 0,9932 (0,8128)[41].

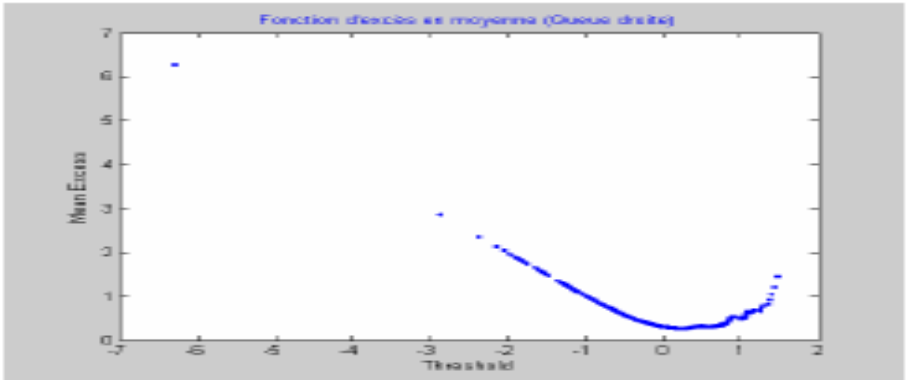

FIGURE 1.10 – Fonction d'excès en moyenne(Queue droite)

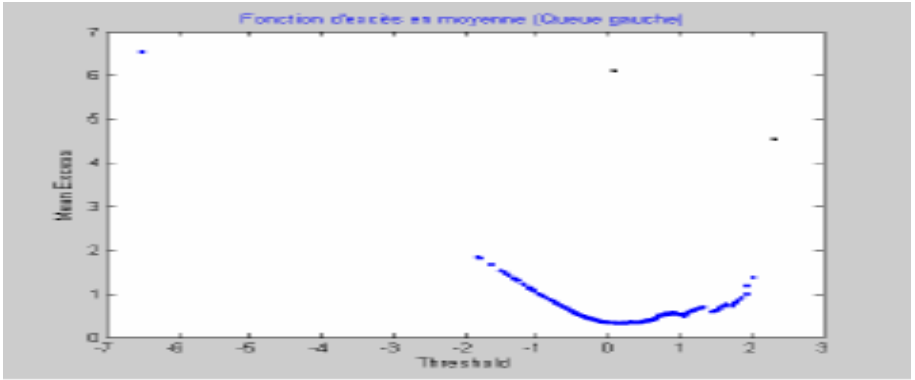

FIGURE 1.11 – Fonction d'excès en moyenne(Queue gauche)

1.4 Tests d'ajustement(Goodness of fit test)

Un test d'adéquation effectué sur la base d'un échantillon permet de déterminer s'il est correct d'approcher la distribution observée, par une loi de probabilité donnée (loi normale, loi de Poisson, etc.). On veut donc savoir si l'échantillon observé peut être issu d'une population qui suit cette loi de probabilité.

Aspects mathématiques : L'objectif d'un test d'adéquation est de déterminer si un échantillon observé peut être considéré comme issu d'une population qui suit une loi de probabilité particulière. Le problème peut être posé sous la forme d'un test d'hypothèse, de la manière suivante :

$$\begin{array}{ll} \text{Hypothèse nulle} & H_0 : "F = F_0", \\ \text{Hypothèse alternative} & H_1 : "F \neq F_0". \end{array}$$

où F est la fonction de répartition inconnue de la population sous-jacente et F_0, la fonction de répartition présumée.

Le test d'adéquation consiste à comparer la distribution empirique avec la distribution présumée. On rejette l'hypothèse nulle, si la distribution empirique n'est pas suffisamment proche de la distribution présumée. Les règles précises de rejet ou d'acceptation de l'hypothèse nulle dépendent du type de test utilisé.

Il existe plusieurs tests d'adéquation dont le test d'adéquation du chi-carré et le test de Kolmogorov- Smirnov [22].

1.4.1 Test du χ^2 (Chi-square goodness of fit test)

Le test d'adéquation du chi-carré est le plus ancien et le plus connu des tests d'adéquation, il a été présenté pour la première fois en 1900 par Karl Pearson.

Soient $X_1, ..., X_n$, un échantillon de n observations. Les étapes d'un test d'adéquation du chi-carré sont les suivantes :

1. Poser les hypothèses. L'hypothèse nulle sera de la forme $H_0 : F = F_0$ où F_0 est la fonction de répartition présumée de la distribution.
2. Répartir les observations en k classes disjointes $[a_{i-1}, a_i]$. On note n_i le nombre d'observations contenues dans la i^{eme} classe, i=1,..., k.
3. Calculer les probabilités théoriques pour chaque classe sur la base de la fonction de répartition présumée F_0 :

$$p_i = F_0(a_i) - F_0(a_{i-1}), i = 1, ..., k$$

4. En déduire les fréquences estimées pour chaque classe

$$e_i = np_i, i = 1, ..., k$$

où n est la taille de l'échantillon.

5. Calculer la statistique χ^2 (chi-carré)

$$\chi^2 = \sum_{i=1}^{n} \frac{(n_i - e_i)^2}{e_i}.$$

Si H_0 est vraie, la statistique χ^2 suit une loi du chi-carré avec v degrés de liberté où $v = (k - 1 - \text{nombre de paramètres estimés})$.

6. Rejeter H_0 si l'écart entre les fréquences observées et les fréquences estimées est grand, c'est-à-dire : si

$$\chi^2 > \chi^2_{v,\alpha}.$$

où $\chi^2_{v,\alpha}$ est la valeur donnée dans la table du χ^2 pour un seuil de signification α particulier.

Domaines et limitations : En vue d'appliquer le test d'adéquation du chi-carré, il faut que n soit assez grand et que les fréquences estimées, e_i, ne soient pas trop petites. On admet habituellement que les fréquences estimées doivent être supérieures à 5, sauf dans les classes extrêmes où elles peuvent être inférieures à 5, mais supérieures à 1. Si cette contrainte n'est pas satisfaite, il faut regrouper les classes pour satisfaire cette règle[22].

1.4.2 Test de Kolmogrov-Smirnov

Les tests d'adéquation permettent de déterminer si la distribution de probabilité de la population dont on connaît un échantillon est d'un certain type. Le test de Kolmogorov-Smirnov est un test d'adéquation pour des variables aléatoires continues. On considère donc n variables aléatoires indépendantes identiquement distribuées dont la fonction de répartition F, inconnue, est continue. Notons F_n la fonction de répartition empirique définie par l'échantillon $(x_1, ..., x_n)$ où les x_i représentent des réalisations de la variable aléatoire en question. La fonction F_n est utilisée comme estimateur de F.
Nous avons :

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} h_i(x), \tag{1.22}$$

où

$$h_i(x) = \begin{cases} \mathbb{1}, & \text{si } x > x_i\,; \\ 0, & \text{sinon.} \end{cases} \tag{1.23}$$

On désire effectuer un test d'hypothèses concernant F. Les hypothèses à tester sont :
H_0 : "$F = F_0$", avec F_0 une fonction de répartition continue spécifiée ;
H_1 : "$F \neq F_0$".
La statistique D_n de Kolmogorov-Smirnov est définie par

$$D_n = sup|F_n(x) - F_0(x)|. \tag{1.24}$$

Sous H_0 pour $n \to \infty$, $D_n \to 0$. Ainsi on rejette H_0 si $D_n > c$ et on l'accepte sinon. La valeur c est déterminée par l'équation :

$$Pr(D_n > c|H_0) = \alpha \tag{1.25}$$

où α est le seuil de signification du test. Pour $n \leq 100$ et quelques valeurs de α les valeurs critiques c sont tabulées[23].

De manière équivalente, on utilse la statistique de test D_n définie comme :

$$D_n = max(D_n^+, D_n^-) \tag{1.26}$$

où

$$D_n^+ = sup(F_n(x) - F_0(x))$$
$$D_n^- = sup(F_0(x) - F_n(x)).$$

1.5 Conclusion

Dans ce chapitre, nous avons mis l'accent sur les lois puissance et sur la théorie des valeurs extrêmes qui se base sur les observations extrêmes qui est un outil fondamental pour étudier des queues de distributions en fonction de l'indice de variabilité ξ.

Nous avons aussi présenté les tests d'ajustement qui nous permettent de déterminer s'il est correct d'approcher la distribution observée, par une loi de probabilité usuelle.

Chapitre 2

Estimation de la densité de probabilité par la méthode du noyau

Contents

2.1 Introduction

Soit $x_1, x_2, \ldots, x_n$ n observations équipondérées issues d'une variable aléatoire réelle X de densité de probabilité réelle $f(x)$ inconnue. Comment obtenir une estimation de $f(x)$ à partir de la seule information contenue dans l'échantillon ?.

Ce problème, que l'on désigne généralement par *estimation non paramétrique de la densité de probabilité* a fait l'objet de multiples travaux par des méthodes diverses, citons :

- L'estimateur par histogramme
- L'estimateur par les séries orthogonales
- Les estimateurs par histogrammes modifiés
- Les méthodes à base de splines
- L'estimateur par la méthode du noyau.

Dans ce chapitre, nous allons présenter une étude détaillée de l'estimateur par la méthode du noyau ainsi que ses propriétés statistiques.

2.2 Estimateur à noyau de Parzen-Rosenblatt

En 1962, Parzen [46] a étudié les propriétés fondamentales de l'estimateur à noyau de la densité, juste après son introduction par Rosenblatt [50]. A partir de ce moment, cet estimateur à noyau de la densité est devenu un objet classique étudié par les statisticiens. Pour les statisticiens, il est déjà devenu un exemple canonique d'estimateur non paramétrique de courbe, qui utilise des résultats de la théorie d'approximation et l'analyse harmonique.

L'estimateur de la densité de probabilité par la méthode du noyau est le plus répandu aujourd'hui, car il répond au problème du choix des différents paramètres dans l'estimation à histogramme et possède de bonnes propriétés. L'idée consiste à évaluer la densité f au point x en comptant le nombre d'observations tombées dans un certain voisinage de x sur $\mathbb{R}$.

Définition 2.1. Soit $x_1, ..., x_n$ un échantillon de loi $f(x)$ sur $\mathbb{R}$, de fonction de répartition $F(x) = \int_{-\infty}^{x} f(t)dt$. On appelle fonction de répartition empirique associé à $x_1, ..., x_n$, la fonction aléatoire $F_n : \mathbb{R} \to [0, 1]$ définie pour tout $x \in \mathbb{R}$ par $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{x_i<x\}}$. On peut également écrire de manière équivalente

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1(x_i)_{]-\infty,x[}. \tag{2.1}$$

La fonction de répartition empirique F_n est un estimateur simple de F. Il s'avère que cette fonction est un très bon estimateur de F.

$$nF_n(x) = \sum_{i=1}^{n} 1_{\{x_i<x\}} \xrightarrow{loi} \mathcal{B}(n, F(x)).$$

où $\mathcal{B}$ est la loi binomiale, dont l'espérance et la variance de $F_n(x)$ sont données respectivement par :

$$\mathbb{E}(F_n(x)) = F(x) \qquad \text{et} \qquad \mathbb{V}(F_n(x) = \frac{1}{n}[1 - F(x)]F(x).$$

A partir de la définition d'une densité de probabilité (basée sur la dérivée de la fonction de répartition) et en utilisant l'équation (2.1), la densité f peut s'écrire en ses points de continuité :

$$f_h(x) = \lim_{h \longrightarrow 0} \frac{F_n(x + h) - F_n(x - h)}{2h}. \tag{2.2}$$

$$\begin{aligned} f_h(x) &= \frac{1}{2nh}\sum_{i=1}^{n} \mathbf{1}(x_i)_{]x-h,x+h[} \\ &= \frac{1}{2nh}\sum_{i=1}^{n} \mathbf{1}_{\{x-h<x_i<x+h\}}. \end{aligned}$$

En posant

$$\omega(u) = \begin{cases} 1/2, & -1 < u \leq 1, \\ 0, & \text{sinon.} \end{cases}$$

On peut réécrire (2.2) sous la forme

$$f_h(x) = \frac{1}{nh} \sum_{i=1}^{n} \omega\big(\frac{x - x_i}{h}\big). \tag{2.3}$$

Nous venons de définir l'estimateur à noyau dit de **Rosenblatt** (uniforme).

Parzen[46] a étudié une classe générale d'estimateurs. En remplaçant la fonction w par une fonction noyau K(Kernel) satisfaisant la condition

$$\int_{-\infty}^{\infty} K(u)du = 1. \tag{2.4}$$

Généralement, K est une densité de probabilité. Par analogie avec la définition de l'estimateur de Rosenblatt l'estimateur à noyau (de Parzen) est :

$$f_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - x_i}{h}), \tag{2.5}$$

où $h = h(n)$ est un paramètre qui est fonction de n, appelé paramètre de lissage.
K est une fonction définie sur $\mathbb{R}$ appelée noyau.

Le noyau K vérifie les conditions suivantes :

$$\int_{\mathbb{R}} K(y)dy = 1, \quad \int_{\mathbb{R}} yK(y)dy = 0, \quad \text{et} \quad \int_{\mathbb{R}} y^2K(y)dy = \sigma_K^2 < \infty.$$

On peut vérifier que $f_h(x)$ est une densité de probabilité. Car $f_h(x) \geq 0 \ \forall x$, et

$$\int_{\mathbb{R}} f_h(x)dx = \int_{\mathbb{R}} \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - x_i}{h})dx = \frac{1}{nh} \sum_{i=1}^{n} \int_{\mathbb{R}} K\big(\frac{x - x_i}{h}\big)dx = \frac{1}{n} \sum_{i=1}^{n} \int_{\mathbb{R}} K(u)du = 1.$$

Pour assurer la convergence de l'estimateur $f_h(x)$, les seules conditions imposées sont :

$$h(n) \longrightarrow 0 \quad \text{et} \quad nh(n) \longrightarrow \infty \quad \text{quand} \quad n \longrightarrow \infty.$$

2.3 Propriétés de l'estimateur à Noyau

Cette section est consacrée à quelques résultats théoriques sur les propriétés de l'estimateur à noyau, à savoir :

– Le comportement asymptotique du biais et de la variance.

– La convergence en moyenne quadratique et en moyenne quadratique intégrée.
– La convergence uniforme (en probabilité,presque complète).
– La convergence en norme L_1.

2.3.1 Espérance, Biais et Variance de l'estimateur

• L'espérance mathématique de $f_h(x)$ est :

$$\begin{aligned}\mathbb{E}f_h(x) &= \frac{1}{nh}\mathbb{E}\sum_{i=1}^{n} K\big(\frac{x-x_i}{h}\big) \\ &= \frac{1}{h}\int_{-\infty}^{\infty} K\big(\frac{x-u}{h}\big) f(u)du.\end{aligned}$$

En posant $y = \frac{x-u}{h} \Rightarrow dy = -\frac{du}{h}$

$$\mathbb{E}f_h(x) = \int_{-\infty}^{\infty} K(y)f(x-hy)dy \tag{2.6}$$

• Le biais de $f_h(x)$ est :

$$Biais\ f_h(x) = \mathbb{E}f_h(x) - f(x) = \int_{-\infty}^{\infty} K(y)f(x-hy)dy - f(x). \tag{2.7}$$

• La variance de $f_h(x)$ est :

$$\begin{aligned}\mathbb{V}f_h(x) &= \mathbb{V}\sum_{i=1}^{n}\frac{1}{nh}K\big(\frac{x-x_i}{h}\big) \\ &= \frac{1}{n^2h^2}\sum_{i=1}^{n}\mathbb{V}K\big(\frac{x-x_i}{h}\big) \\ &= \frac{1}{n^2h^2}\sum_{i=1}^{n}\left[\mathbb{E}\left(K\big(\frac{x-x_i}{h}\big)\right)^2\right] - \frac{1}{n^2h^2}\sum_{i=1}^{n}\left[\mathbb{E}K\big(\frac{x-x_i}{h}\big)\right]^2 \\ &= \frac{1}{nh^2}\int_{-\infty}^{\infty}\big[K\big(\frac{x-y}{h}\big)\big]^2 f(y)dy - \frac{1}{n}\left(\frac{1}{h}\int_{-\infty}^{\infty}K\big(\frac{x-y}{h}\big)f(y)dy\right)^2.\end{aligned}$$

Avec le changement de variable, $y = \frac{x-u}{h}$, on obtient :

$$\mathbb{V}f_h(x) = \frac{1}{nh}\int_{-\infty}^{\infty}\big(K(y)\big)^2 f(x-hy)dy - \frac{1}{n}\left(\int_{-\infty}^{\infty}K(y)f(x-hy)dy\right)^2. \tag{2.8}$$

En faisant le développement de *Taylor* à l'ordre 2 au point $y = 0$ de $f(x-hy)$.
On obtient :

$$f(x-hy) = f(x) - \frac{hy}{1}f'(x) + \frac{h^2y^2}{2!}f''(x) + o(h^2).$$

$$\begin{aligned}\mathbb{E}f_h(x) &= \int_{-\infty}^{\infty} K(y)[f(x) - huf'(x) + \frac{h^2y^2}{2!}f''(x)]dy + o(h^2) \\ &= f(x)\int_{-\infty}^{\infty} K(y)dy - hf'(x)\int_{-\infty}^{\infty} yK(y)dy + \frac{h^2}{2}f''(x)\int_{-\infty}^{\infty} y^2K(y)dy + o(h^2).\end{aligned}$$

Si le noyau K est une fonction symétrique par rapport à 0 c'est-à-dire :

$$\int_{-\infty}^{\infty} yK(y)dy = 0 \quad et \quad \int_{-\infty}^{\infty} y^2K(y)dy < \infty$$

Alors les expressions finales sont données par :

$$\mathbb{E}f_h(x) = f(x) + \frac{h^2}{2}f''(x)\mu_2(K) + o(h^2), \quad \mu_2(K) = \int_{-\infty}^{\infty} y^2K(y)dy. \tag{2.9}$$

$$biais\ f_h(x) = \mathbb{E}f_h(x) - f(x) = \frac{h^2}{2}f''(x)\mu_2(K), \quad \mu_2(K) = \int_{-\infty}^{\infty} y^2K(y)dy. \tag{2.10}$$

$$\mathbb{V}f_h(x) = \frac{f(x)}{nh}\int_{-\infty}^{\infty} K^2(y)dy - \frac{f'(x)}{n}\int_{-\infty}^{\infty} yK^2(y)dy - \frac{1}{n}\Big(f(x) + biais f_h(x)\Big)^2. \tag{2.11}$$

2.3.2 Comportement asymptotique du biais et de la variance

Comportement asymptotique du biais

Théorème. *(Parzen [46])*
Si on a :

1. $\lim_{n\to+\infty} h(n) = 0 \quad et \quad \lim_{y\to+\infty} |yK(y)| = 0$
2. $\sup_y |K(y)| < \infty \quad et \quad \int_{-\infty}^{\infty} |K(y)|dy < \infty$
3. $\int_{-\infty}^{\infty} K(y)dy = 1$

Alors, l'estimateur $f_h(x)$ est asymptotiquement sans biais c'est -à-dire :

$$\lim_{n\longrightarrow\infty} \mathbb{E}f_h(x) = f(x).$$

pour tout point x pour lequel la densité f est continue.

Comportement asymptotique de la variance

Théorème. *(Parzen [46])*
Si on a

1. $\lim_{n\to+\infty} h(n) = 0 \quad et \quad \lim_{y\to+\infty} |yK(y)| = 0.$

2. $\sup_y |K(y)| < \infty \quad et \quad \int_{-\infty}^{\infty} |K(y)|dy < \infty.$
3. $\int_{-\infty}^{\infty} K(y)dy = 1.$

Alors

$$\lim_{n\longrightarrow\infty} nh\mathbb{V}f_h(x) = f(x)\int_{-\infty}^{\infty} K^2(y)dy.$$

pour tout point x pour lequel la densité f est continue.

2.3.3 Convergence de l'estimateur à noyau

Dans ce qui suit, nous allons énoncer quelques résultats qui nous indiquent les différents types de convergence de l'estimateur à noyau.

Convergence en moyenne quadratique

En remplaçant les expressions finales des deux termes, le biais et la variance dans l'équation de $MSE(f(x), f_h(x)) = \mathbb{V}(f_h(x)) + Biais^2 f_h(x)$, on obtient :

$$MSE(f(x), f_h(x)) = \frac{f(x)}{nh}\int_{-\infty}^{\infty} K^2(y)dy + \frac{1}{4}h^4(f''(x))^2\left(\int_{-\infty}^{\infty} y^2K(y)dy\right)^2 + o(\frac{1}{nh} + h^2). \tag{2.12}$$

Théorème. *(Parzen [46])*
Si, $\lim_{n\to\infty} h(n) = 0$ *et* $\lim_{n\to\infty} nh(n) = \infty$,
et K satisfait aux conditions suivantes :

- $\sup_y |K(y)| < \infty \quad et \quad \lim_{y\longrightarrow\infty} |yK(y)| = 0,$
- $\int_{-\infty}^{\infty} |K(y)|dy < \infty \quad et \quad \int_{-\infty}^{\infty} K(y)dy = 1,$

Alors l'estimateur $f_h(x)$ est consistant en moyenne quadratique c'est-à-dire :

$$\lim_{n\longrightarrow\infty} MSE(f_h(x), f(x)) = 0,$$

pour tout point x pour lequel la densité f est continue.

Convergence en moyenne quadratique intégrée

Théorème. *(Parzen [46])*
Si K est un noyau de Parzen-Rosenblatt, c'est-à-dire K verifie :

- $\int_{\mathbb{R}} K(x)dx = 1.$
- $\int_{\mathbb{R}} \mid K(x) \mid dx < \infty.$
- $\sup_x \parallel K(x) \parallel dx < \infty.$
- $\lim_{|x|} K(x) = 0.$

$$\lim_{n\longrightarrow\infty} h(n) = 0, \quad \lim_{n\longrightarrow\infty} nh(n) = \infty \Longleftrightarrow (\forall f \in \mathbb{L}^p), \lim_{n\longrightarrow\infty} MISE(f_h, f) = 0.$$

On note par $\mathbb{L}^p$: l'ensemble des fonctions f définies sur $\mathbb{R}$, telles que $\int |f(x)|^p dx < \infty$.

Convergence uniforme en probabilité

Théorème. *(Parzen [46])*

$$Si \lim_{n\longrightarrow\infty} nh(n)^2 = \infty,$$

si la fonction K satisfait aux conditions suivantes :

1. $\sup_y |K(y)| < \infty \quad et \quad \lim_{y\longrightarrow+\infty} |yK(y)| = 0,$
2. $\int_{-\infty}^{\infty} |K(y)|dy < \infty \quad et \quad \int_{-\infty}^{\infty} K(y)dy = 1,$

et si la transformée de Fourier $\tilde{K}(z) = \int_{-\infty}^{\infty} \exp(-izy)K(y)dy$ est absolument intégrable, alors $f_h(x)$ est un estimateur uniformément consistant en probabilité c'est-à-dire :

$$\forall\epsilon > 0, \quad P\left(\sup_{x\in\mathbb{R}} |f_h(x) - f(x)| < \epsilon\right) = 1.$$

Convergence uniforme presque complète

Théorème. *(Nadaraya [43])*
Si K est un noyau positif à variation bornée et f est uniformément continue,

$$si \lim_{n\longrightarrow\infty} h(n) = 0 \quad et \quad \sum_{n=1}^{\infty} \exp(-\gamma nh(n)^2) < \infty, \quad \forall\gamma > 0,$$

alors :

$$\sup_x |f_h(x) - f(x)| \longrightarrow 0 \quad avec\ une\ probabilité\ 1.$$

Silverman [56] a donné le même théorème sur la convergence presque complète en remplaçant la condition $\sum_{n=1}^{\infty} \exp(-\gamma nh^2) < \infty$, par les deux conditions suivantes :

$$\lim_{n\longrightarrow\infty} h(n) = 0 \quad \text{et} \quad \lim_{n\longrightarrow\infty} \frac{\log n}{nh(n)} = 0.$$

Théorème. *(Silverman[56])*
Si on a :

$$\lim_{n\longrightarrow\infty} h(n) = 0 \quad et \quad \lim_{n\longrightarrow\infty} \frac{\log n}{nh(n)} = 0,$$

et K satisfait aux conditions suivantes :

- *K est uniformement continue et à variation bornée sur $\mathbb{R}$,*
- *Supposons aussi que f est uniformément continue,*
- $\int_{-\infty}^{\infty} |K(y)|dy < \infty, \quad \int_{-\infty}^{\infty} \sqrt{|y\log|y||}\ |dK(y)| < \infty,$
- $\int_{-\infty}^{\infty} K(y)dy = 1,$

alors :

$$\lim_{n\longrightarrow\infty} \sup_x |f_h(x) - f(x)| = 0 \quad Presque\ Sûrement.$$

Convergence L_1 presque complète

Théorème. *(Devroye[19])*
Si :

$$\lim_{n\longrightarrow\infty} h(n) = 0, \quad \lim_{n\longrightarrow\infty} nh(n) = \infty,$$

alors

$$\forall(f \in \mathcal{F}), \lim_{n\longrightarrow\infty} \int |f_h(x) - f(x)|dx = 0, \ \textit{Presque Complètement},$$

où $\mathcal{F}$: Ensemble des densités de probabilité.

Comportement asymptotique

Théorème. *(Parzen[46])*
Si h satisfait : $\lim_{n\longrightarrow\infty} nh(n) = \infty$,
et si le noyau K satisfait aux conditions suivantes :

- $\int_{-\infty}^{\infty} |K(y)|dy < \infty$ *et* $\sup_{y\in\mathbb{R}} |K(y)| < \infty$,
- $\int_{-\infty}^{\infty} K(y)dy = 1$ *et* $\lim_{y\longrightarrow\infty} |yK(y)| = 0$,

alors $f_h(x)$ est un estimateur asymptotiquement normal c'est-à-dire :

$$f_h(x) \xrightarrow{cv.loi} \mathcal{N}(\mathbb{E}f_h(x), \mathbb{V}f_h(x)).$$

2.4 Choix du noyau

L'estimation non paramétrique d'une densité de probabilité par la méthode du noyau nécessite le choix du noyau K. Dans cette section nous allons faire une brève présentation de quelques noyaux usuels.

Noyau Uniforme (Rosenblatt)

Ce noyau a été proposé par Rosenblatt en 1956 [50], l'avantage de ce noyau est la simplicité de sa forme. Il s'écrit sous la forme :

$$K(u) = \begin{cases} \frac{1}{2}, & \text{Si } |u| \leq 1\,; \\ 0, & \text{Sinon.} \end{cases} \tag{2.13}$$

Noyau Box(boite)

$$K(u) = \begin{cases} \frac{1}{2\sqrt{3}}, & \text{si } -\sqrt{3} \leq u \leq \sqrt{3}\,; \\ 0, & \text{Sinon.} \end{cases} \tag{2.14}$$

Noyau Triangulaire

Ce noyau a un avantage par rapport au noyau de Rosenblatt, il est continu partout, ce qui conduit à une estimation de f_h continue. Ce noyau s'écrit sous la forme :

$$K(u) = \begin{cases} (1 - |u|), & \text{Si } -1 \leq u \leq 1\,; \\ \\ 0, & \text{Sinon.} \end{cases} \tag{2.15}$$

Noyau Cosine

$$K(u) = \begin{cases} \frac{\pi}{4}\cos(\frac{\pi u}{2}), & \text{Si} -1 \leq u \leq 1\,; \\ \\ 0, & \text{Sinon.} \end{cases} \tag{2.16}$$

Noyau Gaussien

L'avantage du noyau gaussien est que plus la valeur de h est élevée plus on élargit la fenêtre, ce qui a un effet de lissage globale important ; mais le coût de calcul dans le cas de ce noyau est très élevé du fait de son support infini. Ce noyau s'écrit sous la forme :

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}; \qquad \forall u \in \mathbb{R}. \tag{2.17}$$

Noyau Biweight (Tukey)

Le noyau de *Tukey* ou *biweight* est le plus intéressant car donnant une estimation dérivable partout tout en étant simple à mettre en œuvre. En fait, il s'agit du noyau le plus simple parmi les noyaux de forme polynômial dérivable partout. Ainsi, il assure le lissage locale de la fonction f_h. Ce noyau est d'une forme très proche du noyau gaussien, il est donc préférable. il s'écrit sous la forme :

$$K(u) = \begin{cases} \frac{15}{16}(1 - u^2)^2, & \text{Si } -1 \leq u \leq 1\,; \\ \\ 0 & \text{Sinon.} \end{cases} \tag{2.18}$$

Noyau Triweight

Le noyau triweight s'écrit sous la forme :

$$K(u) = \begin{cases} \frac{35}{32}(1 - u^2)^3, & \text{Si } -1 \leq u \leq 1\,; \\ \\ 0 & \text{Sinon.} \end{cases} \tag{2.19}$$

Noyau Epanechnikov

En 1969, Epanechnikov [25], a donné la forme du noyau K_E défini par :

$$K_E = \begin{cases} \frac{3}{4\sqrt{5}}\left(1 - \frac{x^2}{5}\right), & \text{Si} \quad x \in \left[-\sqrt{5}, \sqrt{5}\right]; \\ 0 & \text{Sinon.} \end{cases} \tag{2.20}$$

qui minimise le MISE asymptotique.

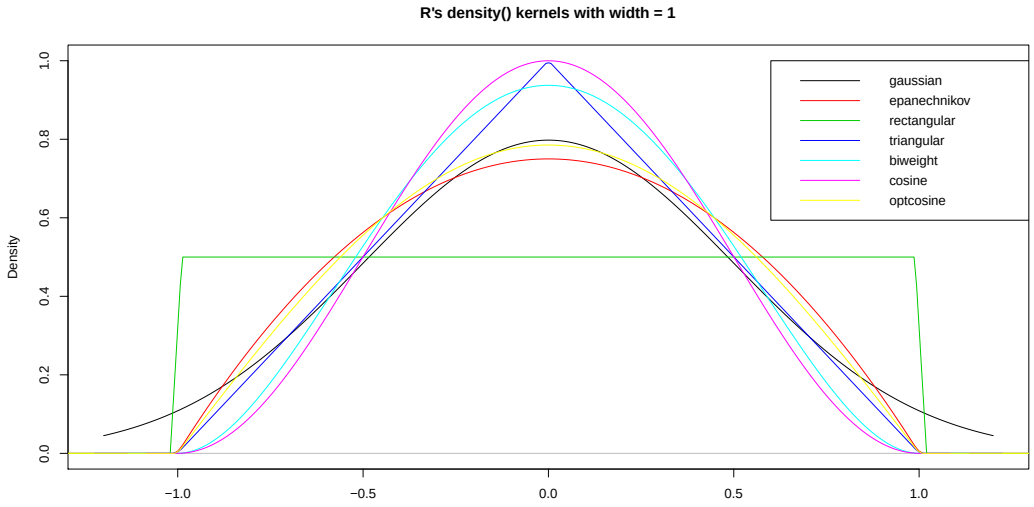

FIGURE 2.1 – Les courbes des différents noyaux usuels

2.5 Choix du paramètre de lissage

D'après la formule (2.5) on constate que l'estimateur $f_h(x)$ de $f(x)$ ne dépend pas seulement du noyau K mais aussi du paramètre h, appelé paramètre de lissage ou fenêtre (bandwidth or window). Une petite perturbation de ce dernier est suffisante pour que $f_h(x)$ change complètement ses caractéristiques (performances numériques ou graphiques), ce qui signifie $f_h(x)$ est fortement lié à ce paramètre. C'est pour cette raison que plusieurs travaux ont été consacrés au choix de ce paramètre.

Il existe plusieurs méthodes de sélection de ce paramètre que l'on peut regrouper en deux familles :

- Méthodes de plug-in (re-injection)
- Méthodes de Cross-Validation (Validation-croisée).
- L'approche bayésienne

La multitude de ces méthodes et leurs diversités du point de vue principe, sont dues au fait que ces méthodes restent incomplètes ou autrement dit, ces méthodes ont toujours des inconvénients [58], soit au sens de la qualité de l'estimateur f_h par rapport à une norme d'erreur bien déterminée, soit par l'allure graphique de la courbe (lissée ou non).

Dans cette section on va présenter la méthode de validation croisée non biaisée.

2.5.1 Méthodes Cross-Validation (Validation Croisée)

Validation croisée non biaisée

Cette méthode appelée Validation Croisée non Biaisée a été proposée par Rudemo [51] en 1982 et Bowman [8] en 1984. Le critère consiste à choisir le paramètre de lissage qui minimise un estimateur convenable de :

$$UCV(h) = \int_{\mathbb{R}} \big[f_h(x) - f(x)\big]^2 dx - \int_{\mathbb{R}} f^2(x)dx = \int_{\mathbb{R}} f_h^2(x)dx - 2\int_{\mathbb{R}} f_h(x)f(x)dx.$$

Puisque $\int_{\mathbb{R}} f^2(x)dx$ ne dépend pas du paramètre de lissage h. On peut choisir le paramètre de lissage de façon à ce qu'il minimise un estimateur de :

$$\int_{\mathbb{R}} f_h^2(x)dx - 2\int_{\mathbb{R}} f_h(x)f(x)dx.$$

Aprés avoir déterminer l'estimateur de $\int_{\mathbb{R}} f_h(x)f(x)dx$ et de le remplacer dans l'équation , le critère $UCV(h)$ sera donnée, pour un noyau K, sous la forme suivante :

$$UCV(h) = \frac{R(K)}{nh} + \sum_{i=1}^{n}\sum_{i\neq j,j=1}^{n}\left[\int \frac{1}{n^2h^2}K\big(\frac{x-x_i}{h}\big)K\big(\frac{x-x_j}{h}\big)dx - \frac{2}{n(n-1)h}K\big(\frac{x_i-x_j}{h}\big)\right]. \tag{2.21}$$

Nous noterons h_{ucv} l'estimateur de h qui minimise $UCV(h)$

De plus, si K est un noyau gaussien alors le critère $UCV(h)$ est donné par la proposition suivante :

Proposition. *Soit $X_1, X_2, ..., X_n$ un n-échantillon i.i.d issu d'une variable aléatoire X de fonction de densité f. Utilisant le noyau gaussien on obtient :*

$$\begin{aligned} UCV(h) &= \frac{1}{2n^2h\sqrt{\pi}}\left(n + 2\sum_{i=1}^{n}\sum_{i\neq j,j=1}^{n}\exp\left(-\big(\frac{x_i-x_j}{2h}\big)^2\right)\right) \\ &- \frac{2}{\sqrt{2}\pi n(n-1)h}\sum_{i=1}^{n}\sum_{i\neq j,j=1}^{n}\exp\left(-\frac{(x_i-x_j)^2}{2h^2}\right). \end{aligned}$$

Remarque 2.1. (Inconvénients de la méthode UCV)

Cette méthode présente deux problèmes majeurs (ou points faibles) : d'une part son manque de robustesse par rapport aux changements de taille de l'échantillon c'est-à-dire le résultat de simulation peut se révéler extrêmement variable d'un échantillon à l'autre, d'autre part, la fonctionnelle à minimiser a souvent tendance à présenter plusieurs minimums locaux [29].

Validation croisée biaisée

Un critère de validation croisée biaisée, a été introduit par Scott et Terrell [55] en 1987 pour remédier aux problèmes de validation croisée non biaisée. Il s'agit d'introduire un biais dans le UCV afin de réduire sa variance.

L'Erreur Quadratique Intégrée Moyenne Asymptotique s'écrit sous la forme :

$$AMISE = \frac{h^4}{4}\sigma_K^4 R(f'') + \frac{R(K)}{nh}.$$

Le paramètre de lissage basé sur la méthode de validation croisée biaisée est la valeur h qui minimise un estimateur du $AMISE$. On peut estimer le $AMISE$ si l'on estime $R(f'')$. Un estimateur naturel de ce terme est donné par $R(f_h'')$ où f_h est l'estimateur de la densité qui utilise la méthode du noyau.

Lemme 2.1. *(Scott et Terrell [55])*
Supposant que le noyau K satisfait aux conditions suivantes :

$$\int K''(u)du = 0, \quad \mu_1(K'') = \int uK''(u) = 0, \quad \mu_2(K'') = \int u^2K''(u) = 2.$$

On obtient le développement asymptotique :

$$\mathbb{E}\big[R(f_h'')\big] = R(f'') + \frac{R(K'')}{nh^5} + O(h^2).$$

Proposition. *(Scott et Trell[55])*
Soit $X_1, X_2, ..., X_n$ un n-échantillon i.i.d issu d'une variable aléatoire X de fonction de densité f. Pour un noyau K on obtient :

$$BCV(h) = \frac{R(K)}{nh} + h^4\frac{\mu_2^2(K)}{4n^2}\sum_i\sum_{j,j\neq i} K_h^{(2)}K_h^{(2)}(X_i - X_j). \tag{2.22}$$

Proposition.
Soit $X_1, X_2, ..., X_n$ un n-échantillon i.i.d issu d'une variable aléatoire X de fonction de densité f. en choisissant le noyau gaussien on obtient :

$$BCV(h) = \frac{1}{2nh\sqrt{\pi}} + \frac{1}{64n^2h\sqrt{\pi}}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\left[\left(\frac{x_i - x_j}{h}\right)^4 - 12\left(\frac{x_i - x_j}{h}\right)^2 + 12\right]\exp\left[-\frac{(x_i - x_j)^2}{4h^2}\right]. \tag{2.23}$$

2.6 Effet du biais aux bornes

Dans la pratique, généralement, on utilise le noyau gaussien en raison de sa simplicité et ses propriétés asymptotiques qui sont établies, pour les variables aléatoires $i.i.d$ ou des séries chronologiques, par plusieurs auteurs (Silverman (1986) [57], Pagan et Ullah (1999)[45] et Fan et Yao (2003) [26].

Cependant le crucial inconvénient de ce noyau est qu'il attribue des poids positif pour des coordonnées (valeurs de x) qui sont à l'extérieur du support lorsque on cherche l'estimateur de la densité de probabilité d'une variable aléatoire bornée ou semi bornée (cas de la loi exponentielle). Cela cause le problème du biais aux bornes et donne un estimateur non consistant.

Ce problème, connu sous le problème de biais aux bornes à donné un essor pour des nouvelles méthodes et de nouveaux noyaux pour l'estimation d'une densité de probabilité par la méthode du noyau pour le cas des données $i.i.d$. Parmi ces méthodes on peut citer :

Les méthodes de réflexion (noyau miroir) de Schuster (1985) [54](voir l'exemple 2.1) et la méthode de rénormalisation local de Diggle (1985) [34] et Härdle(1990)[21]. D'autres auteurs ont proposé d'utiliser les noyaux adaptés aux bornes et les noyaux standard (classique) à l'intérieur du support.

Exemple 2.1. *(Noyau Miroir (Schuster))*

L'idée de cette méthode, développée par Deheuvels et Hominal (1979)[16] et Schuster (1985) [54], est d' ajouter une "masse manquante" par réflexion de l'échantillon et qui concerne les données aux frontières. Elles se focalisent sur le cas où les variables sont positives, c'est-à-dire, dont le support est $[0, \infty[$. *Formellement et sous sa forme plus simple, il consiste à remplacer* $K\left(\frac{x-X_i}{h}\right)$ *par* $K\left(\frac{x-X_i}{h}\right) + K\left(\frac{x+X_i}{h}\right)$. *L'estimateur de la densité est alors de la forme :*

$$f_h = \frac{1}{nh}\sum_{i=1}^{n}\left[K\left(\frac{x-X_i}{h}\right) + K\left(\frac{x+X_i}{h}\right)\right]. \tag{2.24}$$

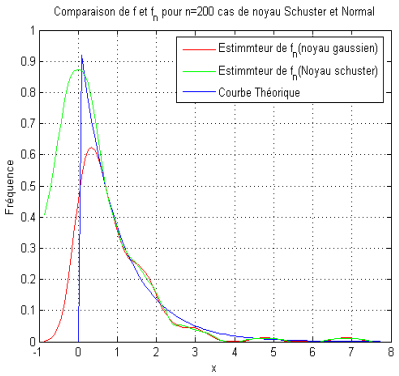

FIGURE 2.2 – Noyau Miroir (Schuster)

Dans le cas des densités dont le support est $[0, 1]$*, la non consistance peut être corrigée aux bornes. Mais le taux de convergence du biais reste très faible, il est d'ordre* $0(h)$ *aux bornes qui est grand par rapport au taux usuel qui est d'ordre* $0(h^2)$*.*

2.7 Noyaux Asymétriques

Quoique les méthodes précédentes diminuent le biais, aux bornes, elles restent peu efficaces car le biais reste considérable si on le compare aux biais de l'intérieur du support. Pour obtenir un biais aux bornes de même ordre que celui de l'intérieur, Devroy et Gyorfi (1985)[20] et Marron et Ruppert (1994)[39], ont proposé d'appliquer une transformation sur les données originales de telle façon que la dérivée d'ordre 1 de la densité des variables transformées soit égale à zéro et ensuite utiliser la méthode de réflexion pour estimer la densité des données transformées. L'objectif étant de trouver cette fois un biais du même ordre mais sans transformation des données. Plusieurs autres auteurs ont proposé d'utiliser les noyaux adaptés dans la région des bornes et le noyau standard à l'intérieur du support (voir Jones (1993) [36]). Pour l'estimation à noyau aux bornes, Müller (1991) [42] pour l'estimateur à noyau optimal aux bornes et Lejeune et Sarda (1992) [37] pour l'estimation linéaire locale.

L'inconvénient de ces estimateurs est qu'ils attribuent des poids négatifs aux valeurs du voisinage des bornes.

La solution la plus récente est d'utiliser des noyaux asymétriques et adaptés qui n'assignent aucun poids à l'extérieur du support. Chen (1999)[12] et Chen (2000)[13] propose respectivement le noyau *Beta* pour les densités à support compact et le noyau gamma pour les densités à variables à support positif (c'est-à-dire sur $[0, +\infty[$).

2.7.1 Noyau Beta

Dans cette section on présente le noyau bêta proposé par Brown et Chen (1999)[9], et Chen (1999,2000)[12, 13] pour l'estimation non paramétrique de la courbe de régression et

des densités unidimensionnelles définis sur un support compact.

L'idée de Harrell et Davis (1982)[32], Chen (1999,2000)[12, 13] est d'utiliser le noyau bêta pour estimer la densité de probabilité à support compact $[0,\ 1]$ et ainsi de régler le problème du biais aux bornes. L'estimateur de la densité sera alors de la forme :

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K(X_i, \frac{x}{h} + 1, \frac{(1-x)}{h} + 1), \tag{2.25}$$

où $K(., \alpha, \beta)$ représente la densité de la distribution *Beta* de paramètres α et β,

$$K(x, \alpha, \beta) = \frac{x^{\alpha}(1-x)^{\beta}}{\mathcal{B}(\alpha, \beta)}, \qquad x \in [0, 1],$$

avec,

$$\mathcal{B}(\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}.$$

Le noyau bêta a deux avantages, premièrement il peut parfaitement estimer les densités à support compact et deuxièmement il possède une forme flexible qui change le lissage dans le sens naturel quand on s'éloigne des bornes. Par conséquent, le noyau bêta élimine le biais des bornes et fourni une réduction de la variance (voir figure (FIG.2.3)). Charpentier, Fermanian et Scaillet [11] ont montré par simulation que l'estimateur à noyau bêta est plus performant quand on le compare à d'autre estimateurs avec des noyaux standards.

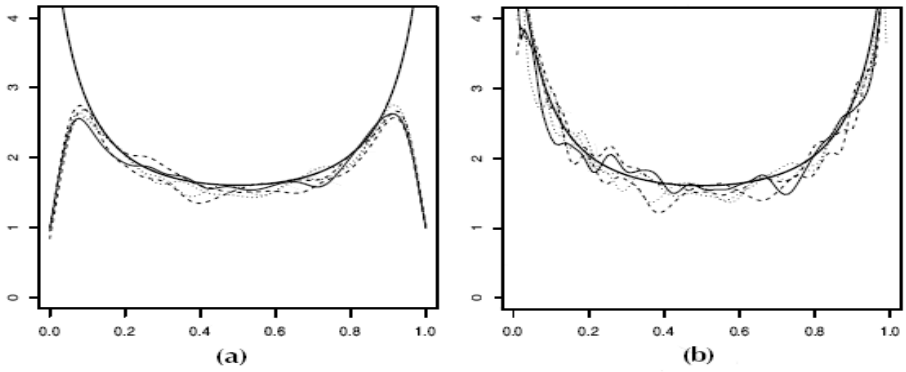

FIGURE 2.3 – Estimation d'une densité de probabilité à support compact $[0, 1]$ pour $n = 10000$:(a) quand on utilise le noyau standard (gaussien) , (b) quand on utilise le noyau bêta.

2.7.2 Noyaux Gamma

On observe $x_1, x_2, ..., x_n$ à partir d'une densité f inconnue. L'objectif est d'estimer la fonction $f(x)$ (par la méthode de noyau) définis un support $x \in [0, +\infty[$. La première forme

du noyau gamma est définie comme suit (voir Bouezmarni et Scaillet [7]) :

$$K_{(\frac{x}{h}+1,h)}(t) = \frac{t^{(x/h)}e^{-(t/h)}}{h^{(x/h)+1}\Gamma((x/h)+1)}. \tag{2.26}$$

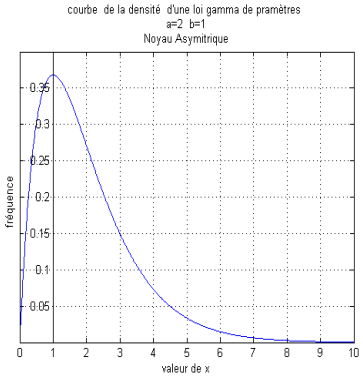

FIGURE 2.4 – Courbe de la densité gamma

L'estimateur à noyau gamma est défini comme suit :

$$\hat{f}_{G_1}(x) = \frac{1}{n}\sum_{i=1}^{n} K_{(\frac{x}{h}+1,h)}(X_i).$$

Biais :

$$Biais\{\hat{f}_{G_1}(x)\} = h\{f'(x) + \frac{1}{2}xf''(x)\} + \circ(h) \tag{2.27}$$

Variance : La variance de $\hat{f}_{G_1}(x)$ est :

$$Var\{\hat{f}_{G_1}(x)\} = n^{-1}Var\{K_{x/h+1,h}(X_i)\} = n^{-1}E\{K_{x/h+1,h}(X_i)\}^2 + O(n^{-1}) \tag{2.28}$$

Soit η_x une variable aléatoire de distribution gamma de paramètres $(2x/h+1, h)$, On aura :

$$E\{K_{x/h+1,h}(X_i)\}^2 = B_h(x)E\{f(\eta_x)\};$$

$$B_h(x) = \begin{cases} \frac{1}{2\sqrt{\Pi}}n^{-1}h^{-1/2}x{-}1/2f(x), & \text{Si } x/h \to \infty\,; \\ \\ \frac{\Gamma(2K+1)}{2^{1+2x}\Gamma^2(K+1)}n^{-1}h^{-1}f(x), & \text{Si } x/h \to K. \end{cases}$$

et $k > 0$;

MSE :

$$MSE\{\hat{f}_{G_1}(x)\} \simeq h^2\{f'(x) + \frac{1}{2}xf''(x)\}^2 + n^{-1}B_h f(x) \tag{2.29}$$

MISE :

$$MISE\{\hat{f}_{G_1}(x)\} = h^2 \int_0^\infty \{f'(x)+\frac{1}{2}xf''(x)\}^2dx+\frac{1}{2\sqrt{\pi}}n^{-1}h^{\frac{-1}{2}} \int_0^\infty x^{\frac{-1}{2}}f(x)dx+\circ(n^{-1}h^{\frac{-1}{2}}+h^2) \tag{2.30}$$

Paramètre de lissage optimal :

$$h^*_{G_1} = \frac{\left[\frac{1}{2\sqrt{\Pi}}\int_0^\infty x^{\frac{-1}{2}}f(x)dx\right]^{\frac{2}{5}}}{4^{\frac{2}{5}}\left[\int_0^\infty\{f'(x)+\frac{1}{2}xf''(x)\}^2dx\right]^{\frac{2}{5}}}n^{\frac{-2}{5}}. \tag{2.31}$$

MISE Optimal :

$$MISE^*\{f^*_{G_1}\} = \frac{5}{4^{\frac{4}{5}}}\left[\frac{1}{2\sqrt{\Pi}}\int_0^\infty x^{\frac{-1}{2}}f(x)dx\right]^{\frac{4}{5}} \quad \left[\int_0^\infty\{f'(x)+\frac{1}{2}xf''(x)\}^2dx\right]^{\frac{1}{5}} n^{\frac{-4}{5}} \tag{2.32}$$

En raison de la contribution indésirable de f' dans le biais de l'estimateur $f^*_1(x)$ (voir la forme du biais donné dans la formule 2.27 et figure 2.5 à gauche), une autre version de $f^*_{G_1}(x)$ notée $f^*_{G_2}(x)$ avait été donnée par Chen.

$$f^*_{G_2}(x) = \frac{1}{n}\sum_{i=1}^{n} K_{(\rho h(x),h)}(X_i) \tag{2.33}$$

avec

$$K_{(\rho_h(x),h)}(t) = \frac{t^{\rho_h(x)-1}e^{-t/h}}{h^{\rho_h(x)}\Gamma(\rho_h(x))}; \tag{2.34}$$

et

$$\rho_h(x) = \begin{cases} x/h, & \text{Si } x \geq 2h\,; \\ \frac{1}{4}(x/h)^2+1, & \text{Si } x \in [0,2h[. \end{cases}$$

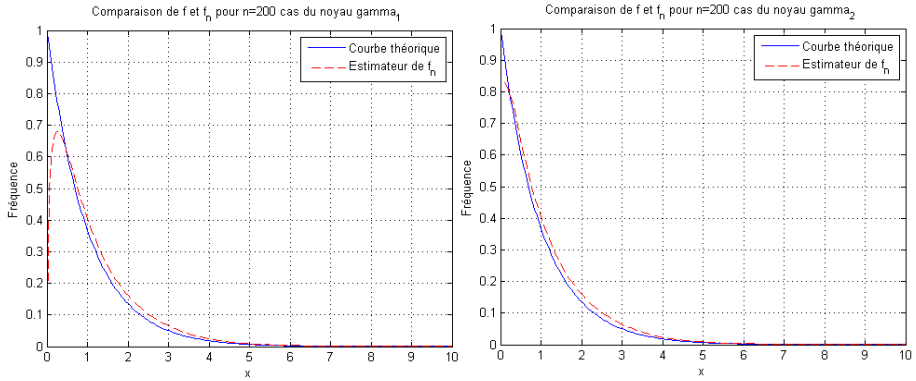

FIGURE 2.5 – Noyaux Gamma

Biais :

$$Biais\{\hat{f}_{G_2}(x)\} = \begin{cases} \frac{1}{2}xf''(x)h + \circ(h), & \text{Si } x \geq 2h\,; \\ \\ \xi_h(x)hf'(x) + \circ(h), & \text{Si } x \in [0, 2h). \end{cases} \tag{2.35}$$

Où $\xi_h(x) = (1-x)\{\rho_h(x) - \frac{x}{h}\}/\{1 + h\rho_h(x) - x\}$, on a $\int_0^\infty \{xf''(x)\}^2 dx < \infty$, $xf''(x)$ converge vers 0 quand $x \longrightarrow \infty$ alors le biais sera minimal en augmentant x.

VAR : La variance de $\hat{f}_{G_2}$ est simulaire à la variance de $\hat{f}_{G_1}$.

MSE :

$$MSE\{\hat{f}_{G_2}(x)\} \simeq \begin{cases} \frac{1}{4}h^2\{xf''(x)\}^2 + n^{-1}B_h(x)f(x), & \text{Si } x \geq 2h\,; \\ \\ h^2\{\xi_h(x)f'(x)\}^2 + n^{-1}B_h(x)f(x), & \text{Si } x \in [0, 2h). \end{cases} \tag{2.36}$$

MISE :

$$MISE\{\hat{f}_{G_2}\} = \frac{1}{4}h^2 \int_0^\infty \{xf''(x)\}^2 dx + \frac{1}{2\sqrt{\pi}} n^{-1} h^{\frac{-1}{2}} \int_0^\infty x^{\frac{-1}{2}} f(x)dx + \circ(n^{-1}h\frac{-1}{2} + h^2) \tag{2.37}$$

Paramètre de lissage optimal :

$$h^*_{G_2} = \frac{\left[\frac{1}{2\sqrt{\pi}} \int_0^\infty x^{\frac{-1}{2}} f(x)dx\right]^{\frac{2}{5}}}{\left[\int_0^\infty \{xf''(x)\}^2 dx\right]^{\frac{2}{5}}} n^{\frac{-2}{5}} \tag{2.38}$$

MISE optimal :

$$MISE^*(\hat{f}_{G_2}) = \frac{5}{4^{4/5}} \left[\frac{1}{2\sqrt{\pi}} \int_0^\infty x^{-1/2} f(x)dx\right]^{4/5} \left[\int_0^\infty \{xf''(x)\}^2 dx\right]^{1/5} n^{-4/5}. \tag{2.39}$$

Remarque 2.2.

1. On constate que la forme du noyau gamma défini dans (2.34) et la qualité du lissage changent selon la position où la densité est estimée (voir figure 2.5 à droite), ce qui implique que l'estimateur du noyau gamma est un estimateur adaptatif de la densité. C'est la différence avec le noyau gaussien, ou tout autre noyau symétrique.
2. Le support du noyau gamma est égal au support de la densité à estimer, ainsi on ne perd pas de poids lorsque on estime la densité aux voisinage des bornes.

2.7.3 Convergence des noyaux gamma

Le noyau gamma est facile à implémenter, il élimine le biais des bornes et souvent non négative. Il atteint le taux de convergence optimale pour les variables $i.i.d$ au sens de MISE dans la classe des estimateurs à noyaux non négatifs. De plus, il permet une réduction de la variance lors du lissage en s'éloignant des bornes. Particulièrement lorsque on utilise la normalité asymptotique de l'estimateur à noyau gamma.

Bouezmarni et Scaillet (2003)[7] ont donné les conditions de convergence faible de l'estimateur à noyau gamma sur un compact $[0, +\infty[$ lorsque f est continue sur ce support et la convergence faible au sens MIAE (Mean Integer Absolute Error). Pour les densités non bornées à l'origine c'est-à-dire au voisinage de zéro, ils ont examiné les performances de cet estimateur par simulation et ils ont prouvé la convergence en probabilité vers l'infini à l'origine.

Fernandez and Monteiro (2005)[27] ont établi le théorème centrale limite pour l'estimateur fonctionnelle pour le noyau gamma. Bouezmarni et Ronbouts (2006)[6] ont démontré la convergence presque sûre au sens du MISE et la normalité asymptotique de cet estimateur.

2.7.4 Noyau inverse gaussien et réciproque de l'inverse gaussien

Soit $X_1, X_2, ..., X_n$ un échantillon aléatoire d'une distribution de densité de probabilité inconnue f, définie sur $[0, \infty[$. On suppose que $f \in C^2$ et $\int_0^\infty \left(x^3 f''(x)\right)^2 dx < \infty$ [53].

Soit $K_{IG(m,\lambda)}$ la densité de de probabilité inverse gaussien de paramètres m et λ, noté $IG(m, \lambda)$, de la variable aléatoire distribuée Y définie comme suit :

$$K_{IG(m,\lambda)}(y) = \frac{\sqrt{\lambda}}{\sqrt{2\pi y^3}} \exp\left(\frac{-\lambda}{2m}\left(\frac{y}{m} - 2 + \frac{m}{y}\right)\right), \quad y > 0 \tag{2.40}$$

L'espérance et la variance de Y valent $E[Y] = m$, $Var[Y] = \frac{m^3}{\lambda}$.

Posons $Z = \frac{1}{Y}$, alors Z suit une loi $RIG(m, \lambda)$ dont sa densité est donnée par :

$$K_{RIG(m,\lambda)}(Z) = \frac{\sqrt{\lambda}}{\sqrt{2\pi Z}} \exp\left(\frac{-\lambda}{2m}\left(mZ - 2 + \frac{1}{mZ}\right)\right), \quad Z > 0. \tag{2.41}$$

La moyenne et la variance de Z valent : $E[Z] = \frac{1}{m} + \frac{1}{\lambda}$, $Var[Z] = \frac{1}{\lambda m} + \frac{2}{\lambda^2}$.

On considère les deux classes des noyaux suivantes :

$$K_{IG(x,\frac{1}{h})}(u) = \frac{1}{\sqrt{2\pi h u^3}} \exp\left(\frac{-1}{2hx}\left(\frac{u}{x} - 2 + \frac{x}{u}\right)\right), \tag{2.42}$$

et

$$K_{RIG(\frac{1}{x-h},\frac{1}{h})}(u) = \frac{1}{\sqrt{2\Pi h u}} \exp\left(-\frac{x-h}{2h}\left(\frac{u}{x-h} - 2 + \frac{x-h}{u}\right)\right) \tag{2.43}$$

Où h doit satisfaire $h + \frac{1}{hn} \to 0$ quand $n \to \infty$.

L'estimateur $\hat{f}_{IG}(x)$ est défini alors comme suit :

$$\hat{f}_{IG}(x) = n^{-1} \sum_{i=1}^{n} K_{IG(x,\frac{1}{h})}(X_i) \tag{2.44}$$

L'estimateur $\hat{f}_{RIG}(x)$ est de la forme suivante :

$$\hat{f}_{RIG}(x) = n^{-1} \sum_{i=1}^{n} K_{RIG(\frac{1}{x-h},\frac{1}{h})}(X_i) \tag{2.45}$$

Il est facile de mettre en application ces estimateurs qui sont semblables aux estimateurs aux noyaux gamma. Il convient à remarquer que $K_{IG(x,\frac{1}{h})}(u)$ tend vers 0, pour tout u pendant que x s'approche de la borne. Ceci induira la contrainte suivante $\hat{f}_{IG}(0) = 0$, ce qui peut être indésirable dans certains cas. $x = 0$, $K_{RIG(\frac{-1}{h},\frac{1}{h})}(u)$ tend vers 0 lorsque u tend vers 0, tandis que $K_{(1,h)}(0) = K_{(\rho_h(0),h)}(0) = \frac{1}{h}$. Pour $x > 0$, tous les noyaux disparaissent si $u = 0$.

Remarque 2.3. Le noyau gamma 2 et RIG sont très semblables, excepté au point $x = 0$, tandis que la différence entre le noyau IG et le noyau gamma 1 est plus remarquée (voir Scaillet [53]).

Biais :

$$Biais\{\hat{f}_{IG}(x)\} = \frac{1}{2}x^3 f''(x)h + \circ(h) \tag{2.46}$$

$$Biais\{\hat{f}_{RIG}(x)\} = \frac{1}{2}x f''(x)h + \circ(h) \tag{2.47}$$

On constate que le biais est grand (respectivement petit) pour l'estimateur à noyau IG pour $x > 1$(respectivement $x < 1$).

Variance En prenant $x \in]0, \infty[$ et K strictement positif.

– Pour $\frac{x}{b} \to \infty$ (x à l'intérieur), $x > 0$, les variances ont les formes suivantes :

$$Var[\hat{f}_{IG}(x)] = \frac{1}{2\sqrt{\pi}} n^{-1} h^{\frac{-1}{2}} x^{\frac{-3}{2}} f(x) + o(n^{-1}h^{\frac{-1}{2}}), \tag{2.48}$$

$$Var[\hat{f}_{RIG}(x)] = \frac{1}{2\sqrt{\pi}} n^{-1} h^{\frac{-1}{2}} x^{\frac{-1}{2}} f(x) + o(n^{-1}h^{\frac{-1}{2}}). \tag{2.49}$$

– Pour $\frac{x}{h} \to K$ (x :borne), $x > 0$, les variances sont égales à :

$$Var[\hat{f}_{IG}(x)] = \frac{1}{2\sqrt{\pi}} n^{-1} h^{-2} K^{\frac{-3}{2}} f(x) + o(n^{-1}h^{-2}), \tag{2.50}$$

$$Var[\hat{f}_{RIG}(x)] = \frac{1}{2\sqrt{\pi}} n^{-1} h^{-1} \left(K^{\frac{-1}{2}} + \frac{7}{16} K^{\frac{-3}{2}} \right) f(x) + o(n^{-1}h^{-1}). \tag{2.51}$$

L'expression de la variance obtenue en utilisant le noyau RIG est égal à l'approximation obtenue par les deux estimateurs des noyaux gamma 1 et 2 sous $\frac{x}{b} \to \infty$. En bornes de x, la variance de l'estimateur RIG a le même ordre que les estimateurs gamma.

Paramètres de lissage optimaux et leurs MISE associés :

$$h^*_{IG} = \frac{\left(\frac{1}{2\sqrt{\pi}} \int_0^\infty x^{\frac{-3}{2}} f(x) dx \right)^{\frac{2}{5}}}{\left(\int_0^\infty \left(x^3 f''(x) \right)^2 dx \right)^{\frac{2}{5}}} n^{\frac{-2}{5}}. \tag{2.52}$$

$$h^*_{RIG} = \frac{\left(\frac{1}{2\sqrt{\pi}} \int_0^\infty x^{\frac{-1}{2}} f(x) dx \right)^{\frac{2}{5}}}{\left(\int_0^\infty \left(x f''(x) \right)^2 dx \right)^{\frac{2}{5}}} n^{\frac{-2}{5}}. \tag{2.53}$$

$$MISE^*_{IG} = \frac{5}{4} \left(\frac{1}{2\sqrt{\pi}} \int_0^\infty x^{\frac{-3}{2}} f(x) dx \right)^{\frac{4}{5}} \left(\int_0^\infty (x^3 f''(x))^2 dx \right)^{\frac{1}{5}} n^{\frac{-4}{5}}. \tag{2.54}$$

$$MISE^*_{RIG} = \frac{5}{4} \left(\frac{1}{2\sqrt{\pi}} \int_0^\infty x^{\frac{-1}{2}} f(x) dx \right)^{\frac{4}{5}} \left(\int_0^\infty (x f''(x))^2 dx \right)^{\frac{1}{5}} n^{\frac{-4}{5}}. \tag{2.55}$$

2.8 Conclusion

Dans ce chapitre nous avons présenté l'estimateur de la densité ainsi que ses principales propriétés par la méthode de noyau. Nous nous sommes intéréssées aux inconvénients et avantages du choix du paramètre de lissage h et du noyau K. Ce qui justifiera le choix de ces deux caractéristiques dans l'application pratique.

Chapitre 3

Application au trafic web

Contents

3.1 Introduction

Nous avons présenté auparavant les lois puissance et heavy-tailed, ainsi que les outils nécessaires pour mesurer leurs indice de variabilité (indice de queue). D'autre part, dans la littérature plusieurs auteurs ont démontré que la distribution d'un trafic web est une loi puissance [10]. A cet effet, comme exemple d'application de ces lois nous avons proposé dans ce chapitre, d'étudier un échantillon de trafic web du serveur de la coupe du monde France 98. Avant de présenter l'application, nous présentons d'abord quelques notions sur le trafic web qui sont utiles pour la compréhension de l'application.

3.2 Quelques notions sur le Web

Le Web est un système hypertexte public fonctionnant sur Internet qui permet de consulter, avec un navigateur, des pages accessibles sur des sites. Une ressource du web est une entité informatique (texte, image, forum Usenet, boîte aux lettres électronique, etc.).

On ne peut accéder à une ressource qu'en respectant un protocole de communication. Les fonctionnalités de chaque protocole varient : réception, envoi, voire échange continu

d'informations.

3.2.1 HTTP

Hyper **T**ext **T**ransfer **P**rotocol est le protocole de communication communément utilisé pour transférer les ressources du Web.

Le HTTP est un protocole simple, basé sur l'émission d'une requête vers un serveur HTTP, en vu de l'obtention d'une ressource. Cette ressource est identifiée par un URL.

L'objectif initial du HTTP est la requête sur des pages hypertexte (HTML), lesquelles servent de base à la construction d'un document composite, par combinaison de divers types de ressources annexes :

- des images
- des feuilles de style
- des scripts clients attachés

L'un des plus importants avantages du protocole http est de "découper" les documents à transférer en blocs. Ces blocs peuvent être acheminés de manière indépendante du serveur vers le client. Ils seront ensuite assemblés au niveau du récepteur.

3.2.2 La transaction HTTP

Une transaction HTTP consiste en un ensemble de requêtes envoyées du client vers le serveur Web suivies des réponses du serveur au client (voir la figure 3.1).

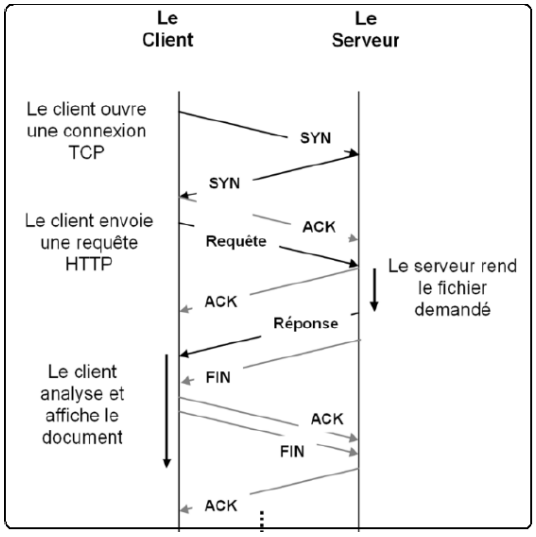

FIGURE 3.1 – La transaction HTTP.

Le client ouvre d'abord une connexion TCP qui résulte de l'échange de paquets SYN comme part des étapes du protocole TCP. Lorsque le client et le serveur terminent les étapes de connexion TCP et se synchronisent, le client envoie une requête HTTP, et le

serveur doit à son tour en envoyant le fichier demandé. Après l'envoi de la réponse, le serveur mis fin à la connexion TCP [52].

3.2.3 URL

Une URL (Uniform Resource Locator) est un format de nommage universel pour désigner une ressource dans le Web. Il s'agit d'une chaîne de caractères qui se décompose en parties :

- **Le nom du protocole :** est, en quelque sorte, le langage utilisé pour communiquer sur le réseau. Le protocole utilisé est le protocole HTTP, suivi du séparateur "//".
- **Le nom du serveur :** il s'agit d'un nom de domaine de l'ordinateur hébergeant la ressource demandée.
- **Le numéro de port :** il s'agit d'un numéro associé à un service permettant au serveur de savoir quel type de ressource est demandée. Le port associé par défaut au protocole HTTP est le port numéro 80.
- **Le chemin d'accès à la ressource :** cette dernière partie permet au serveur de connaître l'emplacement dans lequel la ressource est située, c'est-à-dire de manière générale l'emplacement (répertoire) et le nom du fichier demandé [5].

Une URL a donc la structure suivante :

Protocole	Nom du serveur	Port(par défaut 80)	Chemin
http ://	www.webmaster.com	80	/glossair/glossair.php3

3.2.4 Déroulement d'une requête

Le déroulement d'une requête se fait principalement en sept (07) étapes :

1. Demande d'une connexion ;
2. Attente de la réponse du serveur ;
3. Etablissement de la connexion ;
4. Envoi d'une requête URL ;
5. Réponse du serveur ;
6. Affichage de la réponse ;
7. Fermeture de la connexion.

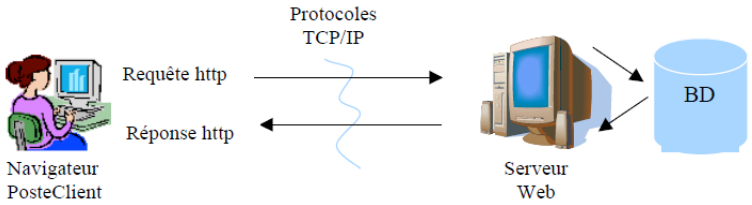

FIGURE 3.2 – Déroulement d'une requête.

3.2.5 Gestion des cas d'erreur

Comme tout protocole se doit être complet, HTTP doit permettre de renseigner le client sur les multiples cas d'erreur qui peuvent se rencontrer au moment de la résolution de l'adresse "logique". Les codes d'erreur du HTTP sont le résultat du catalogage des erreurs possibles (ressource manquante, URL non conforme, etc...). Il utilise un port machine (par défaut le port 80) et cette norme envoie des codes à trois chiffres en tant que réponse [5] : 100, 200,... La signification de chaque code est donnée dans le tableau 3.1.

Code-statut	Phrase-raison	Description
100	Continue	Utilisé pour informer le client que la requête a été reçue par le serveur et n'a pas encore été rejetée, il faut donc continuer l'envoi s'il reste encore les requêtes à envoyer, ou bien ignorer cette réponse dans le cas contraire.
200	OK	Indique la réussite de la requête (l'information retournée avec la réponse dépend de la méthode utilisée).
204	No Content	Pas d'informations à envoyer au client.
304	Bad request	Indique que le serveur n'a pas compris la requête (erreur syntaxique par exemple).
400	Not Modified	Utilisé pour répondre à une requête avec un GET conditionnel (If-Modified-Since)dans le cas de non modification.
404	Not found	Aucune information trouvée.
408	Request Time-out	Indique que le client ne peux pas envoyer une requête après le temps d'attente(fixé à l'avance) de serveur
500	Internal Sever Error	Le serveur a rencontré une condition imprévu qui l'empêche d'accomplir la requête
505	HTTP Version not supported	Version HTTP refusée

TABLE 3.1: Quelques code-statuts et Phrase-raisons associées

3.3 Présentation des données

Les données que nous avons à notre disposition sont une partie d'un fichier trace réel du serveur de la coupe du monde France 98. Chaque fichier trace est séparé dans des fichiers "log" et chaque ligne d'un fichier "log" correspond à une requête "URL" envoyée par un utilisateur, qui contient le nom de la machine, l'instant de la création de la requête, l'URL, la taille des fichiers Web (voir le tableau 3.2).

Nom de la machine	L'instant de création de la requête	URL	Code-Statuts	La taille des fichiers Web
34600	[30/Apr/1998 :21 :30 :17 +0000]	$GET/images$ $/hm\_bg.jpg\ HTTP/1.0$	200	24736

TABLE 3.2: Exemple d'une ligne dans un fichier trace.

3.4 Analyse statistique préliminaire des données

3.4.1 Séparation des données selon la Réussite et l'Échec de téléchargement

Lorsque on a feuilleté les données brutes du trafic web du serveur en question, nous avons constaté que certaines requêtes n'ont pas reçu de réponses pour diverses raisons (voir le tableau (3.1)). L'analyse de ces données au sens de réussite et d'échec nous a fourni les informations résumées dans le tableau suivant :

Mois	Taux de réussite du téléchargement (%)	Taux d'échec du téléchargement (%)					
	200	206	302	304	400	404	500
Avril	81.8462	0.6154	0	17.5385	0	0	0
Mai	83.2079	0.2991	0.0058	16.3499	0.0004	0.1316	0.0052
Juin	85.3917	0.2093	0.0061	13.7529	0.0072	0.6297	0.0032
Juillet	82.5792	0.2565	0.0058	16.7574	0.0003	0.3967	0.0040
04 mois	84.3773	0.2439	0.0060	14.9240	0.0042	0.4407	0.0040

TABLE 3.3: Séparation des données selon la Réussite et l'Echec du téléchargement

A partir des résultats du tableau 3.3, on constate que, parmi les requêtes du téléchargement 85% des requêtes ont été téléchargées avec succès, tandis que les 15% des téléchargements restant ont échoué pour diverses raisons. 86% des échecs sont engendrés par la non compréhension de la requête par le serveur (une "mauvaise requête " ayant le code- statut 304).

Remarque 3.1. Vu que les informations sur les requêtes dans l'état d'échec sont incomplètes, nous prenons en compte que les cas de réussite du téléchargement pour le reste de notre travail.

3.4.2 Analyse statistique descriptive des données

Après l'extraction des données concernant les requêtes qui sont téléchargées avec succès, leurs analyse statistique descriptive nous a fourni les résultats rangés dans le tableau suivant :

Mois	Taille d'échantillon n	Min	Premier Quantile Q_1	Médiane Q_2	Moyenne	Troisième QuantileQ_3	Max	Écart-type σ
Avril	266	42	871	1795	10731	24736	33665	11511
Mai	2791491	4	499	1050	7348.3	3401	2891900	73534
Juin	4896571	4	582	1050	5789	3477	2891900	53659
Juillet	869456	4	749	1420	6356.8	4644	2891900	52753
04 mois	8557784	4	568	1056	6355.5	3676	2891900	60783

TABLE 3.4: Calcul des caractéristiques des données.

On remarque que :
- La taille des échantillons diffère d'un mois à l' autre.
- Les variances des échantillons nous indiquent une forte dispersion des observations autour de la moyenne car les variances sont très grandes. Il est probable que les distributions des données soient des lois puissance, vu que ces dernières ont une variance égale à l'infini (ce qui est confirmé avec la théorie).
- Le trafic est plus important au mois de juin ce qui coïncide avec la période de la coupe du monde France 98.

3.4.3 Hiérarchique du trafic web et leurs classification

Après une analyse des données, nous avons constaté que les fichiers téléchargés sont de différents types : vidéo, image, texte. Les types de fichiers n'ont pas la même taille moyenne (en général, taille vidéo > taille image > taille texte). A cet effet, pour analyser notre échantillon nous proposons de le décomposer en sous échantillons, selon le type, comme il est illustré dans le schéma 3.3.

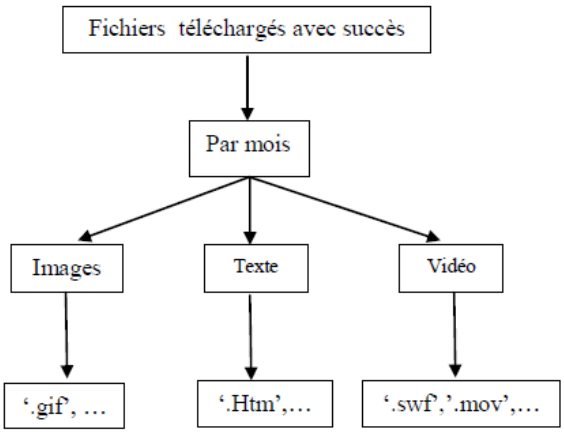

FIGURE 3.3 – Séparation des données selon le type de fichiers existants

Lorsque le serveur Web gère plusieurs catégories de fichiers (vidéo, audio, image, page HTML, fichiers PDF, . . .), il ne peut pas accorder la même charge pour répondre à chacune de ces catégories, vu la variation existante entre la taille de ces fichiers. On note donc à ce niveau qu'il est nécessaire d'adopter une classification du trafic selon la taille des fichiers.

Dans notre cas, nous ne disposons pas de données concernant la charge des types de fichiers sur le serveur web. Par contre, nos données permettent de réaliser la classification des fichiers selon leurs fréquences. Ainsi les résultats de classification obtenus par la méthode ABC (Pareto 80/20) sont présentés dans la figure 3.4.

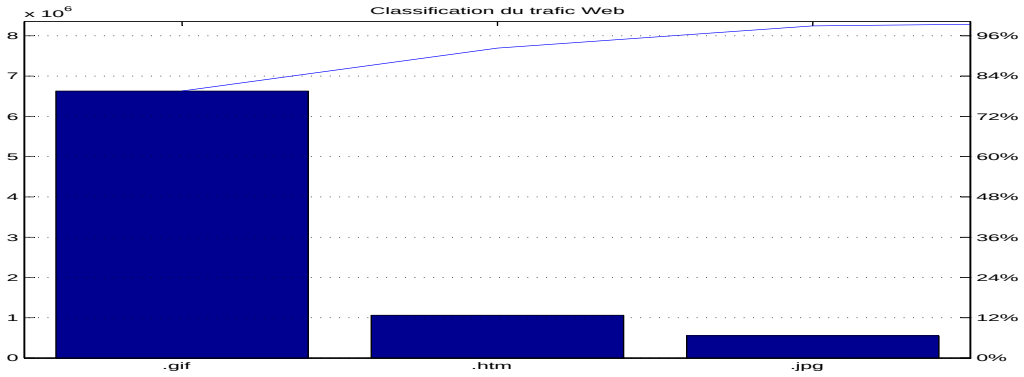

FIGURE 3.4 – Classifications de Pareto de tous les fichiers existants

Pour avoir une idée sur le reste des données(3%) voir la figure 3.5.

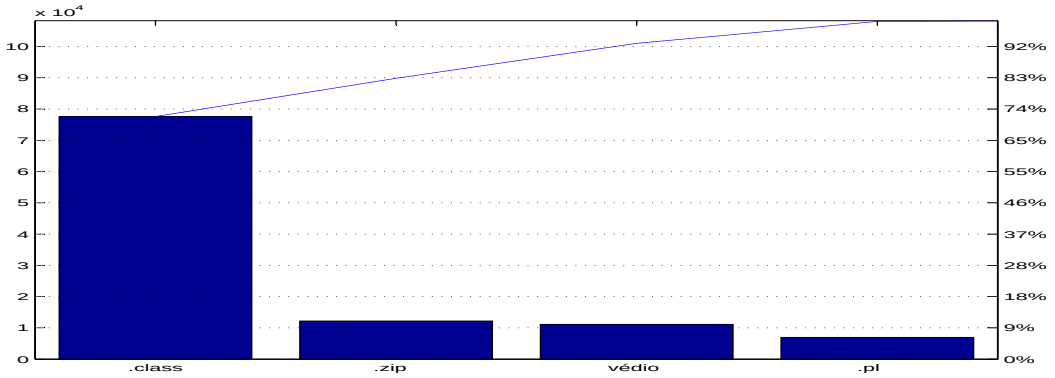

FIGURE 3.5 – Classification de Pareto de reste des fichiers

Discussion
On remarque que les fichiers du type image et htm (".gif", ".jpg" et ".htm") dominent notre échantillon (97%) ce qui conforme aux études existantes.
Dans la littérature, plusieurs travaux ont montré que les requêtes des documents HTML et Image forment 95% de l'ensemble de toutes les requêtes[38]. Plus précisément, les documents Image sont les plus demandés dans le web avec un pourcentage de référencement 65-80% et les documents HTML de 17-28% [38].

Pour déterminer la loi de nos données, il est suffisant de connaître les lois de ces fichiers dominants. A cet effet, pour ce qui suit nous avons opté pour la démarche suivante :

1. estimer les caractéristiques considérées pour chaque type pour chaque mois.
2. estimer les caractéristiques considérées pour chaque type pour quatre (04) mois.
3. estimer les caractéristiques considérées pour l'échantillon global.

3.5 Estimation de l'indice de variabilité

Dans ce qui suit, nous allons présenter les résultats obtenus par les méthodes distribution de survie, QQ-plot, estimateur de Hill et la méthode du maximum de vraisemblance pour l'estimation de l'indice de queue des différents échantillons.

3.5.1 Par la distribution de survie

Afin de déterminer les classes (Pareto, heavy- tailed,...) des distributions de différents fichiers (.gif ,.htm, .jpg,...), nous avons jugé utile d'estimer d'abord leurs fonctions de survie ($\overline{F_n}$) qui nous donnent une idée préalable sur le comportement de la loi des tailles de chaque type de fichier (section 3.4.3). Pour cela, on compare ces dernières avec celles d'une fonction exponentielle $\overline{F_0}$. Les résultats graphiques obtenus pour l'estimation empirique de la distribution de survie des différents fichiers sont présentés dans les figures suivantes :

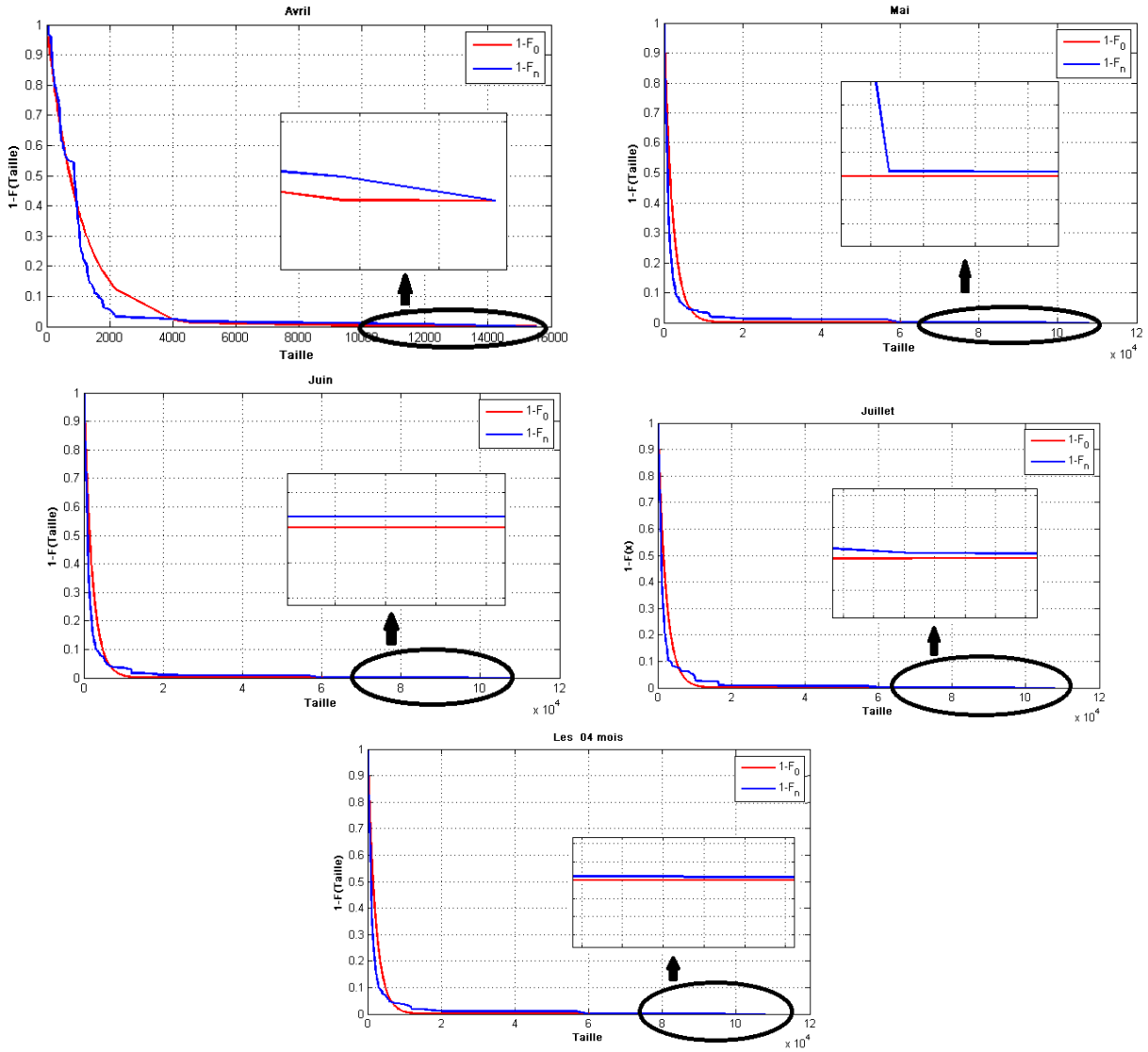

FIGURE 3.6 – Distribution de survie des fichiers du type ".gif"

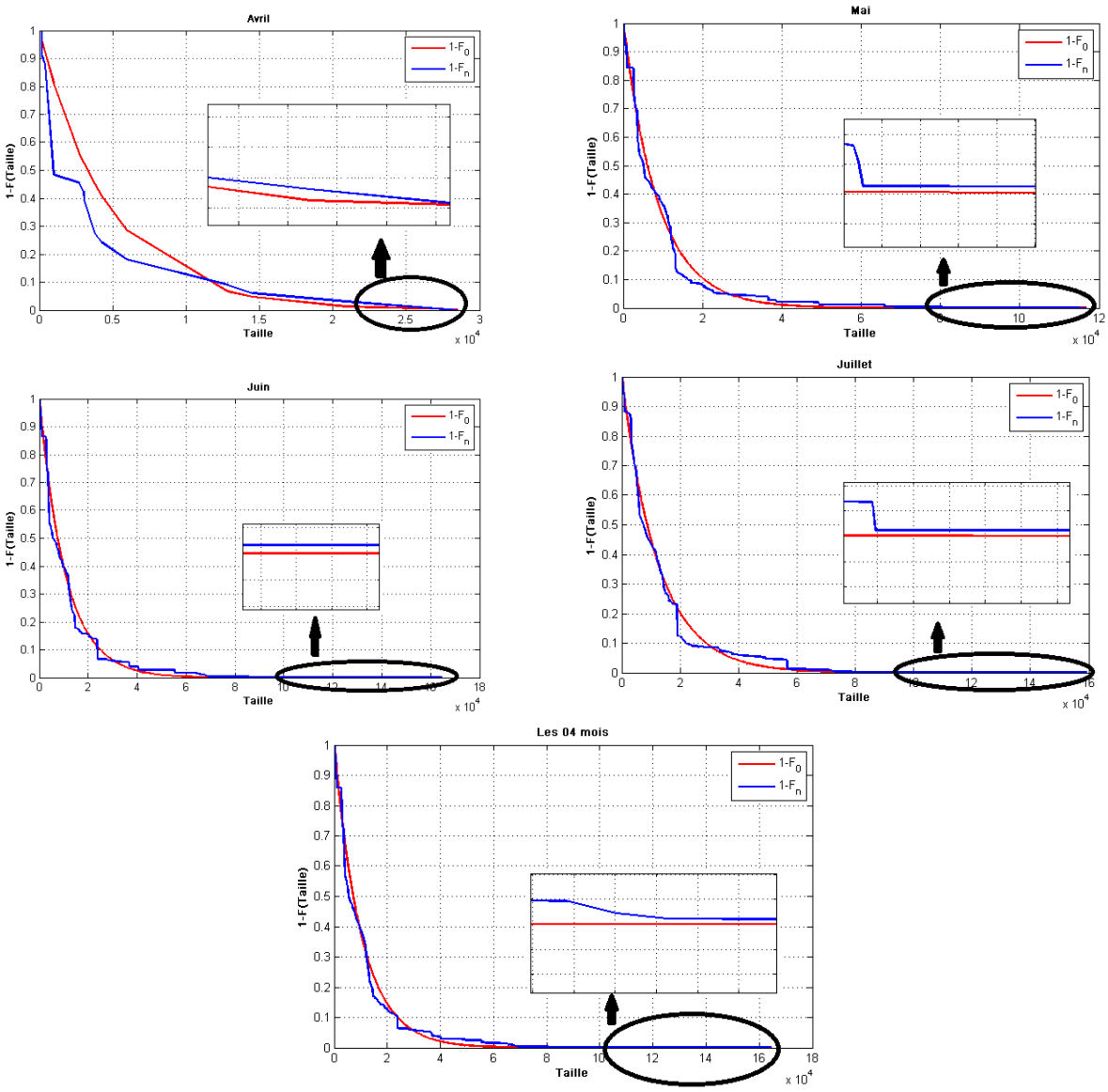

FIGURE 3.7 – Distribution de survie des fichiers du type ".htm"

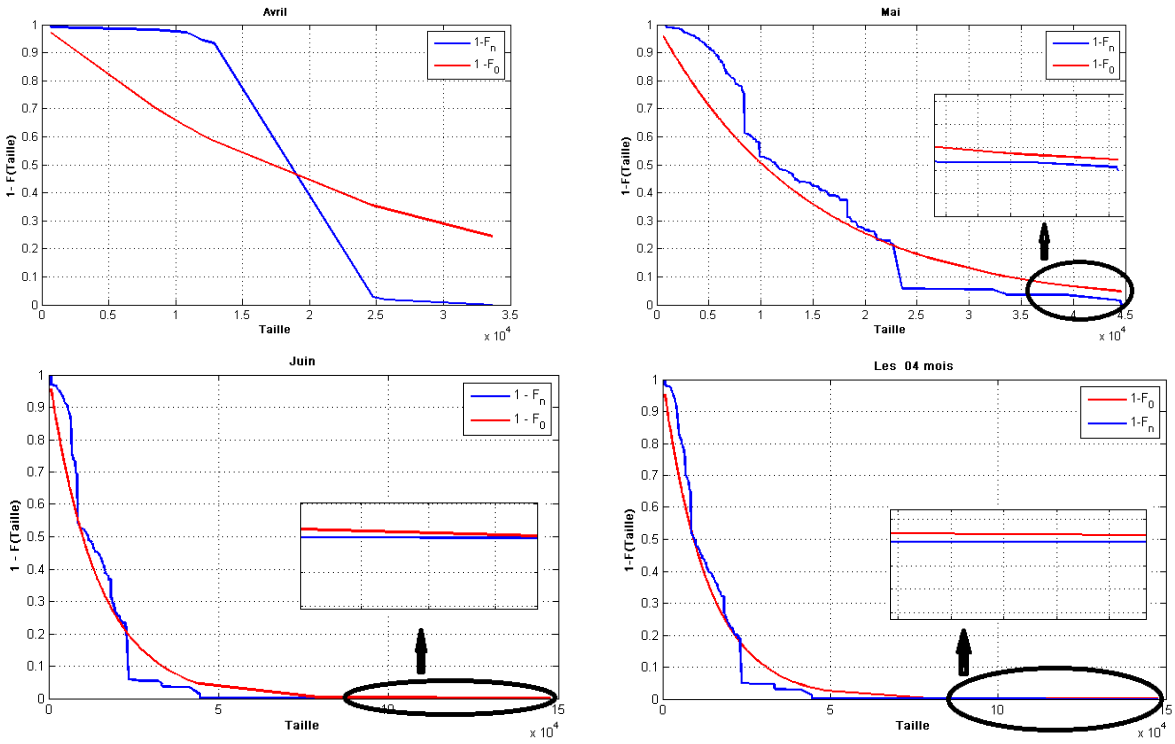

FIGURE 3.8 – Distribution de survie des fichiers du type ".jpg"

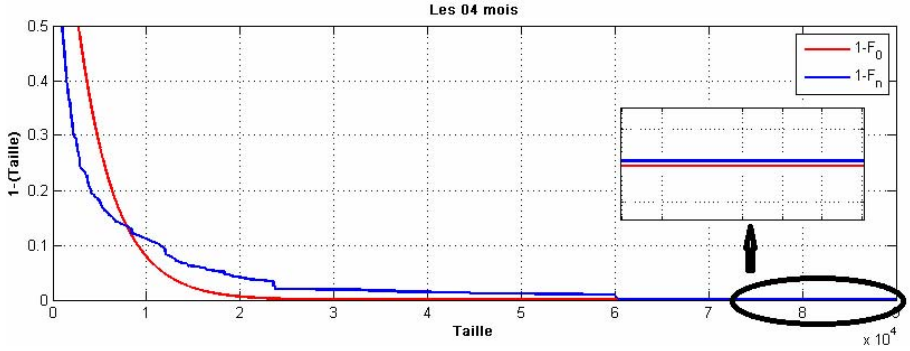

FIGURE 3.9 – Distribution de survie de tous les fichiers.

Discussion

A partir de ces figures, nous constatons que : la forme des distributions des sous échantillons (les types des fichiers) ne sont pas les mêmes.
Les graphiques montrent que les distributions peuvent être classées en trois grandes classes : classe de lois à queue lourde, classe de lois à queue fine et classe de lois sans queue.

Les résultats graphiques montrent que la forme de la distribution de nos données dans tous les cas appartient à la classe des lois à queue lourde, sauf pour les fichiers du type "jpg".

Les détails de cette constatation sont résumés dans le tableau suivant :

Extension	Avril	Mai	Juin	juillet	04 mois
GIF	queue lourde	queue lourde	queue lourde	queue lourde	queue lourde
JPG	sans queue	queue fine	queue fine	queue fine	queue fine
HTM	queue lourde	queue lourde	queue lourde	queue lourde	queue lourde
Global	queue lourde	queue lourde	queue lourde	queue lourde	queue lourde

TABLE 3.5: Classification des distributions de différents fichiers selon la nature de la queue.

3.5.2 Méthode de QQ-plot

Dans cette partie, nous allons utiliser la méthode QQ-plot, pour déterminer le type des distributions correspondantes à nos données. La représentation des quantiles de la distribution empirique de nos différentes données contre les quantiles de la fonction de distribution normale nous a fourni les résultats suivants :

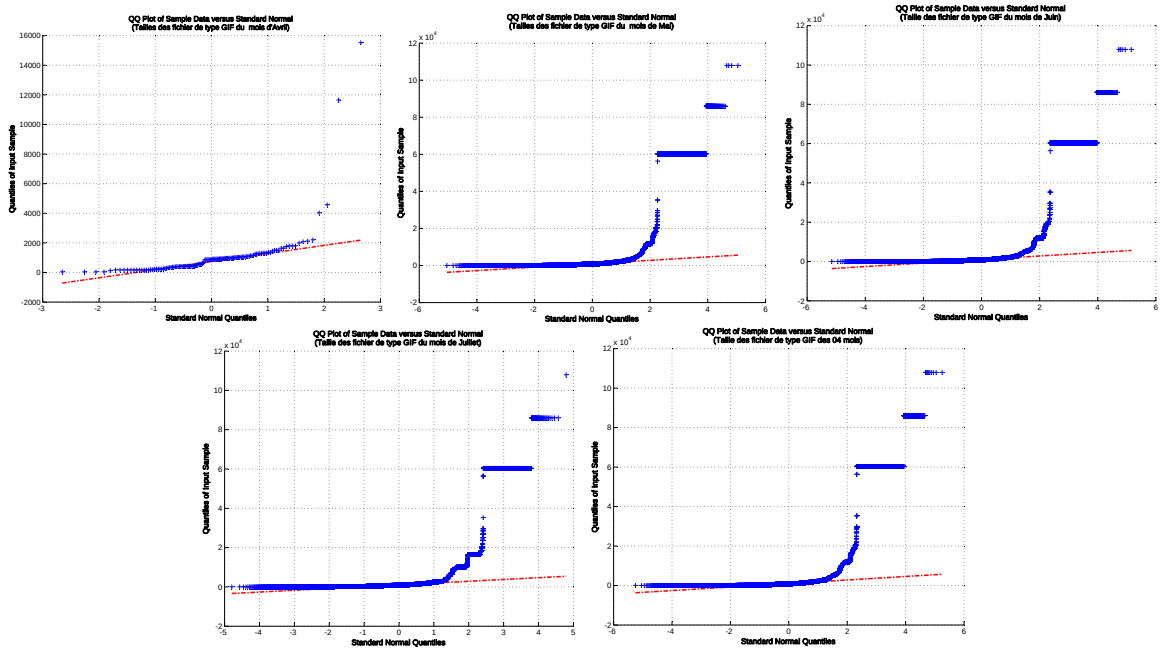

FIGURE 3.10 – QQ-Plot de la taille des fichiers du type '.gif' en fonction de la loi normale.

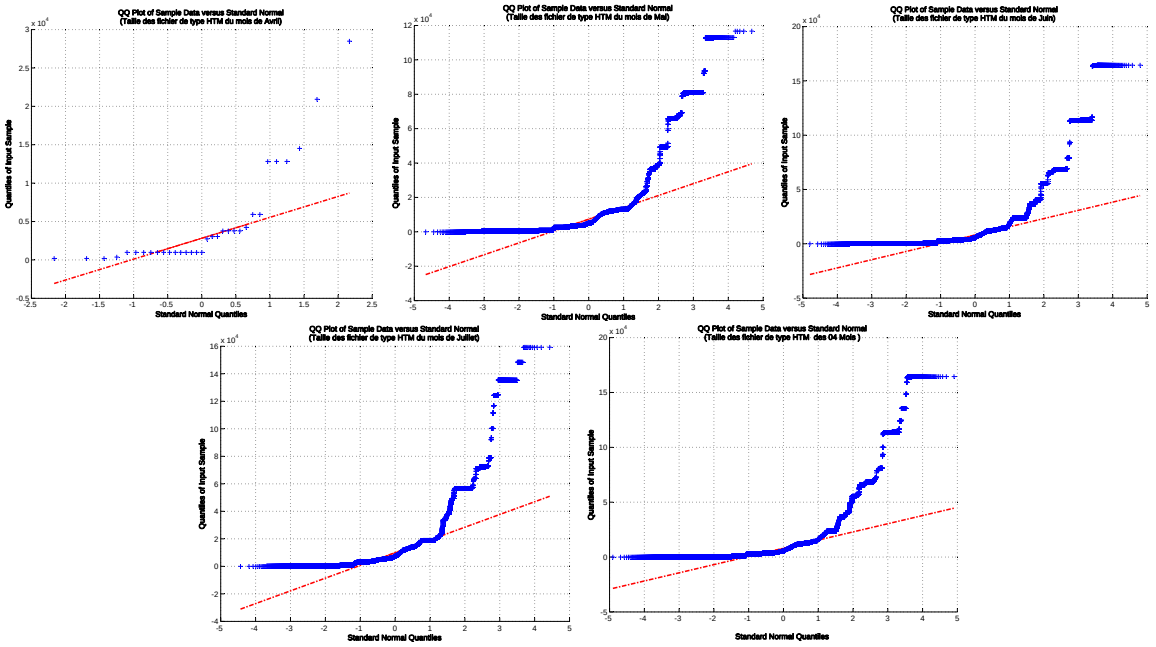

FIGURE 3.11 – QQ-Plot de la taille des fichiers du type '.HTM' en fonction de la loi normale.

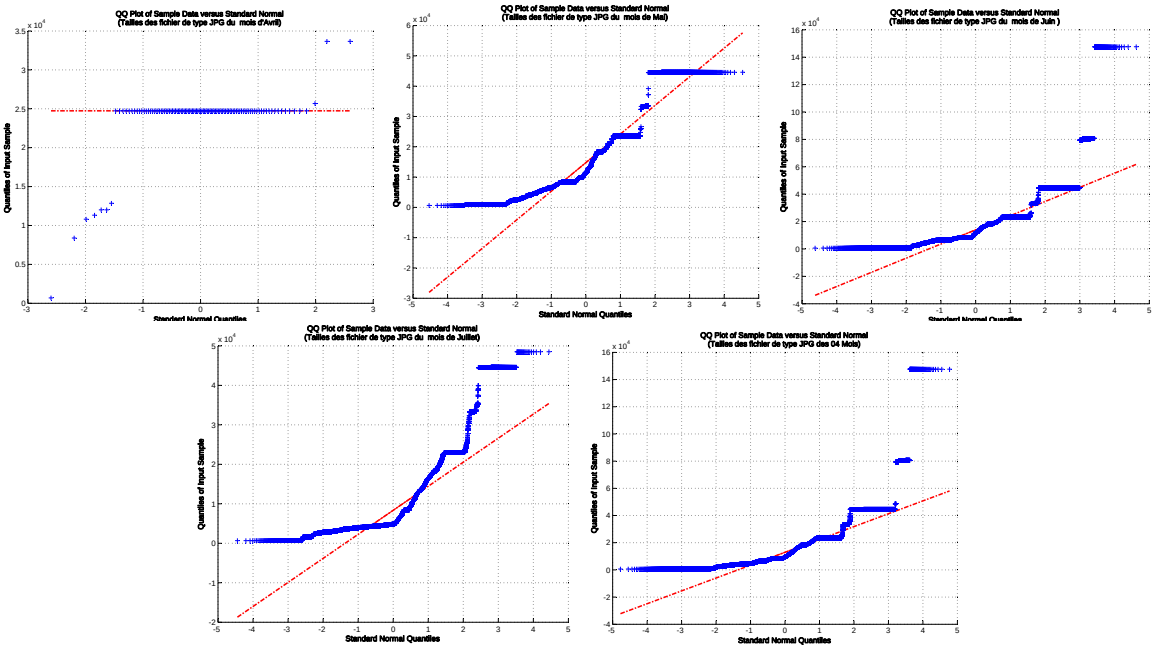

FIGURE 3.12 – QQ-Plot de la taille des fichiers du type '.JPG' en fonction de la loi normale.

Discussion

A partir de ces résultats graphique, on conclue que les distributions de la taille des différents types des fichiers ont des queues plus lourdes que la normale. Mais, comme on la citer auparavant, dans la théorie des valeurs extrêmes, le QQ-plot se base sur la distribution exponentielle. Le QQ-plot sous l'hypothèse d'une distribution exponentielle est la représentation des quantiles de la distribution empirique sur l'axe des X contre les quantiles de la fonction de distribution exponentielle sur l'axe des Y.

Afin de déterminer si :

1. Les données suivent la loi exponentielle (la distribution présente une queue très légère, les points du graphique présentent une forme linéaire).
2. Les données suivent une distribution à queue épaisse *"heavy-tailed distribution"* (le graphique QQ-plot est concave).
3. Les données suivent une distribution à queue légère *"light-tailed distribution"* (le graphique QQ-plot a une forme convexe).

Ainsi, la représentation des quantiles de la distribution empirique de nos différentes données contre les quantiles de la fonction de distribution exponentielle nous a fourni les résultats suivants :

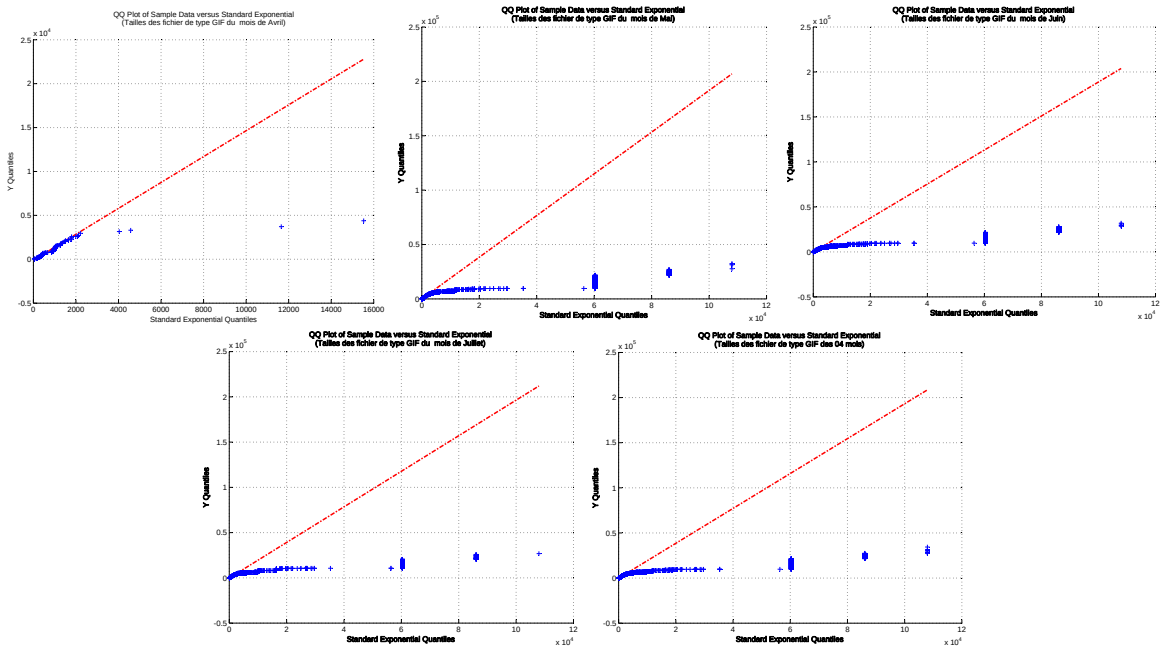

FIGURE 3.13 – QQ-Plot de la taille des fichiers du type '.gif' en fonction de la loi exponentielle.

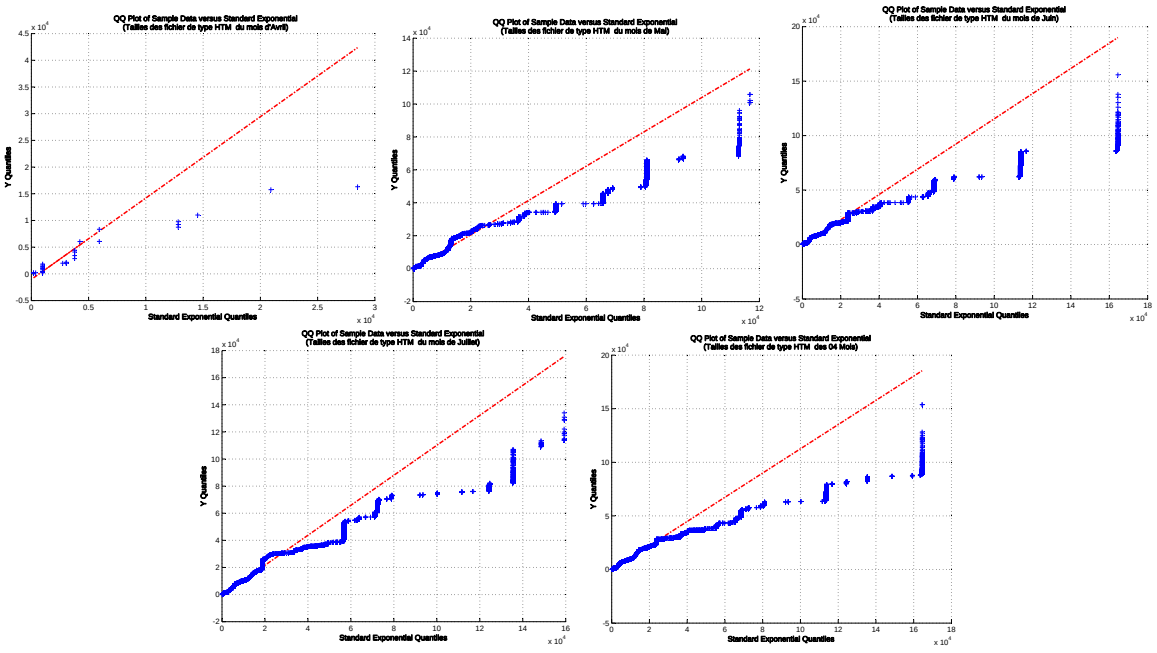

FIGURE 3.14 – QQ-Plot de la taille des fichiers du type '.htm' en fonction de la loi exponentielle.

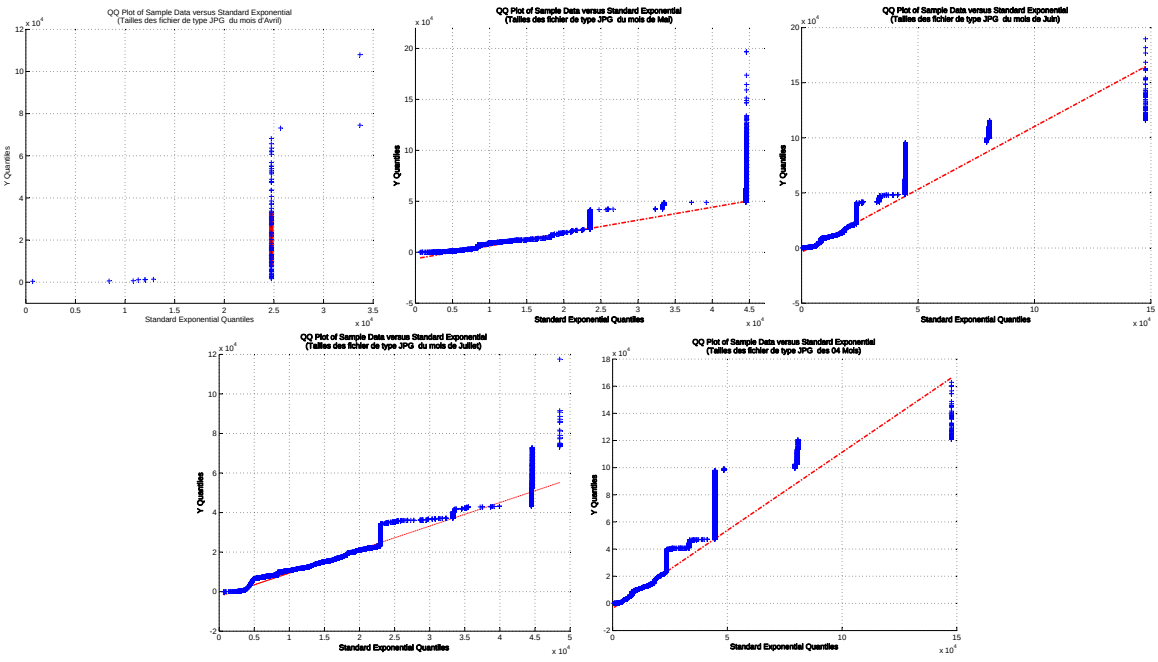

FIGURE 3.15 – QQ-Plot de la taille des fichiers du type '.jpg' en fonction de la loi exponentielle.

Discussion

A partir des figures 3.13, 3.14 et 3.15 on constate que :

- Les distributions de la taille des fichiers 'gif' ont des queues plus lourdes que celle d'une distribution exponentielle.
- Les distributions de la taille des fichiers 'htm' ont des queues plus lourdes que celle d'une distribution exponentielle mais moins lourdes par rapport aux fichiers de type 'gif'.
- Les distributions de la taille des fichiers 'jpg' sont des distributions à queue lourde aux extrémités droites par rapport à une distribution exponentielle, à l'exception de celles du mois d'Avril.

Les deux méthodes précédentes nous ne donnent qu'un aperçu sur la qualité des distributions en question. Afin de confirmer les résultats précédents, on mesure l'indice de queue de ces distributions (leptokurtique) par l'estimateur non paramétrique de Hill et la méthode du maximum de vraisemblance par l'indice de variabilité des valeurs extrêmes (GEV).

3.5.3 Estimateur de Hill

La distribution heavy-tailed figure parmi les distributions statistiques fortes (distribution en loi puissance). On peut déterminer la classe des distributions de la taille des fichiers par une seule valeur, celle du l'indice de variation ξ. Pour l'estimation de cet indice, nous avons fait appel à l'estimation non paramétrique (estimateur de Hill).

Les résultats graphiques obtenus par l'application de cette dernière méthode à nos données nous ont fourni ce qui suit :

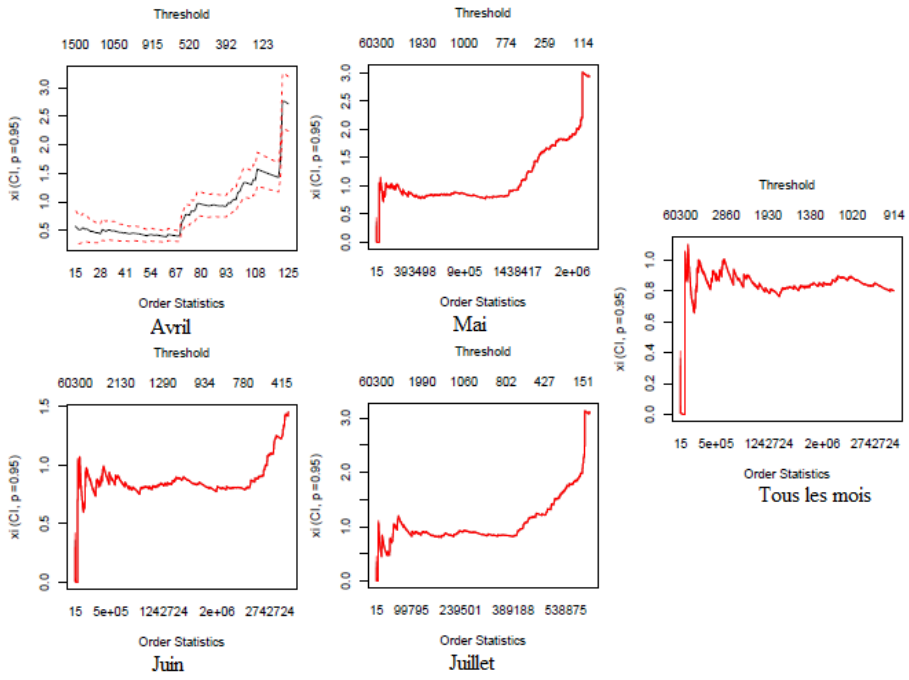

FIGURE 3.16 – Estimation de l'indice de variabilité ξ pour les fichiers du type '.gif'.

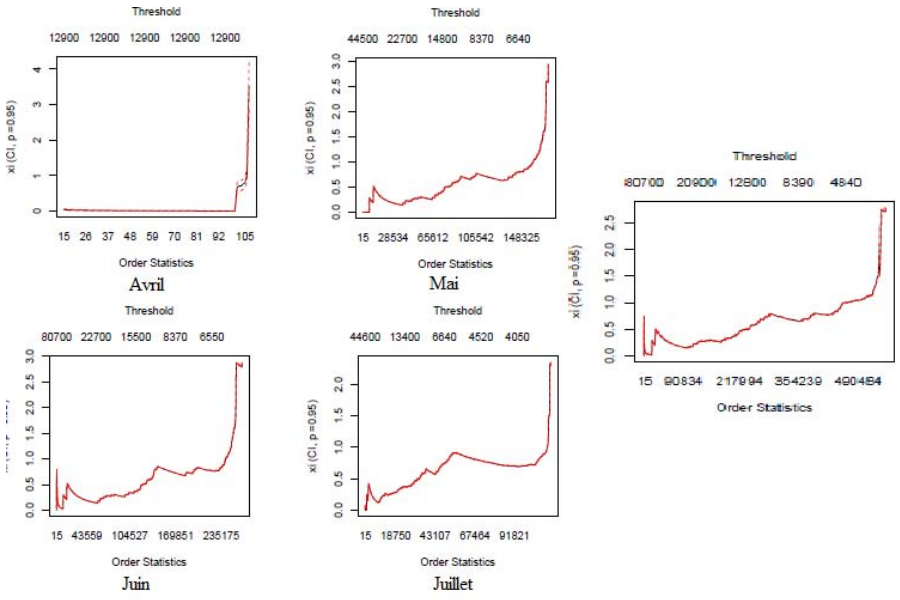

FIGURE 3.17 – Estimation de l'indice de variabilité ξ pour les fichiers du type '.jpg'.

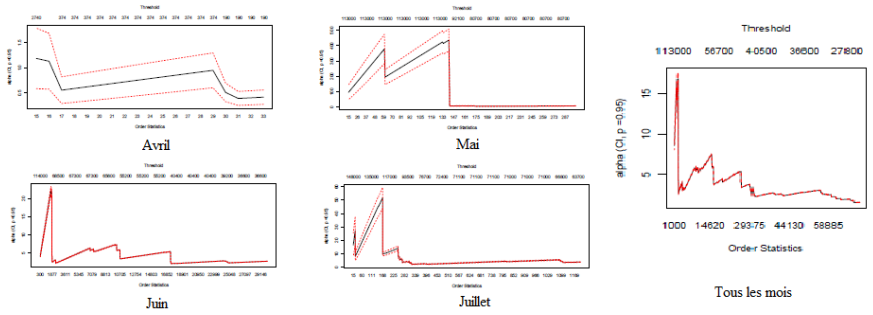

FIGURE 3.18 – Estimation de l'indice de variabilité ξ pour les fichiers du type '.htm'.

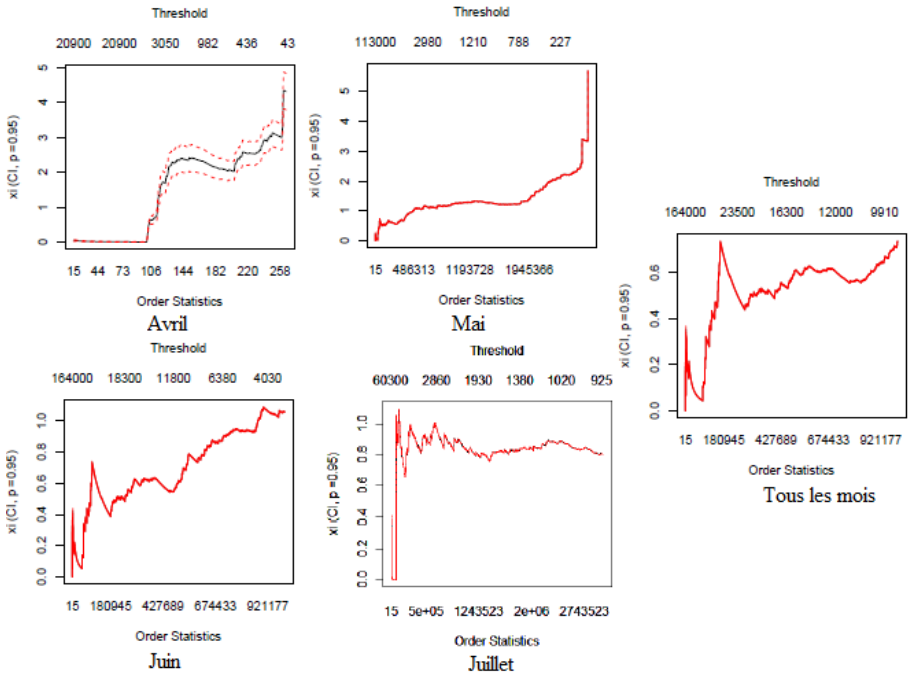

FIGURE 3.19 – Estimation de l'indice de variabilité ξ pour tous les fichiers.

Discussion

Les graphiques précédents présentent les estimateurs de l'indice de variabilité ξ en fonction des valeurs ordonnées de nos différentes données. Les valeurs de l'indice de queue obtenu à partir des graphes précédents sont rangées dans le tableau suivant :

Fichiers téléchargés	Mois	Paramètre ξ
gif	Avril	0.3681
	Mai	0.3402
	Juin	0.3331
	Juillet	0.3216
	4 mois	0.3334
jpg	Avril	0.2847
	Mai	0.3418
	Juin	0.3490
	Juillet	0.4272
	4 mois	0.3592
htm	Avril	0.4182
	Mai	0.1405
	Juin	0.1373
	Juillet	0.1340
	4 mois	0.1380
Image (gif, jpg)	4 mois	0.3131
Tous les fichiers	4 mois	0.1741

TABLE 3.6 – Estimation de l'indice de variabilité.

Discussion

Les valeurs de ξ obtenus nous indiquent que la distribution de la taille les fichiers 'gif' ainsi que celle de la taille des fichiers 'jpg' ont un indice de queue qui vaut approximativement 0.33. Cela signifie, que les distributions en question appartient au MDA de Fréchet.

Les valeurs de ξ associés aux distributions de la taille des fichiers de type 'gif' sont très proches ce qui signifie que la distribution de la taille de ce type de fichier est la même dans tous les mois.

Les valeurs de ξ obtenus à partir des données de la taille des fichiers de type 'htm' nous indique que la distribution de la taille de ce type de fichier appartient au MDA de Fréchet. De plus, la distribution des données du mois d'avril à une queue plus lourde que les autres données correspondantes respectivement aux mois de Mai, de Juin et de juillet. Cet écart peut être expliqué par la taille d'échantillon qui est très petite pour le mois d'avril.

3.5.4 Méthode du maximum de vraisemblance

L'une des autres méthodes d'estimation de l'indice de queue est le maximum de vraisemblance pour l'estimation des paramètres d'une loi GEV. L'application de cette technique à nos données nous a retourné les résultats rangés dans le tableau suivant :

Fichiers téléchargés	Mois	Paramètre ξ	Intervalle de confiance de ξ
gif	Avril	0.4009	[0.2178 , 0.5840]
	Mai	0.7561	[0.7543 , 0.7580]
	Juin	0.6931	[0.6918 , 0.6944]
	Juillet	0.6882	[0.6853 , 0.6911]
	4 mois	0.7116	[0.7106 , 0.7126]
jpg	Avril	-0.4152	[-0.4725 ,-0.3580]
	Mai	0.1432	[0.1389 , 0.1476]
	Juin	0.1296	[0.1263 , 0.1330]
	Juillet	0.4301	[0.4259 , 0.4343]
	4 mois	0.2147	[0.2121 , 0.2174]
htm	Avril	0.9761	[0.4829 , 1.4692]
	Mai	0.5212	[0.5169 , 0.5254]
	Juin	0.5828	[0.5797 , 0.5860]
	Juillet	0.5100	[0.5029 , 0.5172]
	4 mois	0.5636	[0.5612 , 0.5660]
Image (gif, jpg)	4 mois	0.9224	[0.9214 , 0.9234]
Tous les fichiers	4 mois	1.0573	[1.0563 , 1.0583]

TABLE 3.7 – Estimation de l'indice de variabilité par le MLE.

Discussion

On remarque que toutes les valeurs obtenues de ξ de tous les données sont positives sauf pour le mois d'avril (cas des fichiers du type jpg) :

$\xi > 0$ signifie que la distribution de nos données (la majorité) appartient au MDA de Fréchet regroupant la majorité des distributions à queue lourde.

$\xi < 0$ signifie que la distribution de la taille de ce type de fichiers appartient au MDA de Weibull (comme par exemple la distribution de la taille des fichier de type '.jpg' du mois d'Avril).

3.5.5 Comparaison

Les quatre méthodes précédentes graphique (fonction de survie et QQ-plot) et numériques (estimateur de Hill et MLE) nous ont fournis des résultats similaires qui indiquent que les distributions de nos différentes données sont à queue lourde. la différence numérique peut être expliquée par le manque de données et/ou par la limitation d'application de ces méthodes.

A titre d'exemple, on voit clairement que l'estimateur de Hill nous a fourni un estimateur erroné, dans le cas de la taille des fichiers de type 'jpg' du mois d'avril, du fait que la vraie distribution de ce dernier, appartient au MDA Weibull (sans queue) et l'estimateur

de Hill n'est applicable que sur des distributions appartenant au MDA de Fréchet (queue lourde).

Toujours dans le cas de l'estimateur de Hill, la différence numérique peut être expliquée, aussi, par le manque de précision et sa phase de stabilisation, qui devrait correspondre à la valeur de l'indice de queue, et qui est difficile à identifier.

D'un point de vue théorique, toutes les méthodes d'estimation de l'indice de queue partagent les mêmes propriétés de consistance et de normalité asymptotique. Cependant, les simulations montrent qu'il y a de grandes différences entre ces différents estimateurs. En général, il n'y a pas une meilleure méthode que les autres.

3.6 Estimation de la distribution des données

Dans cette partie nous nous sommes intéressées à l'estimation des distributions de la taille de chaque type de fichiers ainsi que tout le trafic.

3.6.1 Estimation paramétrique : Tests d'ajustement

L'idée la plus naturelle, pour la détermination de la loi d'un échantillon est l'utilisation des tests d'ajustement paramétriques. Comme notre échantillon présente un comportement à queue lourde (voir section 3.5), le choix de F_0 sera dans la famille des lois puissance (famille à queue lourde). Les résultats du test Kolmogorov-Smirnov pour un risque α=5% sur les différents échantillons pour les différentes lois sont présentés dans le tableau 3.8.

D'après le test de Kolmogorov-Smirnov effectué sur la taille des fichiers Web, on déduit que la série des tailles des fichiers Web ne provient d'aucune loi connue, ($D_n > D_\alpha$) on rejette l'hypothèse H_0 : "$F = F_0$", avec F_0 une fonction de répartition continue spécifiée dans tous les cas.

A partir des résultats du tableau, on peut conclure aussi que :

- La loi de la la taille des fichiers " .gif " est proche d'une loi log-normale,
- La loi de la la taille des fichiers " .jpg " est proche d'une loi log-normale,
- Tandis que la loi de la taille des fichiers du type " .htm " est proche d'une loi gamma. car D_n est minimal pour ces cas.

Par conséquent, on utilise la méthode d'estimation non paramétrique du noyau pour estimer la distribution de la taille des données en question.

Remarque 3.2. Pour d'autres valeurs du risque α nous avons obtenu les mêmes résultats que précédemment, c'est-à-dire on rejette l'hypothèse H_0 dans tous les cas.

Echantillon	Loi	Paramètre(s)	D_n(calculé)	D_α(valeur tabulé)
GIF	heavy-tailed (Pareto,$\alpha = 1$)	0.1490	0.4642	0.0005
	Pareto	$(41, 0.3334)$	0.1026	
	Log-normale	(6.7126 , 1.2448)	0.1018	
	Expenontielle	2118.5	0.2320	
	Weibull	(1548.8996 , 0.7212)	0.1216	
	Gamma	(0.6462 , 3278.239)	0.1733	
JPG	heavy-tailed	0.1082	0.5440	0.0018
	Pareto	$(637, 0.3592)$	0.1699	
	Log-normale	(9.2410 , 0.7723)	0.0914	
	Expenontielle	13347.8103	0.1824	
	Weibull	(14845.1797 , 1.4965)	0.1306	
	Gamma	(2.0887 , 6390.4877)	0.1263	
HTM	heavy-tailed	0.1158	0.4961	0.0013
	Pareto	(4 , 0.1380)	0.1143	
	Log-normale	(8.6336 , 1.2104)	0.1230	
	Expenontielle	10338.9863	0.1082	
	Weibull	(9993.9971 , 0.9351)	0.1062	
	Gamma	(0.9518 , 10862.7561)	0.0979	
Tous les fichiers	heavy-tailed	0.1388	0.4493	0.0005
	Pareto	(0.7868 , 1433.9829)	0.0625	
	Log-normale	(7.2021 , 1.5243)	0.0657	
	Expenontielle	6365.1647	0.3848	
	Weibull	(2908.6199 , 0.5916)	0.1047	
	Gamma	(0.4195 , 15171.5290)	0.1966	

TABLE 3.8: Résultats du test de Kolmogorov-Smirnov

3.6.2 Estimation non paramétrique : Méthode du noyau

La méthode du noyau dépend de deux paramètres, le paramètre de lissage h (la fenêtre) et le noyau K. Notre choix pour le couple (h, K) est illustré dans les deux paragraphes suivants :

Choix du noyau K

Comme on l'a cité auparavant (voir section 2.7) si on s'intéresse à l'estimation d'une densité de probabilité par la méthode du noyau définie sur un support positif, ce qui est le cas de nos données, il est préférable d'utiliser les noyaux asymétriques. C'est la raison pour

laquelle notre choix s'est fixé sur les noyaux asymétriques les plus usuels, à savoir : gamma 1, gamma 2, IG et RIG donnés dans les formules 2.26, 2.34, 2.40 et 2.41 respectivement.

Choix du paramètre de lissage h

Dans la pratique la règle du pouce, proposée par Silverman (1986), est fréquemment utilisée pour le choix du h^* optimal au sens du MISE pour le noyau gaussien. Une règle analogue peut être suggérée pour les distributions heavy-tailed (respectivement les distributions Pareto). En effet, si X est une variable aléatoire d'une distribution heavy-tailed de paramètre α (respectivement une distribution Pareto de paramètres α et k) alors le h^* optimal au sens du MISE dans ce cas est donné par la proposition 3.1 (respectivement 3.2).

Proposition 3.1. *Si on suppose que f est une densité d'une distribution de heavy-tailed(Pareto, $\alpha = 1$) de paramètre k. Alors le h^* optimal au sens du MISE est donné par :*
Pour le noyau gamma 1,

$$h^*_{G_1} = \left[\frac{(2k+3)\left(\frac{1}{\sqrt{\pi}}\right)}{k(k+1)\left[4(k+1)(2k+3) - 2(k+2)(2k+3)(2k+1) + (k+1)(k+2)^2(2k+1)\right]} \right]^{\frac{2}{5}} n^{\frac{-2}{5}}. \tag{3.1}$$

Pour le noyau gamma 2,

$$h^*_{G_2} = \left[\frac{(2k+3)}{\sqrt{\pi}k(k+1)^2(k+2)^2(2k+1)} \right]^{\frac{2}{5}} n^{\frac{-2}{5}}. \tag{3.2}$$

si de plus $k > \frac{1}{2}$, alors pour le noyau inverse gaussien,

$$h^*_{IG} = \left[\frac{(2k-1)}{\sqrt{\pi}k(k+1)^2(k+2)^2(2k+1)(2k+3)} \right]^{\frac{2}{5}} n^{\frac{-2}{5}}. \tag{3.3}$$

Pour le noyau réciproque de l'inverse gaussien,

$$h^*_{RIG} = \left[\frac{(2k+3)}{\sqrt{\pi}k(k+1)^2(k+2)^2(2k+1)} \right]^{\frac{2}{5}} n^{\frac{-2}{5}}. \tag{3.4}$$

Proposition 3.2. *Si on suppose que $f(x)$ est une distribution de Pareto de paramètres α et k le h^* optimal au sens du MISE est donné par :*
pour le noyau gamma 1 ;

$$h^{**}_{G_1} = \left[\frac{k\alpha^{\frac{1}{2}}/\sqrt{\pi}(2k+1)}{k(k+1)^2\left(\frac{4}{2k+1} - \frac{4(k+2)}{\alpha(2k+2)} + \frac{(k+2)^2}{\alpha^2(2k+3)}\right)} \right]^{\frac{2}{5}} n^{\frac{-2}{5}}. \tag{3.5}$$

pour le noyau gamma 2 ;

$$h^{**}_{G_2} = \left[\frac{(2k+3)\alpha^{\frac{5}{2}}}{(\sqrt{\pi}(2k+1)(k+1)^2(k+2)^2)}\right]^{\frac{2}{5}} n^{\frac{-2}{5}}. \tag{3.6}$$

pour le noyau inverse gaussien,(si $k > \frac{1}{2}$).

$$h^{**}_{IG} = \left[\frac{k\alpha^{\frac{-5}{2}}}{\sqrt{\pi}(2k+3)} * \frac{2k-1}{(k(k+1)(k+2))^2}\right]^{\frac{2}{5}} n^{\frac{-2}{5}} \tag{3.7}$$

$$h^{**}_{RIG} = \left[\frac{(2k+3)\alpha^{\frac{5}{2}}}{(\sqrt{\pi}(2k+1)(k+1)^2(k+2)^2)}\right]^{\frac{2}{5}} n^{\frac{-2}{5}}. \tag{3.8}$$

Calcul du paramètre de lissage h

Nous constatons que pour estimer les paramètres de lissage cités dans les deux propositions 3.1 et 3.2, il est nécessaire d'estimer d'abord le paramètre de la loi heavy-tailed et ceux de la loi de Pareto. Pour cela, nous avons utilisé la méthode du maximum de vraisemblance pour les estimer.

Estimation de paramètre k de la loi heavy-tailed :

On a

$$L(X_1, ..., X_n; k) = \prod_{i=1}^{n} \frac{k}{X_i^{k+1}}. \tag{3.9}$$

En introduisant le logarithme, on obtient :

$$\log(L(X;k)) = n\log(k) - (k+1)\sum_{i=1}^{n}\log(X_i) \Rightarrow \frac{\partial \log(L(X;k))}{\partial k} = \frac{n}{k} - \sum_{i=1}^{n}\log(X_i) = 0,$$

avec $X = (X_1, ...X_n)$.

ce qui donne

$$\hat{k} = \frac{n}{\sum_{i=1}^{n} log(X_i)}. \tag{3.10}$$

Estimation du paramètre α et k de la loi Pareto :

Estimation de paramètre α :

On a :

$$L(X_1, ..., X_n; \alpha, k) = \prod_{i=1}^{n} \frac{k\alpha^k}{X_i^{k+1}} \tag{3.11}$$

vu que le domaine de définition de la loi de Pareto dépend du paramètre α nous ne pouvons pas estimer α de la manière usuelle. D'autre part, on constate que $L(X;\alpha,k)$ est

décroissante en fonction α, alors son estimateur de maximum de vraisemblance ne peut être que la statistique $\mathrm{T} = \inf(X_i) \quad \Rightarrow \hat{\alpha} = \min(X_i)$

Estimation de paramètre k On a :

$$L(X_1, ..., X_n; \alpha, k) = \prod_{i=1}^{n} \frac{k\alpha^k}{X_i^{k+1}}$$

Après l'introduction de logarithme, on aura :

$$\begin{aligned} \log(L(X;\alpha,k)) &= n\log(k\alpha^k) - (k+1)\sum_{i=1}^{n}\log(X_i) \\ \Rightarrow \quad & \frac{\partial \log(L)}{\partial k} = \frac{n}{k} + n\log(\alpha) - \sum_{i=1}^{n} Log(X_i) = 0. \end{aligned} \tag{3.12}$$

ce qui donne

$$\hat{k} = \frac{n}{\sum_{i=1}^{n} log\left(\frac{X_i}{\alpha}\right)}. \tag{3.13}$$

L'estimation des différents paramètres de lissage pour les différents noyaux considérés sur tous les sous échantillons sont rangés dans le tableau suivant :

		Méthodes de sélection du paramètre du lissage					
Type	Mois	h_{bcv}	h_{ucv}	$h^*_{G_1}$	$h^*_{G_2}$	$h^{**}_{G_1}$	$h^{**}_{G_2}$
GIF	Avril	362.945	79.71153	0.2211	0.1698	0.1647	2.5691
	Mai	279.4622	201.6705	0.0045	0.0035	0.0035	0.0534
	Juin	335.1984	335.1984	0.0036	0.0028	0.0028	0.0419
	Juillet	469.6119	433.0935	0.0076	0.0059	0.0060	0.0889
	04 mois	123.9476	311.5385	0.0029	0.0022	0.0022	0.0337
HTM	Avril	3742.956	390.7652	0.4012	0.3204	0.3464	18.6953
	Mai	960.5359	960.5359	0.0104	0.0085	0.0083	0.0139
	Juin	363.7889	273.7803	0.0082	0.0067	0.0066	0.0110
	Juillet	1514.396	1661.327	0.0167	0.0137	0.0134	0.0223
	04 mois	354.8974	909.386	0.0066	0.0054	0.0053	0.0088
JPG	Avril	1831.586	191.2026	0.4012	0.3204	0.3522	49.4086
	Mai	940.456	98.3652	0.0104	0.0085	0.0162	2.2250
	Juin	97.10973	930.6331	0.0082	0.0067	0.0135	1.8624
	Juillet	198.3123	82.62597	0.0167	0.0137	0.0168	2.4787
	04 mois	535.0769	80.04327	0.0066	0.0054	0.0099	1.3714
image	04 mois	71.82692	352.1172	0.0028	0.0022	0.0023	0.0334
tous les fichiers	04 mois	130.9746	350.5085	0.0027	0.0021	0.0021	0.0037

TABLE 3.9 – Différents paramètres de lissage optimaux

On constate que les paramètres du lissage obtenus par les méthodes UCV et BCV sont trop grands et ceux des G1 et G2 (resp RIG) sont trop petits. A priori, les paramètres h_{ucv} et h_{bcv} nous fournissent des estimateurs sur-lissés, tandis que ceux h^* et h^{**} nous fournissent des estimateurs sous-lissés.

Cas des noyaux gamma

Les résultats graphiques obtenus pour l'estimation de la densité des différents sous-échantillons et l'échantillon global sont présentés dans les figures suivantes :

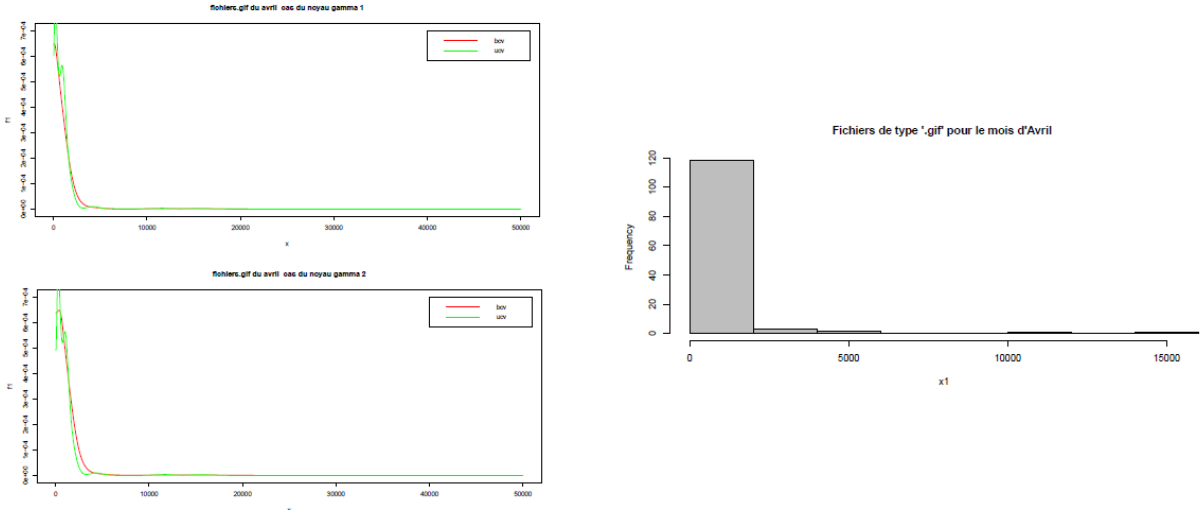

FIGURE 3.20 – Distribution de la taille des fichiers du type ".gif" pour le mois d'avril

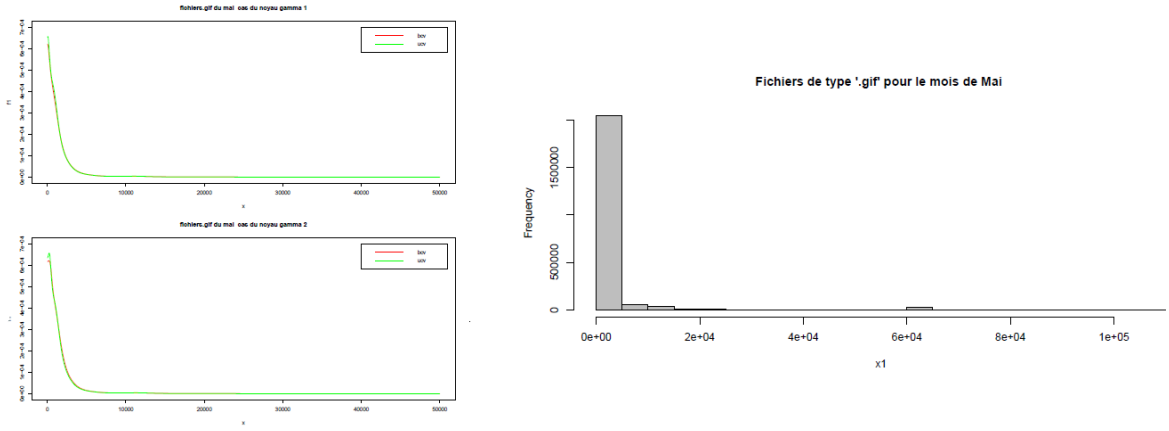

FIGURE 3.21 – Distribution de la taille des fichiers du type ".gif" pour le mois de mai

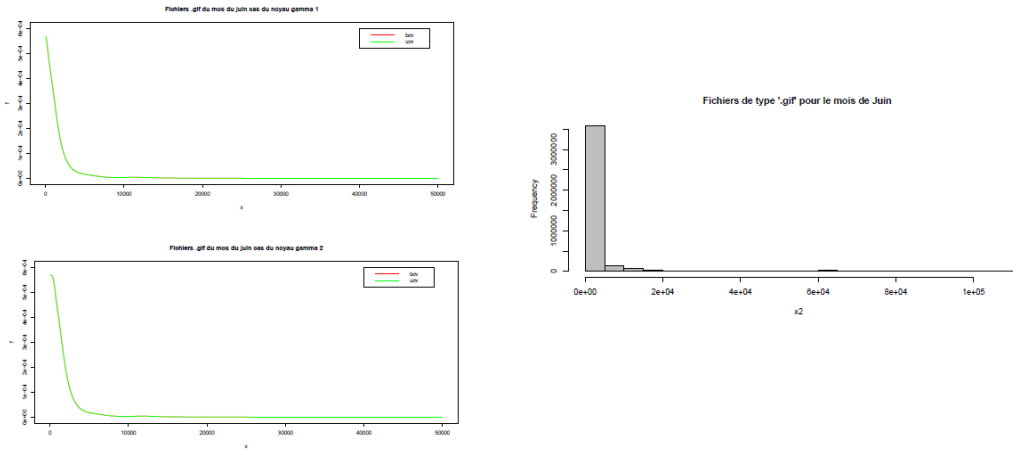

FIGURE 3.22 – Distribution de la taille des fichiers du type ".gif" pour le mois de juin

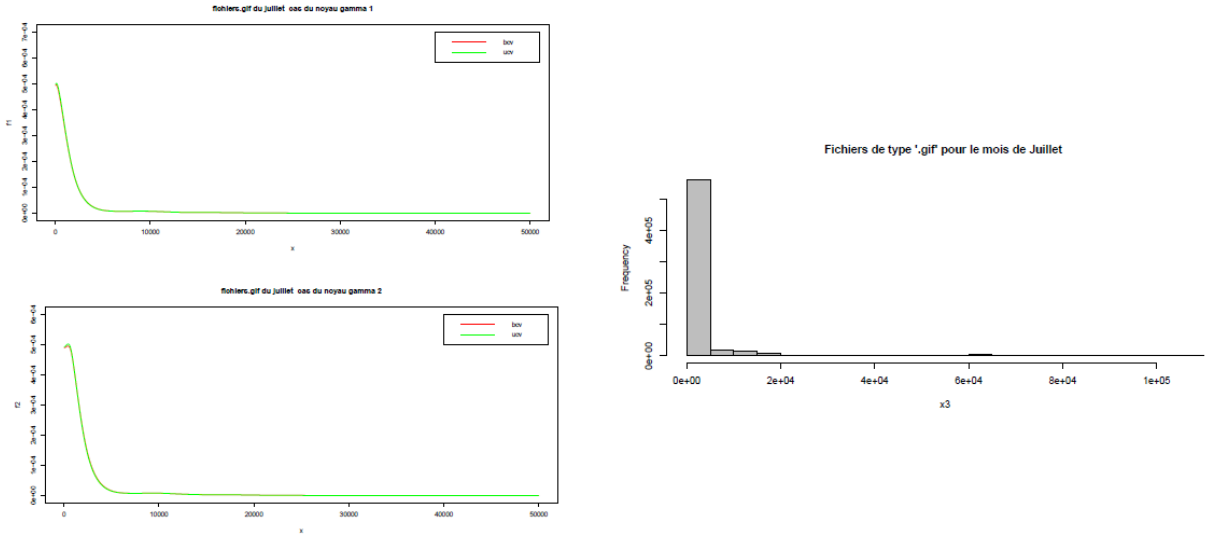

FIGURE 3.23 – Distribution de la taille des fichiers du type ".gif" pour le mois de juillet.

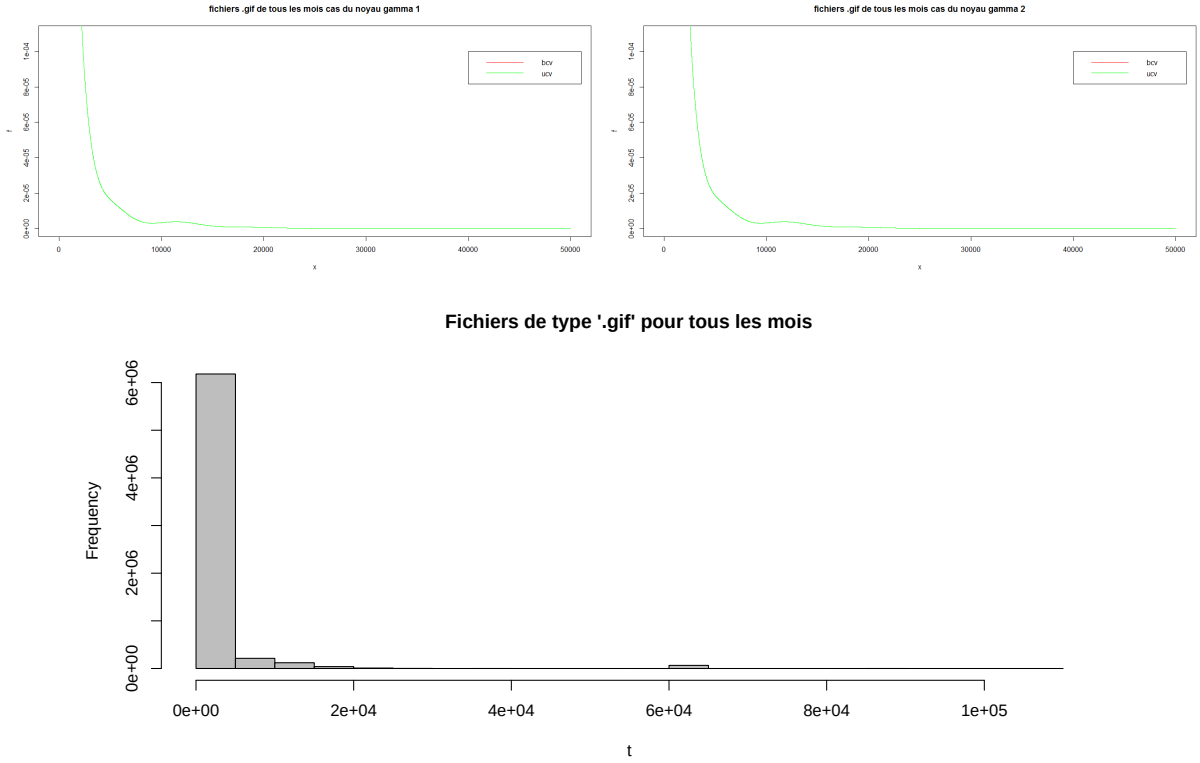

FIGURE 3.24 – Distribution de la taille des fichiers du type ".gif" pour tous les mois.

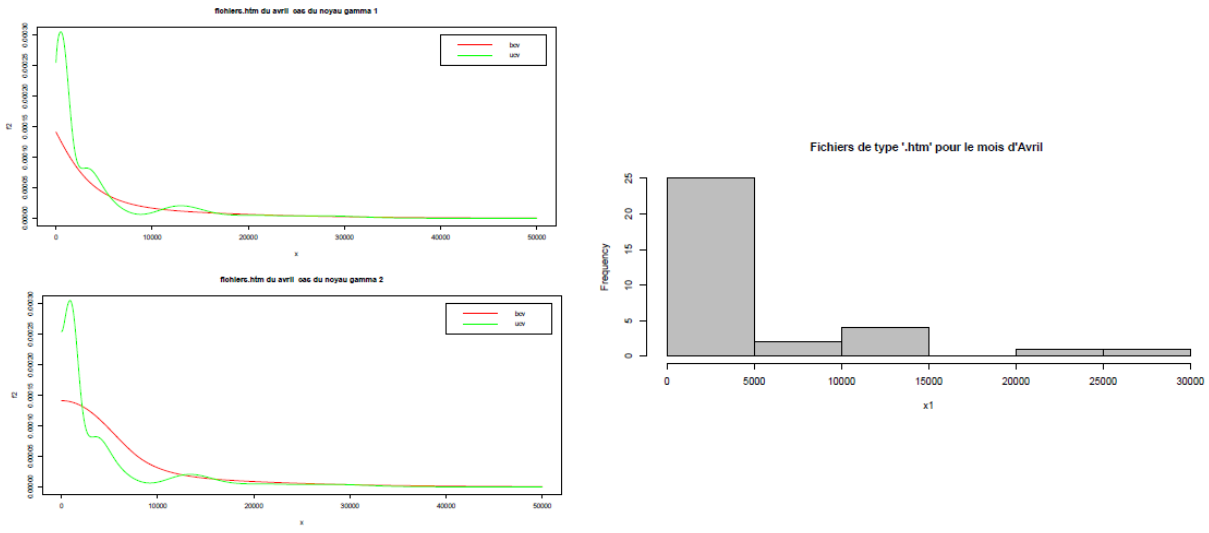

FIGURE 3.25 – Distribution de la taille des fichiers du type ".htm" pour le mois d'avril

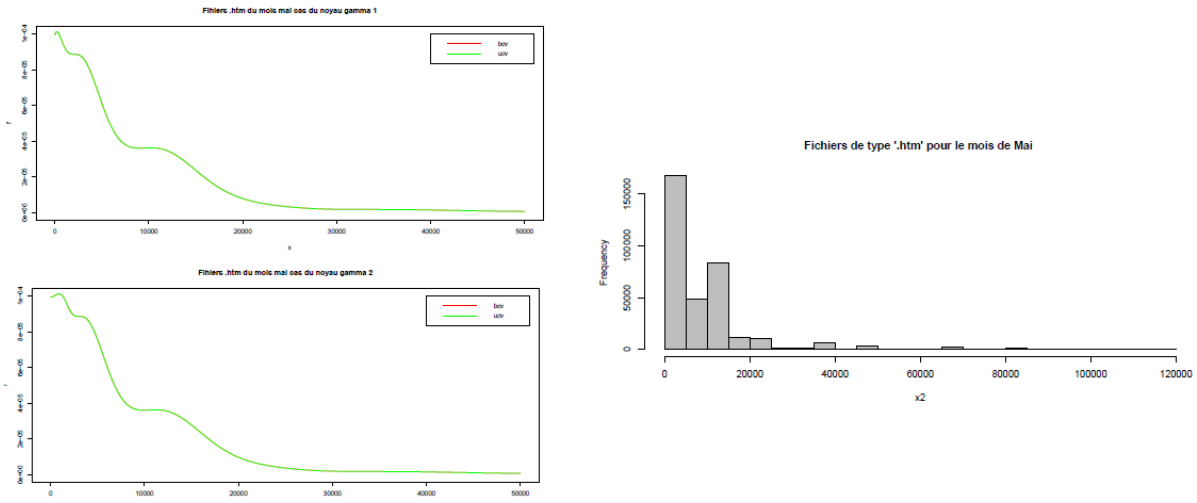

FIGURE 3.26 – Distribution de la taille des fichiers du type ".htm" pour le mois mai

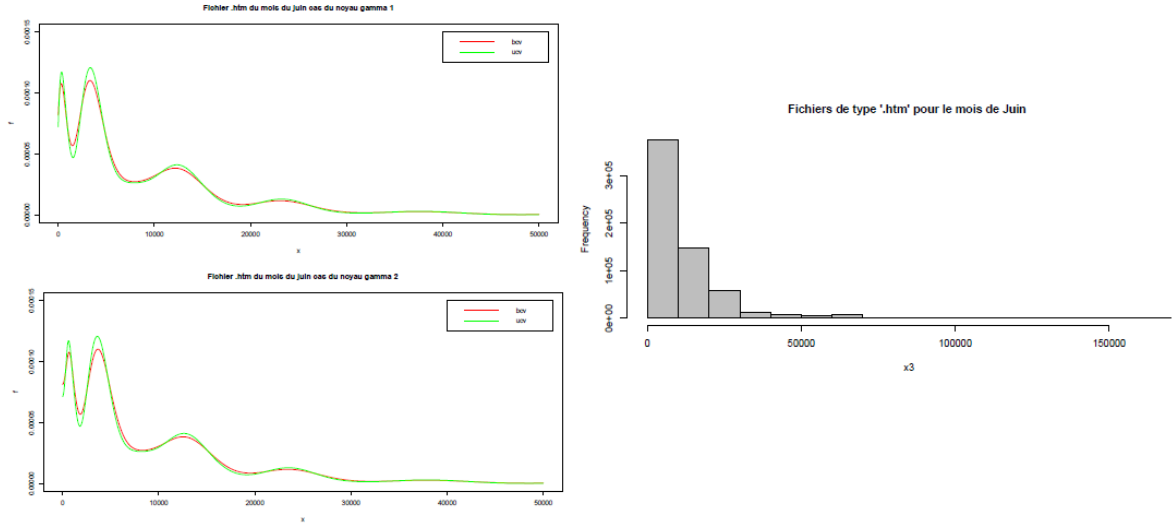

FIGURE 3.27 – Distribution de la taille des fichiers du type ".htm" pour le mois juin

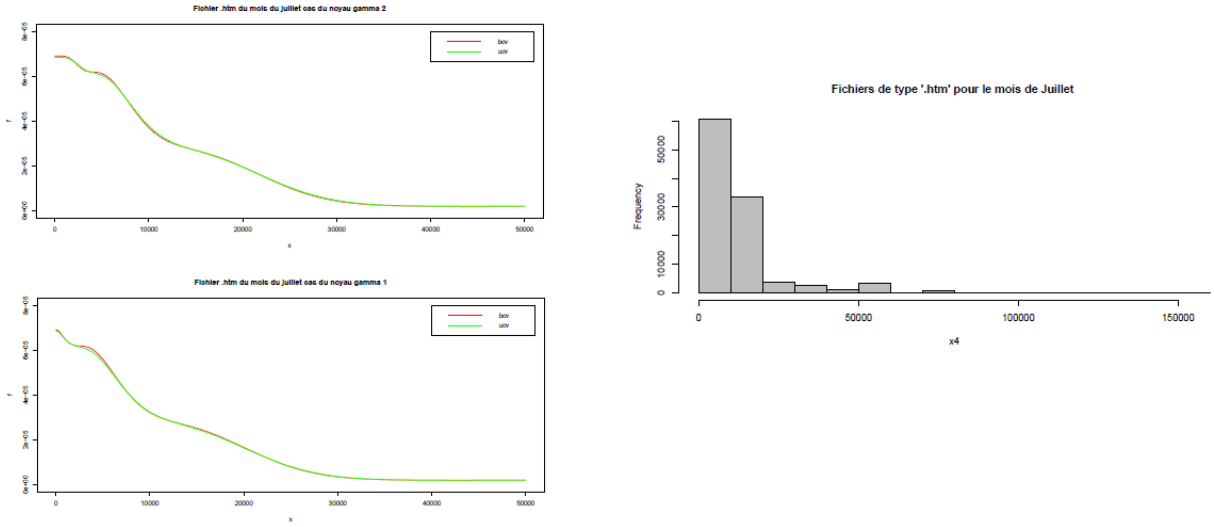

FIGURE 3.28 – Distribution de la taille des fichiers du type ".htm" pour le mois juillet

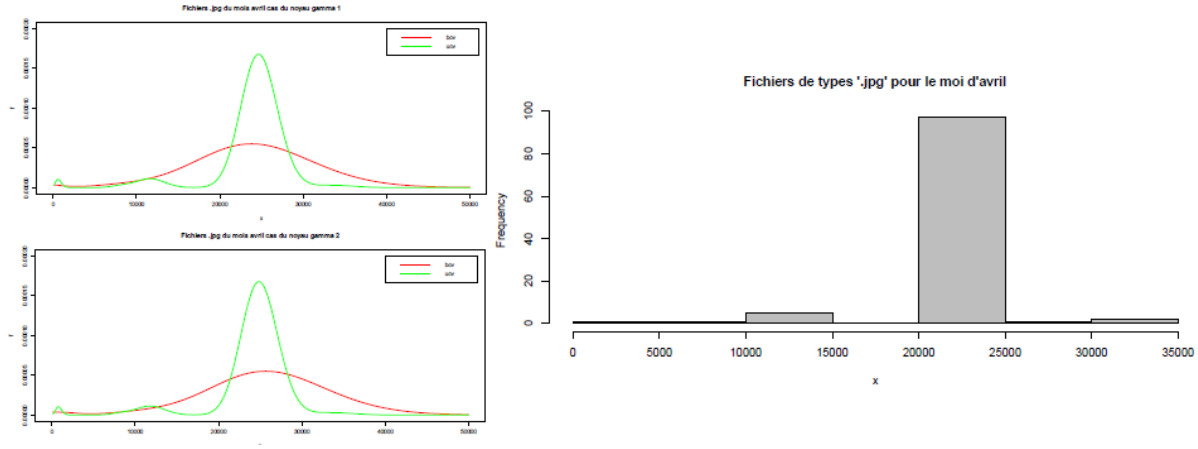

FIGURE 3.29 – Distribution de la taille des fichiers du type ".jpg" pour le mois d'avril

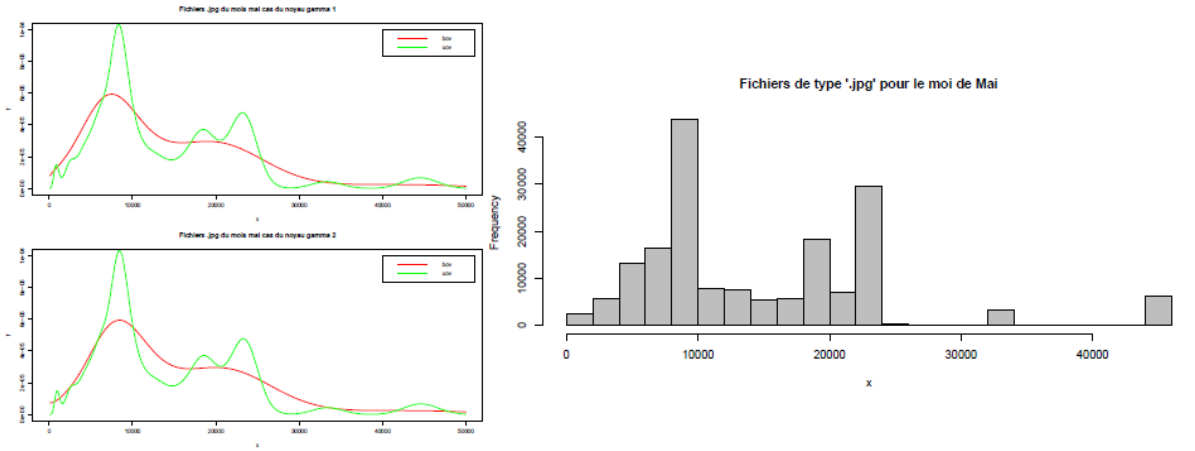

FIGURE 3.30 – Distribution de la taille des fichiers du type ".jpg" pour le mois mai

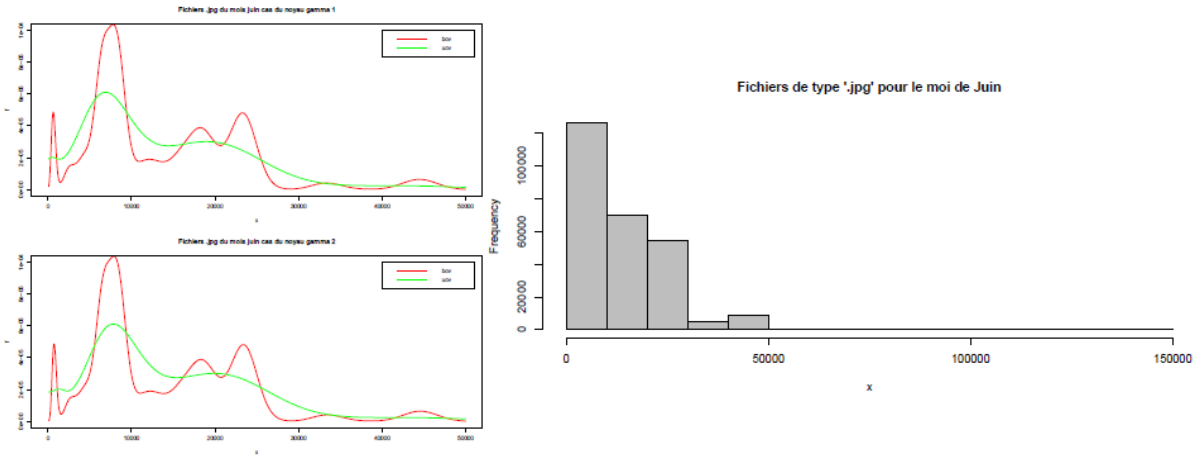

FIGURE 3.31 – Distribution de la taille des fichiers du type ".jpg" pour le mois juin

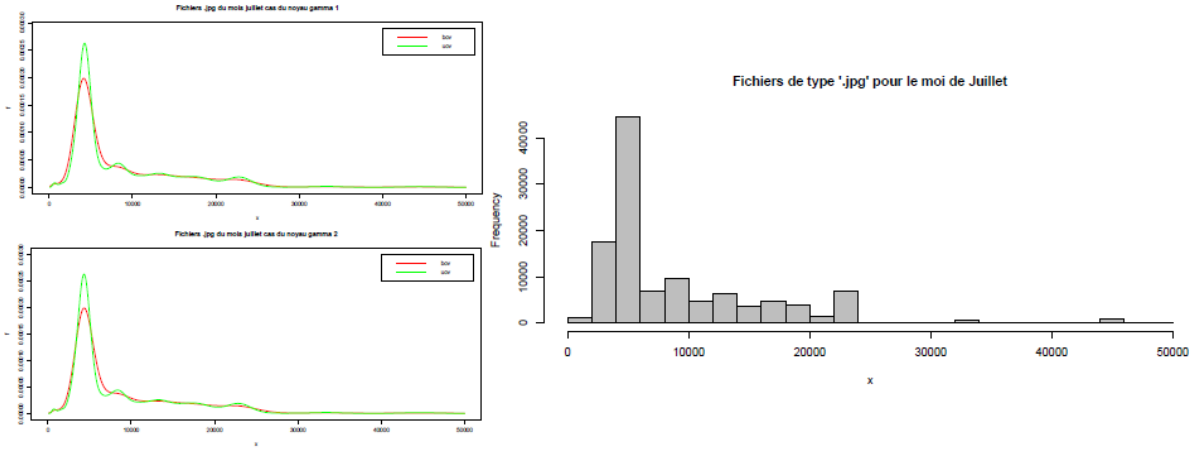

FIGURE 3.32 – Distribution de la taille des fichiers du type ".jpg" pour le mois juillet

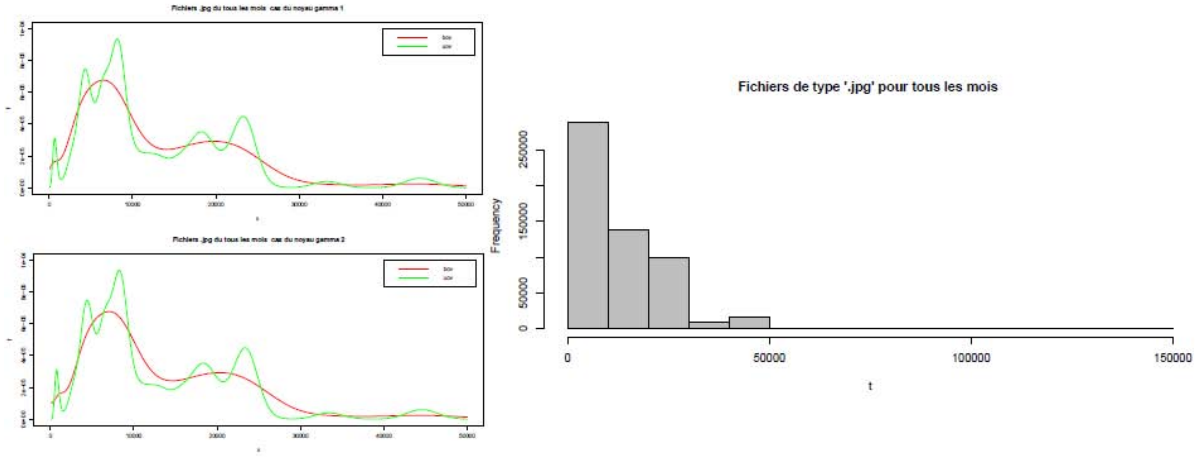

FIGURE 3.33 – Distribution de la taille des fichiers du type ".jpg" pour tous les mois

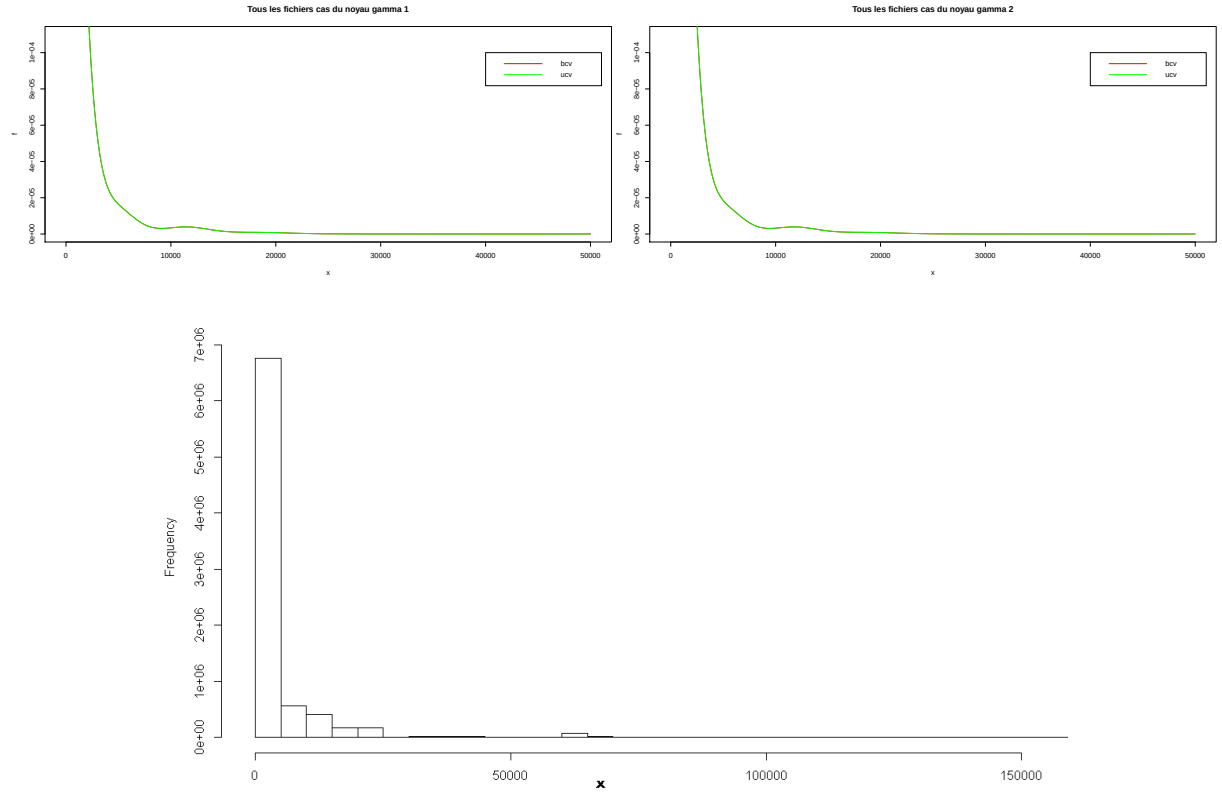

FIGURE 3.34 – Distribution de la taille de tous les fichiers pour tous les mois

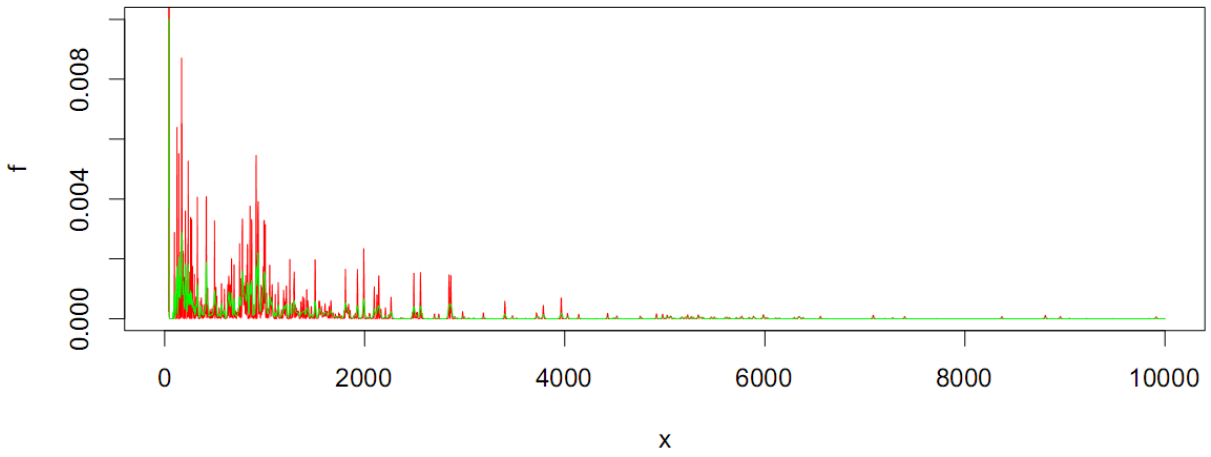

FIGURE 3.35 – Distribution de la taille des fichiers du type ".gif" cas du h^{**} pour le noyau gamma

Discussion

D'après les résultats graphiques, on constate que :

Les estimateurs obtenus par les noyaux gamma pour le même paramètre de lissage sont très semblables dans tous les cas.

Cas des fichiers du type '.jpg' : Les paramètres de lissage h_{bcv} nous fournissent des estimateurs sur-lissés (à l'exception des données du mois de juin et tous les mois où h_{ucv} nous fournit un estimateur sur-lissé). On peut conclure qu'il est préférable d'utiliser le paramètre h_{ucv} pour l'estimateur de densité de la taille des fichiers du type '.jpg'.

Cas des fichiers du type '.htm' : A l'exception des données du mois d'avril, h_{ucv} nous fournit un estimateur sur-lissé. h_{ucv} et h_{bcv} nous fournissent des bons estimateurs de la densité de la taille des fichiers '.htm' et cela pour les deux noyaux gamma.

Cas des fichiers du type '.gif' : Dans tous les mois h_{ucv} et h_{bcv} nous donnent des bons estimateurs, de plus ils sont semblables à l'exception du mois d'avril où il y a une légère différence et cela pour les deux noyaux gamma, vu que la taille de l'échantillon pour ce mois est petite.

Cas de tous les fichiers : D'après les résultats obtenus dans la section 3.4.3 nous avons constaté que notre échantillon est dominé par les fichiers du type 'gif' qui forment 80% de l'échantillon global. On constate sur les figures que la densité de la taille de tous les types des fichiers est très semblable à la loi de la taille des fichiers 'gif'.

Les paramètres de lissage optimaux au sens du MISE en remplaçant $f(x)$ par la densité Pareto (respectivement Heavy-Tailed) (méthode rule of thumb) nous fournissent des estimateurs sous lissés dans tous les cas considérés. L'échec de cette méthode peut être expliqué par le fait que les différentes méthodes plug-in ne donnent des bons résultats que pour des densités suffisamment lisses et régulières, et leurs performances s'amoindrissent en présence de cibles plus complexes(Pour plus de détails voir [58]), comme par exemple des densités multimodales qui est le cas de nos données.

Cas des noyaux IG et RIG

Les résultats graphiques obtenus pour l'estimation de la densité des différents sous-échantillons et l'échantillon global en utilisant cette fois-ci les noyaux IG et RIG sont présentés dans les figures suivantes :

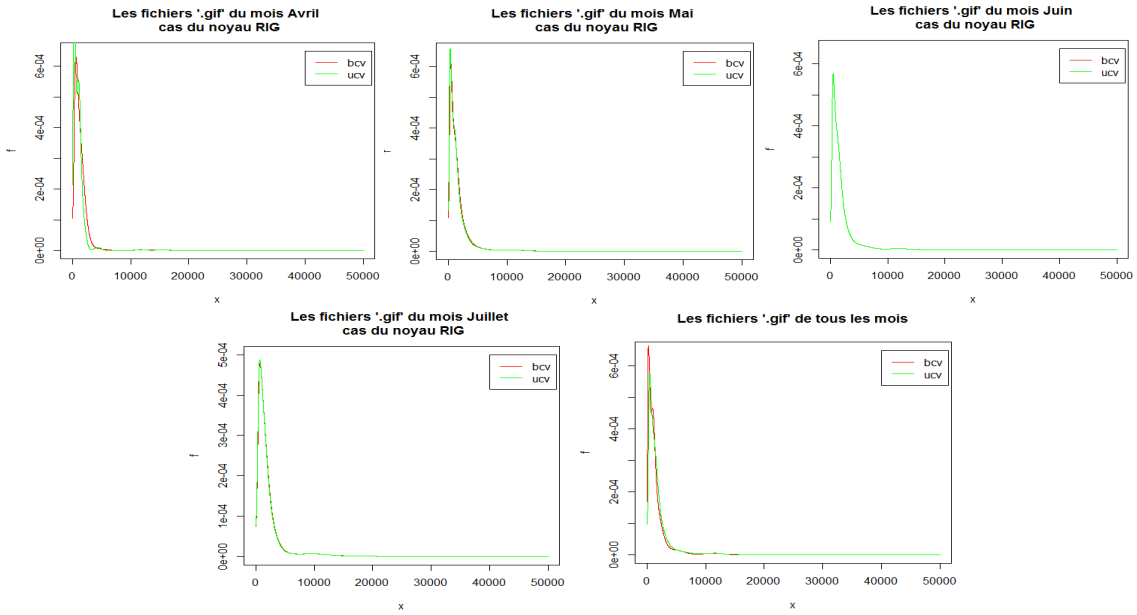

FIGURE 3.36 – Distribution de la taille des fichiers du type ".gif"

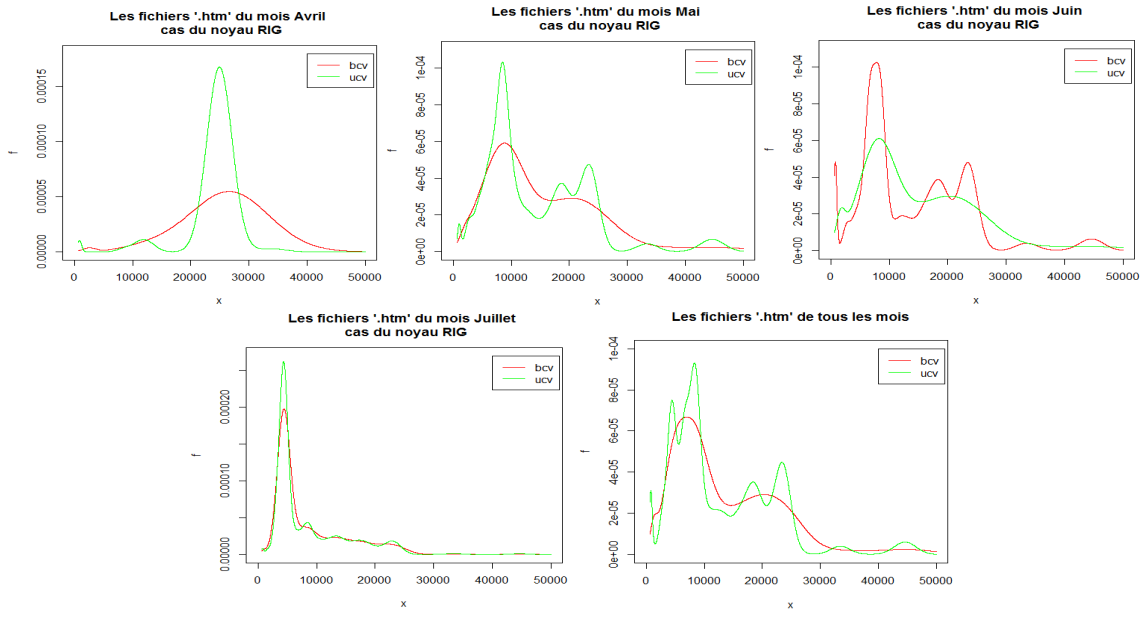

FIGURE 3.37 – Distribution de la taille des fichiers du type ".jpg"

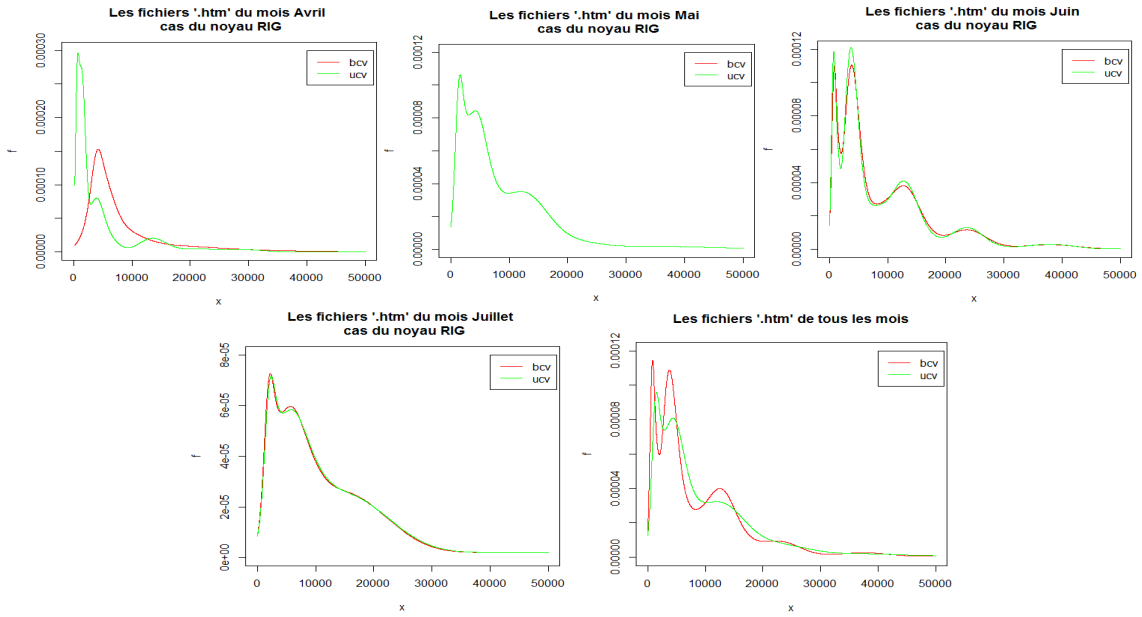

FIGURE 3.38 – Distribution de la taille des fichiers du type ".htm"

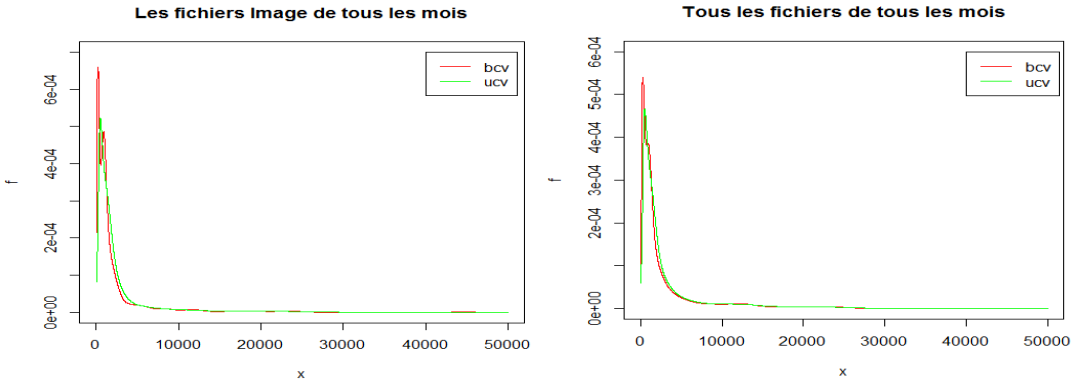

FIGURE 3.39 – Distributions de la taille des Images et la taille de tous les fichiers pour tous les mois cas du noyau RIG.

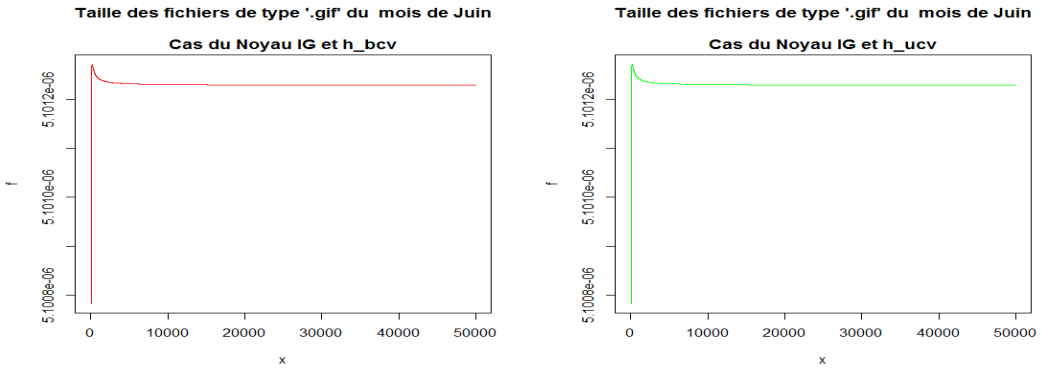

FIGURE 3.40 – Distribution de la taille des fichiers '.gif' du mois de Juin cas du noyau IG.

Discussion

A partir des estimateurs, de la densité de la taille des fichiers, obtenus par le noyau RIG et le noyau IG trois importantes remarques peuvent être tirées :

⋆ Le noyau RIG nous fournit des estimateurs pratiquement semblable aux estimateurs des noyaux gamma, pour le cas des fichiers '.htm' et 'jpg'.

⋆ Le noyau RIG est introduit dans la littérature comme solution pour le problème d'effet du biais aux bornes, mais de la figure 3.36 on peut déduire que ce noyau n'est pas vraiment une solution efficace autant que les noyaux gamma.

⋆ Le noyau IG nous fournit des estimateurs divergents.

3.7 Conclusion

Dans ce chapitre nous avons réalisé une analyse statistique de quelques caractéristiques du trafic web du serveur de la coupe de monde France 98. Les principaux résultats obtenus peuvent être résumés en quatre points.

⋆ L'évaluation de l'indice de queue, des tailles des différents types de fichiers, indique que ces derniers sont issus de la famille des lois à queue lourde.

⋆ Le test d'ajustement $K.S.$ nous a indiqué que le comportement de la taille des fichiers ne suit pas une distribution usuelle. Ce qui nous ramène à l'utilisation d'estimation non paramétrique d'une densité de probabilité.

⋆ L'utilisation des noyaux asymétriques ($gamma$, RIG et IG) dans l'estimation des distributions de la taille des fichiers par la méthode du noyau nous indique que :
les paramètres de lissage optimaux au sens du MISE en remplaçant $f(x)$ par la densité de Pareto (respectivement Heavy-Tailed) (méthode rule of thumb) nous fournissent des estimateurs sous lissés dans tous les cas considérés.

Les paramètres obtenus par les méthodes de sélection "ucv" et "bcv" nous donnent des bons estimateurs (graphiquement) dans la majorité des cas.

⋆ Les caractéristiques statistiques (distribution, indice de queue,...) du trafic web du serveur de la coupe du monde 98 sont engendrées par les fichiers du type GIF (image) qui forme plus de 80% des fichiers qui circulent dans le trafic web.

Conclusion Générale

Les distributions en "lois puissance" fascinent des chercheurs de toutes disciplines. De nombreuses études ont montré que de telles distributions se retrouvaient sur la fréquence des signes en linguistique, les revenus en économie, la taille des villes, la durée de vie en biologie ou encore les fluctuations turbulentes ou la distribution des degrés des nœuds de nombreux graphes. Dans ce travail, nous avons proposé de modéliser un échantillon du trafic web du serveur de la coupe du monde France 98 par une loi puissance (Heavy-Tailed). Notre objectif est axé dans deux directions : l'une sur l'estimation du paramètre de variabilité, l'autre sur l'estimation de la distribution des données.

Avant de réaliser l'étude en question, nous avons exposé, les principales notions sur les lois puissance et les lois Heavy-Tailed, qui sont suivies de quelques outils statistiques qui nous permettent d'identifier ou de quantifier leurs caractéristiques, à savoir : les méthodes de mesure de l'indice de variabilité (méthode paramétrique, semi-paramétrique et non paramétriques) et les tests d'ajustement qui nous permettent de déterminer s'il est correct d'approcher une distribution observée par une telle loi. L'échec des tests précédents est envisageable, pour cela nous avons mis l'accent sur la méthode du noyau qui s'utilise dans ce genre de situation. En effet, nous avons introduit le principe de base de l'estimation à noyau de Parzen-Rosenblatt, ses propriétés et les méthodes de sélection du noyau et du paramètre de lissage. Nous avons également abordé le problème de l'effet du biais aux bornes des estimateurs à noyaux symétriques (pour des données bornées ou semi-bornées) suivi des estimateurs à noyaux asymétriques (gamma, IG et RIG).

Lorsque on a feuilleté les données brutes du trafic web du serveur en question nous avons constaté que certaines requêtes n'ont pas été téléchargées(échec de téléchargement) pour diverses raisons(ressource manquante, URL non conforme, domaine inexistant, accès interdit, etc). Pour réaliser l'étude nous n'avons pris en considération que le cas de réussite du téléchargement.

Après une analyse des données, nous avons constaté que les fichiers téléchargés sont de différents types : vidéo, image, texte, etc. Les types de fichiers n'ont pas la même taille moyenne (en général, taille vidéo > taille image > taille texte). A cet effet, pour analyser notre échantillon nous avons proposé de le décomposer, selon le type des fichiers. De plus,

les résultats de la classification de Pareto (80/20) de nos données, nous ont suggéré de nous limiter à l'étude des fichiers dominants GIF, HTM et JPG (97% de l'échantillon global).

Ensuite, nous nous sommes intéressées à la détermination de la famille de la distribution des tailles de chaque type de fichiers au sens de son leptokurtique. Ainsi, l'estimation de l'indice de variabilité des différents échantillons à l'aide des méthodes(fonction de survie, Hill,...) nous confirme que les distributions de la taille des fichiers font partie de la famille des distributions à queue lourde.

La dernière partie est consacrée à la modélisation d'une situation réelle(trafic web) par les lois puissance (Heavy-Tailed).

L'échec du test de K-S pour l'identification d'une loi usuelle du comportement de la taille d'un type de fichier a fait de l'estimation non paramétrique (estimateurs à noyau) l'objet du reste du travail.

Pour l'estimation de la densité en question, nous sommes contraints d'utiliser les noyaux asymétriques (gamma), vu que nos données sont définies sur un support positif. Pour le choix du paramètre de lissage, nous avons proposé d'utiliser le h optimal (qui minimise le MISE pour une distribution de Pareto (respectivement Heavy-Tailed)). Cette approche pour le choix de h ne nous fournit que des estimateurs sous lissés. D'autre part, le h_{ucv} et le h_{bcv} nous donnent dans la majorité des situations considérées de bon estimateurs. Cela signifie, que le choix de la méthode de sélection du paramètre de lissage, pour nos données, est très important, qui est pareil même pour le choix du noyau. En effet, si les noyaux gamma nous fournissent de bons et semblables estimateurs, il n'est pas le cas pour le noyau RIG et le noyau IG. Le noyau RIG ne remédié pas totalement le problème d'effet du biais aux bornes et le noyau IG fournit des estimateurs divergents.

Le comportement (caractéristiques) du trafic Web, selon nos données, est fortement lié aux caractéristiques des fichiers du type GIF. Cela peut être expliqué par une dominance (80%) des fichiers du type GIF par rapport aux autres types de fichiers.

Parmi les perspectives de ce travail, nous pouvons dégager plusieurs axes intéressants, tant sur le plan théorique que pratique :

- ⋆ Il serait intéressant, de faire une projection de la présente démarche sur les données actuelles du web.
- ⋆ Il serait intéressant, d'étudier le taux de charge selon le type de fichier.
- ⋆ Inclure lors de l'étude le cas des données censurées (cas d'échec), ou encore la dépendance physique des données (exemple : une page htm téléchargées entraine automatiquement le téléchargement d'autre type de fichiers).
- ⋆ Appliquer les noyaux associés sur tout le trafic web du serveur de la coupe du monde 98 (de 30/04/1998 jusqu'au 26/07/1998).
- ⋆ Développer d'autres résultats théoriques, telle la technique de choix du paramètre de lissage pour des distributions appartenant à la famille heavy-tailed.

Bibliographie

[1] J. Beirlant, Y. Goegebeur, J. Segers, and J. Teugels. *Statistics of Extremes*. Wiley, 2004.

[2] J. Beirlant and J.L. Teugels. *Extreme Value Theory*, chapter Asymptotic Normality of Hill's estimator. Springer-Verlag, New York, 1987.

[3] J. Beirlant, P. Vynckier, and J. L. Teugels. Excess function and estimation of the extreme-value index. *Bernoulli*, 2 :293–318, 1996.

[4] J. Beirlant, P. Vynckier, and J.L. Teugels. Tail index estimation, pareto quantile plots, and regression diagnostics. *Journal of the American Statistical Association*, (91) :1959–1667, 1996.

[5] N. Benhamida. *Sur les politiques de gestion du cache d'un serveur web. Mémoire magister, Informatique Option : Réseaux et Systèmes Distribués*. Université de Béjaia,2007.

[6] T. Bouezmarni and J.V.K. Rombouts. Nonparametric density estimation for positive time series. *Econometric Theory*, September 21, 2006.

[7] T. Bouezmarni and O. Scaillet. Consistency of asymmetric kernel density estimators and smoothed histograms with application to income data. *Econometric Theaory*, (21) :390–412, 2003.

[8] A.W. Bowman. An alternative method of cross-validation for the smoothing density estimates. *Biometrika*, (71) :353–360, 1984.

[9] B.M. Brown and S. Chen. Beta-bernstein smoothing for regression curves with compact supports. *Scandinavian Journal of Statistics*, (26) :47–59, 1999.

[10] Y. Caron. *Contribution de la loi de Zipf à l'analyse d'images. Thèse doctorat en Informatique*. PhD thesis, septembre 2004.

[11] A. Charpentier, J. D. Fermanian, and O. Scaillet. *The estimation of copulas : theory and practice*. Ensae-Crest and Katholieke Universiteit Leuven, BNP-Paribas and Crest, HEC Genève and Swiss Finance Institute,(Revised Proof Ref : 33259e), September 29 2006.

[12] S. Chen. A beta kernel estimation for density functions. *Computational Statistics and Data Analysis*, (31) :131–145, 1999.

[13] S. Chen. Beta kernel for regression curve. *Statistica Sinica*, (10) :73–92, 2000.

[14] S. Csörgö and D.M. Mason. Central limit theorems for sums of extreme values. *In Mathematical Procedings of the Cambridge philosophical Society*, 98 :547–558, 1985.

[15] L. de Haan and S. Resnick. A simple asymptotic estimate for the index of a stable distribution. *Journal of the Royal Statistical Society, Ser. B*, (42) :83–87, 1980.

[16] P. Deheuvels and P. Homminal. Estimation non paramétrique de la densité compte tenu d'informations sur le support. *Revue de statistique Appliquée*, (27) :47–68, 1979.

[17] A.L.M. Dekkers and L. de Haan. Optimal choice of simple fraction in extreme-value estimation. *Journal of Multivariate Analysis*, (47) :173–195, 1993.

[18] T. Delleji and M.S. Bouhlel. Evaluation de deux métriques pour la mesure de la qualité des images compressées par la norme jpeg. Institut Supérieur de Biotechnologie de Sfax (ISBS), Mars 2005.

[19] L. Devroye. The equivalence of weak, strong and complete convergence in l1 for kernel density estimates. *The Annals of Statistics*, (11) :896–904, 1983.

[20] L. Devroye and L. Györfi. Nonparametric density estimation : The l1, view, new york. *John Wiley*, 1985.

[21] P. Diggle. A kernel method for smoothing point process data. *Applied Statistics*, (34,138-147), 1985.

[22] Y. Dodge. *Statistique Dictionnaire encyclopédique*. Université de Neuchâtel Suisse, 2007.

[23] Y. Dodge and G. Melfi. Premiers pas en simulation. Technical report, Université de Neuchâtel,Suisse, 2008.

[24] G. Draisma, L. de Haan, L. Peng, and T.T. Pereira. A bootstrap-based method to achieve optimality in estimating the extreme-value index. *Extremes*, (2) :367–404, 1999.

[25] V. A Epanechnikov. Nonparametric estimation of a multidimensional probability density. *Theory Probab. Appl*, (14) :153–158, 1969.

[26] J. Fan and Q. Yao. Nonlinear time series. *Springer-Verlag, New York.*, 2003.

[27] M. Fernandez and P. Monteiro. Central limit theorem for asymmetric kernel functionals. *Annals of the Institute of Statistical Mathematics*, (57) :425–442, 2005.

[28] P. Gallinari. *Recherche d'information textuelle*. 2012.

[29] P. Hall and J. S. Marron. Local minima in cross-validation function. *Journal of the royal statistical society*, (90) :149–173, 1991.

[30] P. Hall and A.H. Welsh. Adaptive estimates of parameters of regular variation. *The Annals of Statistics*, (13) :331–341, 1985.

[31] P.G. Hall. On some simple estimates of an exponent of regular variation. *J. Royal Stata. Soc. B*, (44) :37–42, 1982.

[32] F.E. Harrell and C.E. Davis. A new distribution-free quantile estimator. *Biometrika*, (69) :635–640, 1982.

[33] B.M. Hill. A simple general approach to inference about the tail of a distribution. *Anal of statistics*, (3) :1163–1174, 1975.

[34] W. Härdle. Applied nonparametric regression. *Cambridge University Press, UK*, 1990.

[35] E. Häusler and J.L. Teugels. On asymptotic normality of hill's estimator for the exponent of regular variation. *Annals of Statistics*, (13) :743–756, 1985.

[36] M.C. Jones. Simple boundary correction for kernel density estimation. *Statistical Computing*, (3) :135–146, 1993.

[37] M. Lejeune and P. Sarda. Smooth estimators of distribution and density functions. *Computational Statistics and Data Analysis*, (14) :457–471, 1992.

[38] A. Mahanti. *Web proxy workload characterization and modeling. Master's thesis,Department of computer science.* université saskatchewan, September 1999.

[39] J. S. Marron and D. Ruppert. Transformations to reduce boundary bias in kernel density estimation. *Journal of the Royal Statistical Society, Series B*, (56) :653–671, 1994.

[40] M. Mason. Laws of large numbers for sums of extreme values. *Annals of Probability*, (10(3)) :754–764, 1982.

[41] I. Mathlouthi and A. Zenaidi. *Théorie des valeurs extrêmes vs méthodes classiques de calcul de la VaR : Application au Tunindex.* ECOFI, Institut des Hautes Etudes Commerciales, Carthage, 2006 Tunisie.

[42] H.G. Müller. Smooth optimum kernel estimators near endpoints. *Biometrika*, (78) :521–530, 1991.

[43] E. Nadaraya. On nonparametric estimation density function and regression. *Theory Probab P.P.L*, (10) :186–190, 1965.

[44] N. Victor NG. *Variation régulière et application aux réseaux de télécommunication, Mémoire Master Recherche en Probabilités et Statistiques.* Université Paul Sabatier, Toulouse III, 2007-2008.

[45] A. Pagan and A. Ullah. Nonparametric econometrics. *Cambridge University Press, UK*, 1999.

[46] E. Parzen. On estimation of a probability density function and mode. *Ann. Math. Statist*, (33) :1065–1076, 1962.

[47] J. Pickands. Statistical inference extreme order statistics. *Anal of statistics*, (3) :119–130, 1975.

[48] O. Powell. *Algorithmes pour le Web.* Université de Genève, 14 décembre 2005.

[49] C. Resnick and S. Starica. Smoothing the moment estimtor of the extreme value parameter. *Extremes*, (1) :263–293, 1999.

[50] M. Rosenblatt. Remarks in some nonparametric estimates of a density function. *Ann. Math. Statist*, (27) :832–837, 1956.

[51] M. Rudemo. Empirical choice of histogram and kernel density estimators. *Scandinavian Journal of Statistics*, (9) :65–78, 1982.

[52] A. Sadoun and Y.Ziane. Sur la caractérisation du trafic web : étude statistique. Master's thesis, Mémoire Master 2 en Recherche Opérationnelle Option Modélisation et Mathématique et Technique de Décision, Département de Recherche Opérationnelle, Université A/Mira de Béjaïa, juin 2011.

[53] O. Scaillet. Density estimation using inverse and reciprocal inverse gaussian kernels. Technical report, HEC Genève and FAME, Université de Genève, Bd Carl Vogt 102, CH - 1211 Genève 4, Suisse., January 2003.

[54] E. Schuster. Incorporating support constraints into nonparametric estimators of densities. *Communications in Statistics - Theory and Methods*, (14, 1123-1136), 1985.

[55] D.W. Scott and G.R. Terrell. Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association*, (82) :1131–1146, 1987.

[56] B. Silverman. Weak and strong uniform consistency of the kernel estimate of density function and its derivates. *Ann. Statist*, (6) :177–184, 1978.

[57] B. Silverman. Density estimation for statistics and data analysis. *London : Chapman & Hall*, 1986.

[58] N. Zougab. Etude comparative des méthodes de sélection du paramétrer de lissage dans l'estimation de la densité de probabilité par la méthode du noyau. Thèse de magister, Université de Bejaia, Mai 2007.