Estimation de l'indice des valeurs extrêmes en présence des données censurées Étude de cas : Les durées de chômage en Algérie

Estimation of extreme values index in the presence of censored data Case study : Unemployment durations in Algeria

ZOUADI Nihad [1], SAIDI Ghania [2]

[1] Laboratoire de Modélisation des Processus Stochastiques, ENSSEA, Koléa , zouadi_nihad@yahoo.fr

[2] Laboratoire de Modélisation des Processus Stochastiques, ENSSEA, Koléa , ghsaidi@yahoo.fr

Received: 14/11/2018 **Revised 30/12/2018** **Accepted: 31/12/2018**

Résumé:

Les valeurs extrêmes sont des événements rares ayant une faible probabilité d'apparition puisqu'il s'agit des valeurs beaucoup plus grandes ou plus petites que celles observées habituellement. Dernièrement, la Théorie des Valeurs Extrêmes (TEV) a reçu beaucoup d'attention aussi bien sur le plan théorique que sur le plan pratique. Et récemment, un intérêt s'est porté sur leur application en présence des données censurées.

Ce problème a été mentionné pour la première fois en 1997 dans le livre de Reiss et Thomas mais il n'a été réellement abordé qu'en 2007 par Beirlant et al. Dans cet article, nous nous sommes intéressés à explorer l'apport de cette théorie en matière d'estimation de l'indice de queue des valeurs extrêmes.

Mots clés: TVE, Censures, Application, Estimation, Indice de queue extrême.

Jel Classification Codes: B23, C13, C15, C24, C55.

Abstract:

Extreme values are rare events with a low probability of occurrence since they are much larger or smaller than those usually observed. Recently, Extreme Values Theory (EVT) has received a lot of attention both theoretically and practically. And recently, there has been interest in their application in the presence of censored data.

This problem was mentioned for the first time in 1997 in the book of Reiss and Thomas, but it was actually addressed in 2007 by Beirlant et al. In this article, we are interested in exploring the contribution of this theory in estimating the tail index of extreme values.

Keywords: TVE, Censorship, Application, Estimation, Extreme Tail Index.

Jel Classification Codes: B23, C13, C15, C24, C55.

Auteur correspondant: ZOUADI Nihad, Email: zouadi_nihad@yahoo.fr

1. Introduction:

La Théorie des Valeurs Extrêmes (TVE), apparue entre 1920 et 1940, grâce à Fréchet, Fisher et Tippett, Gumbel et Gnedenko, joue un rôle de plus en plus important dans le traitement de la modélisation des événements rares, puisque son application fournit une méthode relativement sûre pour l'extrapolation au-delà de ce qui a été observé (Embrechts et al., 1997).

Cependant, dans la réalité, quel que soit le phénomène étudié, on fait face toujours aux données incomplètes que ça soit tronqué ou censuré, comme le cas par exemple dans les applications classiques telles que l'analyse des données de durée de vie (analyse de survie, la théorie de fiabilité, l'assurance), qui sont caractérisées par la présence des données censurées.

Définition : (**La variable de censure)**

La variable de censure **C** est définie par la non-observation de l'événement étudié. Si au lieu d'observer X, on observe C, et que l'on sait que X > C (respectivement X < C, C1 < X < C2), on dit qu'il y a **censure à droite** (respectivement, **censure à gauche**, **censure par intervalle**).

Pour l'individu *i,* considérons :

- Son temps de survie ou bien de réalisation *Xi,*
- Son temps de censure *Ci,*
- La durée réellement observée *Ti.*

Une méthode approximative pour corriger ce problème consiste à définir une variable aléatoire non négative Y, indépendante de la variable d'intérêt X, appelée variable aléatoire censurée, et ensuite de considérer la variable $Z = min(X, Y)$ et la variable indicatrice $\delta = 1\mathrm{I}\,(X \leq Y)$ qui est égale à 1 si la variable X a été vraiment observée et 0 sinon (i.e. elle est censurée).

L'analyse de ce type de données (les données censurées) a reçu beaucoup d'attention et un intérêt croissant ces dernières années, surtout dans l'analyse de survie, la fiabilité et la Bio-statistique. Notamment, l'intérêt s'est porté au cas des censures à droite où certaines observations sont connues d'être au moins plus large que les valeurs reportées.

En se référant à la littérature, on trouve pas mal d'études de recherche qui ont été faites dans ce sens, parmi eux en peut citer:

PATHÉ NDAO, (2013), une thèse de doctorat de l'université Gaston Berger de Saint-Louis, portée sur la modélisation de valeurs extrêmes conditionnelles en présence de censure, où il a supposé une variable d'intérêt *Y* qui sera censuré aléatoirement à droite par une autre variable aléatoire *C (indépendante de Y)*, qui sont mesurée simultanément avec une covariable *X*. À cet effet les deux approches de la TVE sont modélisées, et les propriétés asymptotiques du comportement des estimateurs conditionnels de l'indice de queue obtenu à partir de cette méthode étaient vérifiées via une simulation.

Ou encore celle du **M. Ivette Gomes & M. Manuela Neves, (2010),** un article porté sur l'estimation de l'indice de valeur extrême pour les données censurées aléatoirement, où ils accordent une attention particulière à cette estimation (en présence des censures) tout en adaptant les estimateurs semi-paramétriques connue dans le cas normal. La performance de ces estimateurs était illustrée à travers une simulation suivie par une application dans la méthodologie sur (03) trois ensembles de données de survie disponibles dans Klein and Moeschberger (2005), concernant : le cancer du larynx, le cancer de la langue et leucémie.

Dans notre étude, on a essayé d'apporter un plus, une petite contribution, afin d'améliorer la performance et la qualité de l'estimation de l'indice extrême en présence des données censurées ; en particulier nous avons cherché à déterminer un estimateur de l'indice extrême lorsque les données sont soumises à des censures aléatoires à droite, et cela en répondant à la problématique suivante :

Comment se fait l'estimation des valeurs extrêmes, en particulier l'indice de la queue, en présence des données censurées ?

Pour bien cerner la problématique posée ci-dessus, les questions suivantes seront le noyau de ce travail :

- Où se réside l'intérêt à s'intéresser à ce type d'analyse ?
- Par quelles mesures la théorie des valeurs extrêmes (TVE- unie varié), permet la mise en place d'un système d'évaluation, d'intervention et de prévention face aux événements extrêmes? Et comment peut-on savoir le niveau de retour attendu d'être dépassé en moyenne une fois chaque T période ?
- Comment la TVE prend en considération la présence des données censurées ? et quel impact aura-t-il sur les estimateurs traditionnels de l'indice de queue extrême ?

Afin d'arriver à nos objectifs et répondre à notre problématique, nous nous sommes basés sur les hypothèses suivantes :

- Les outils probabilistes traditionnels développés dans un univers gaussien sont inadaptés à la compréhension des comportements extrêmes ;
- Vu ses soubassements théoriques fondés sur des théories mathématiques et statistiques approfondies, la TVE est la méthode adéquate pour modéliser les événements extrêmes ;
- Toutes les formules utilisées sont valables et basées sur l'hypothèse *iid* des observations.

A cet effet, nous avons essayé de proposer un nouvel estimateur de l'indice des valeurs extrêmes dans le cas des données censurées aléatoirement à droite en appliquant l'algorithme de Newton-Raphson sur une vraisemblance adaptée aux censures. Ensuite, afin d'évaluer la performance de l'estimateur proposé, nous avons fait des simulations, où nous avons étudié à la fois la précision de cet estimateur par rapport aux deux autres estimateurs en fonction de la taille de l'échantillon et le pourcentage/niveau de censure, et à la fin avec une illustration réelle sur les durées de chômage en Algérie.

2. Généralités sur la théorie des valeurs extrêmes :

La théorie des valeurs extrêmes a pour but d'étudier la loi du maximum d'une suite de variables aléatoires réelles si, et spécialement si, la loi du phénomène n'est pas connue.

Considérons $X_1, \cdots, X_n$ une suite de n variables aléatoires indépendantes et identiquement distribuées (*iid*) de fonction de répartition *F* définie par :

$$F(x) = P(X_i \leq x) \quad pour\ i = 1, \cdots, n.$$

Pour étudier le comportement extrême des évènements, on considère la variable aléatoire $M_n = max(X_1, \cdots, X_n)$ le maximum d'un échantillon de taille n.

Comme les variables aléatoires sont *iid,* alors la fonction de répartition de M_n est donnée par :

$$F_{M_n}(x) = P(M_n \leq x) = \big(F(x)\big)^n.$$

Il est à signaler que la loi d'une variable aléatoire parente X est rarement connue avec précision et, même si la loi de cette variable parente X est connue avec exactitude, la loi du terme maximum n'est pas toujours facilement calculable. Pour ces raisons, il est intéressant de considérer les comportements asymptotiques du maximum convenablement normalisé.

Historiquement, la première approche développée dans l'analyse des valeurs extrêmes pour une population donnée est celle des blocs maxima connus par la distribution des Valeurs Extrêmes Généralisées (GEVD). Cette approche est apportée aux données qui consistent en un ensemble des maximums : annuels journaliers, semestriels journaliers, trimestriels journaliers, etc, et qui regroupent trois lois des valeurs extrêmes à savoir la loi Gumbel, loi de Weibull et loi de Fréchet.

Cependant, cette approche a été critiquée dans la mesure où l'utilisation d'un seul maxima conduit à une perte d'information contenue dans les autres grandes valeurs de l'échantillon. Pour pallier ce problème, (Pickands,1975) a introduit une nouvelle approche dans l'analyse des valeurs extrêmes connue par la Distribution de Pareto Généralisée (GPD) (*Peaks Over Threshold)*, qui est apportée aux données qui dépassent un certain seuil élevé bien déterminé.

2.1 Excès au –delà d'un seuil :

La méthode des excès au-delà d'un seuil (ou Peak Over Threshold, POT) consiste à observer non pas le maximum ou les plus grandes valeurs, mais toutes les valeurs des réalisations qui excèdent un certain seuil élevé. Cette méthode initialement développée par Pickands en 1975 et étudiée par divers auteurs tels que Smith en 1987, Davison et Smith 1990 et Reiss et Thomas en 2007.

On définit un seuil u réel suffisamment élevé, $N_u = card\{i: i = 1, \cdots, n, X_i > u\}$ et $Y_i = X_i - u > 0$ pour $1 \leq i \leq N_u$ où N_u est le nombre de dépassements du seuil u pour les $(X_i)_{1 \leq i \leq n}$ et $Y_1, \cdots, Y_{N_u}$ les excès correspondants.

On définit la loi conditionnelle des excès F_u (la loi conditionnelle F_u par rapport au seuil u pour les variables aléatoires dépassant ce seuil) par :

$$F_u(y) = P(X - u \leq y / X > u) = \frac{F(u+y) - F(u)}{1 - F(u)}, \quad y \geq 0$$

Figure N° 1 : Les dépassements de X au-delà d'un certain seuil u

Source : Réalisé par les auteurs

Y_i représente l'excès de la variable X au-dessus du seuil u quand $> u$, définie par $X_i - u$. Ce qui signifie qu'on s'intéresse à la loi de probabilité de Y_i sachant que $X > u$.

Le théorème de Pickands-Balkema-de Haan ci-après donne la forme de la loi limite pour les valeurs extrêmes sous certaines conditions de convergence, la loi limite est une loi de Pareto Généralisée (GPD).

2.2 Théorème de Pickands-Balkema-de Hann :

Une fonction de répartition F appartient au domaine d'attraction maximale de G_{ξ}, si et seulement si, il existe une fonction positive $\beta(u)$ telle que :

$$\lim_{u \to x_F} Sup_{0 \le y \le x_F - u} \left| F_u(y) - G_{\xi,\beta(u)}(y) \right| = 0 \qquad (1)$$

Où $F_u(y)$ est la fonction de répartition conditionnelle des excès pour u élevé, x_F est le point terminal de F, $x_F = \{x \in \mathbb{R}: F(x) < 1\}$ et $G_{\xi,\beta}$ est la GPD pour un seuil assez élevé donnée par :

$$G_{\xi,\beta}(y) = \begin{cases} 1 - \left(1 - \xi \frac{y}{\beta}\right)^{-\frac{1}{\xi}} & , \ \xi \neq 0 \\ 1 - exp\left(-\frac{y}{\beta}\right) & , \ \xi = 0 \end{cases}$$

Où $y \geq 0$ pour $\xi \geq 0$ et $0 \leq y \leq \frac{-\beta}{\xi}$ pour $\xi < 0$.

β est le paramètre d'échelle et ξ est le paramètre de forme.

La dérivée de la distribution cumulative de GPD donne la fonction de densité de probabilité suivante :

$$g_{\xi,\beta}(y) = \begin{cases} \beta^{-1}(1 + \xi \frac{y}{\beta})^{-\frac{1}{\xi}-1}, & \xi \neq 0 \\ \beta^{-1} \exp(-\frac{y}{\beta}) & , \quad \xi = 0 \end{cases}$$

Cela signifie que la manière la plus naturelle de modéliser la fonction de distribution des excès au-delà d'un seuil suffisamment élevé est l'utilisation de la loi Pareto Généralisée qui a les propriétés suivantes :

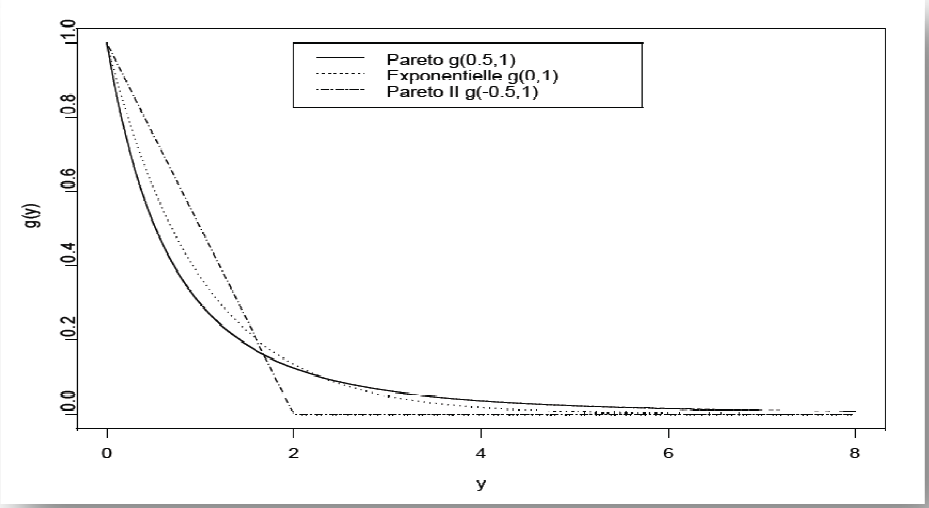
$$E(Y) = \frac{\beta}{1-\xi}, (\xi < 1) \quad \text{et} \quad V(Y) = \frac{\beta^2}{(1-\xi)^2(1-2\xi)}, (\xi < \frac{1}{2})$$

Lorsque $\xi \geq 1$, la moyenne n'est pas définie, et lorsque $\xi \geq \frac{1}{2}$ c'est la variance qui n'est pas définie.

Selon les valeurs du paramètre de forme ξ, la GPD regroupe les trois distributions suivantes :

- ✓ Lorsque $\xi > 0$, on obtient la loi de Pareto usuelle,
- ✓ Lorsque $\xi < 0$, on obtient la loi de Pareto du type II,
- ✓ Lorsque $\xi = 0$, on obtient la loi exponentielle de paramètre β.

Figure N° 2 : Densité des lois des valeurs extrêmes, avec $\xi = 0,5$ pour la loi de Pareto,$\xi = 0$ pour la loi exponentielle et $\xi = -0,5$ pour la loi Pareto II



Source : BECHIR RAGGAD (2009), fondements de la théorie des valeurs extrêmes, ses principales applications et son apport à la gestion des risques du marché pétrolier ,Mathematics and Social Sciences (47e année, n° 186), p. 39.

Néanmoins, dans la pratique, l'application de cette approche (les excès) n'est pas facile, puisqu'il n'existe pas une méthode exacte pour déterminer le seuil de référence, c'est-à-dire le seuil à partir duquel on considère une observation comme une valeur extrême (différents seuils conduisent à différents résultats).

Pour un seuil élevé, le comportement des estimateurs extrêmes est connu pour être régi par un paramètre essentiel de la distribution qui est le paramètre de forme appelé aussi l'indice des Valeurs Extrêmes. Ce paramètre est très important puisqu'il mesure la lourdeur de la queue. Dans la littérature, l'estimation de cet indice a été largement étudiée par plusieurs auteurs à savoir (Hill, 1975), (Smith,1987), (Dekkers et al.,1989) et (Drees et al.,2006).

3. Estimation de l'IVE sans censure :

Pour la majorité des fonctions de répartition F la loi asymptotique du maximum $X_{n:n}$ est une loi des valeurs extrêmes qui étant indexée par le paramètre de queue ξ, ce paramètre apporte une information sur la forme de la queue de distribution de F.

Dans la littérature de la TVE, il existe de nombreux auteurs qui se sont intéressés à l'estimation de l'indice des valeurs extrêmes ξ et des quantiles extrêmes. Dans ce qui suit, nous exposerons uniquement trois estimateurs de ξ.

3.1 Estimateur de Pickands :

Cet estimateur a été introduit en 1975 par James Pickands, pour toute $\xi \in \mathbb{R}$.

Définition : (Estimateurs de Pickands)

Soient $X_1, X_2, \cdots, X_n$ n variables aléatoires indépendantes et identiquement distribuées de fonction de répartition $F \in D(H_\xi)$, où $\xi \in \mathbb{R}$. Soit $k = k_n$ une suite d'entiers avec $1 < k < n$, l'estimateur de Pickands est défini par :

$$\hat{\xi}^P = \hat{\xi}^P(k) = \frac{1}{\log 2}\log\left(\frac{X_{n-k+1:n} - X_{n-2k+1:n}}{X_{n-2k+1:n} - X_{n-4k+1:n}}\right)$$

L'auteur a démontré la consistance faible de son estimateur. La convergence forte ainsi que la normalité asymptotique ont été démontrées par Dekkers et de Haan .

3.2 Estimateur de Hill :

Cet estimateur a été introduit par Hill en 1975 afin d'estimer le paramètre de queue des distributions appartenant au domaine d'attraction de Fréchet $D\left(\Phi_{\frac{1}{\xi}}\right)$, c'est-à-dire, quand la queue de distribution a une forme de Pareto.

Définition : (Estimateur de Hill)

Soient $X_1, X_2, \cdots, X_n$ n variables aléatoires indépendantes et identiquement distribuées de fonction de répartition $F \in D\left(\Phi_{\frac{1}{\xi}}\right)$, où $\xi < 0$. Soit $k = k_n$ une suite d'entiers avec $1 < k < n$, l'estimateur de Hill est défini par

$$\hat{\xi}^H = \hat{\xi}^H(k) = \frac{1}{k}\sum_{i=1}^{k} \log X_{n+i-1:n} - \log X_{n-k:n}$$

Un grand nombre de travaux théoriques ont été consacrés à l'étude des propriétés de l'estimateur de Hill. La consistance faible a été établie par (Mason, 1982), la consistance forte fut établie par (Deheuvels et al.,1988) et plus récemment par (Necir, 2006) . La normalité asymptotique est due entre autres à (Davis et Resnick ,1984) , (Csargö et Mason, 1985) et (Häusler et Teugels, 1985) .

3.3 Estimateur des Moments :

Un inconvénient de l'estimateur de Hill est qu'il est conçu seulement pour l'IVE des distributions à queues lourdes. En 1989, Dekkers et al. ont proposé une extension de tous types de distribution, appelé estimateur des moments.

Définition : (Estimateur des Moments)

Pour $\xi \in \mathbb{R}$, l'estimateur des moments est

$$\hat{\xi}^M = \hat{\xi}^M(k) = {M_n}^{(1)} + T_n = {M_n}^{(1)} + 1 - \frac{1}{2}\left(1 - \frac{\left({M_n}^{(1)}\right)^2}{{M_n}^{(2)}}\right)^{-1}$$

Avec :

$${M_n}^{(r)} = {M_n}^{(r)}(k) = \frac{1}{k}\sum_{i=1}^{k}(\log X_{n-i+1:n} - \log X_{n-k:n})^r, \quad r = 1,2$$

Où ${M_n}^{(1)}$ est l'estimateur de Hill $\hat{\xi}^H$.

Les propriétés asymptotiques de cet estimateur ont été étudiées par (Dekkers et al., 1989).

4. Estimation de l'IVE avec censure :

On s'intéresse dans cette section à l'estimation de l'IVE en présence de données censurées aléatoirement à droite. Ce problème est très récent dans la littérature, les premiers qui ont mentionné le sujet sont (Beirlant et al.,1996) et (Reiss et Thomas,2007), mais sans résultats asymptotiques. Certains estimateurs des paramètres de la queue ont été proposés par (Beirlant et Guillou, 2001) pour les données tronquées et étendues à la censure aléatoire par (Beirlant et al.,2007) et l'année suivante par (Einmahl et al., 2008) .

En réalité, l'estimation des valeurs extrêmes en présence de données censurées aléatoirement à droite revient à dire que l'échantillon $X_1, X_2, \cdots, X_n$ (les durées réelles de vie) n'est pas observé, mais qu'il est censuré par un deuxième échantillon $Y_1, Y_2, \cdots, Y_n$, qui est supposé être indépendant du premier, où les X_i et Y_i sont des variables aléatoires iid de lois F et G respectivement. Autrement dit, les variables que nous observons sont issues d'une part des variables aléatoires Z_i définies par $Z_i = min(X_i, Y_i),\ i = 1,2, \cdots, n$, et d'autre part des indicateurs de censure $\delta_i = \mathbb{I}_{\{X_i \leq Y_i\},}\ i = 1, \cdots, n.$

Toutefois, il convient de signaler que, les différents estimateurs proposés de l'indice des valeurs extrêmes en prenant en considération la présence des censures ont été tous construits de la même manière. Leurs estimateurs sont basés sur un estimateur standard de l'indice de queue (non adapter à la censure) divisé par la proportion de données non censurées dans les plus grands k variables aléatoires $Z_1, Z_2, \cdots, Z_n$.

$$\hat{\xi}_1^{(.,c)} = \hat{\xi}_1^{(.,c)}(k) = \frac{\hat{\xi}^{\cdot}}{\hat{p}},$$

Où $\hat{p} = \hat{p}(k) = \frac{1}{k}\sum_{i=1}^{k} \delta_{[n-i+1:n]}$, $\delta_{[j:n]}$ est le concomitant de la j-ème statistique d'ordre, c'est-à-dire, $\delta_{[j:n]} = \delta_i$ si $Z_{j:n} = Z_i, 1 \leq i \leq n$.

$\hat{\xi}^{\cdot}$ peut être n'importe quel estimateur non adapté à la censure, en particulier $\hat{\xi}^H, \hat{\xi}^M, \cdots$, et $\hat{p}$ l'estimateur de la proportion des données observées dans la queue à droite de distribution avec $k = k_n$ (k_n est une suite d'entiers liée à la taille de l'échantillon n tel que $\lim_{n\to\infty} k_n = \infty$ et $\lim_{n\to\infty} {k_n}/{n} = 0$) .

(Beirlant et al., 2007) sont les premiers qui ont introduit cette méthodologie dans le cas d'estimateurs de Hill et de moment. De plus, ils ont proposé les estimateurs des quantiles extrêmes et ont discuté leurs propriétés asymptotiques lorsque les données sont censurées pour un seuil déterministe. (Einmahl et al. ,2008) ont adapté différents estimateurs de l'IVE au cas où les données sont censurées par un seuil aléatoire et ils ont proposé une méthode unifiée pour établir leur normalité asymptotique.

Cependant, les propriétés asymptotiques et la normalité de la plupart des estimateurs proposés restent difficiles à obtenir.

Vu les solutions proposées concernant le problème de l'estimation de l'indice des valeurs extrêmes en présence des données censurées aléatoirement à droite, nous allons présenter dans ce qui suit un apport concernant l'estimateur non paramétrique de cet indice en appliquant l'algorithme de Newton-Raphson.

❖ **L'écriture de la vraisemblance dans les modèles de durée :**

Supposons que nous disposons d'un échantillon de taille *N* des durées observées (complètes et/ou censurées) $t_1, \cdots, t_N$. Cela revient à disposer, en plus de la valeur de t_i, d'une variable indicatrice de censure C_i telle que $C_i = 1$ si la durée t_i est complète, et 0 sinon (censurée). Les observations sont donc des réalisations de $T_i = min(Z_i, C_i)$ et de la variable indicatrice de non censure $\delta_i = \mathbb{I}[Z_i < C_i]$. Les n variables aléatoires observées T_i et C_i sont i.i.d. et les C_i et Z_i sont supposées indépendants entre elles.
La vraisemblance du modèle s'ecrit:

$$L = \prod_{i=1}^{N} f(t_i)^{\delta_i} \; S(t_i)^{1-\delta_i}$$

Où $S(t_i)$ = 1- $F(t_i)$.

En effet, la probabilité qu'une durée soit censurée en t_i, donc supérieure ou égale à t_i est la valeur de la survie $S(t_i)$.

La log-vraisemblance a donc pour forme :

$$log\, L = \sum_{i=1}^{N} \delta_i \; \log \; f\,(t_i) + \sum_{i=1}^{N} (1 - \delta_i) \log \; S\,(t_i)$$

Afin de simplifier l'expression ci-dessus, on utilise la relation du taux de hasard :

$$\lambda(t_i) = f\,(t_i) \,/\, S\,(t_i)$$

Ce qui donne :

$$log\, L = \sum_{i=1}^{N} \delta_i \; \log \; \lambda\,(t_i) + \sum_{i=1}^{N} \log \; S\,(t_i).$$

Lorsqu'on spécifie une forme particulière pour λ et *S*, on obtient la valeur de la fonction à maximiser en calculant log $\lambda\,(t_i)$ et log $S\,(t_i)$.

Remarque :

Sous l'hypothèse de censure non informative, on remarque qu'il est équivalent de chercher l'estimateur du maximum de vraisemblance de θ en maximisant l'expression :

$$L = \prod_{i=1}^{N} f(t_i).$$

Dans le cas d'un modèle des valeurs extrêmes basé sur l'approche POT, la vraisemblance s'écrit, en tenant compte des censures, comme suit :

$$L = \prod_{i=1}^{K} [\, f_{GPD}(E_i)]^{\delta_i} \;\; [1 - F_{GPD}(E_i)]^{1-\delta_i}$$

Avec $E_i = Z_j - u \quad si \quad Z_j > u$, u est le seuil $1 - F_{GPD}(E_i) = \left(1 + \frac{\xi}{\sigma} E_i\right)^{-\frac{1}{\gamma}}$ et la densité associée f_{GPD} est égale $\frac{1}{\sigma}\left(1 + \frac{\xi}{\sigma} E_i\right)^{-\frac{1+\xi}{\xi}}$.

Ainsi, on obtient la vraisemblance suivante :

$$L(\xi, \sigma) = \prod_{i=1}^{K} [\frac{1}{\sigma}\left(1 + \frac{\xi}{\sigma} E_i\right)^{-\frac{1+\xi}{\xi}}]^{\delta_i} \; [\left(1 + \frac{\xi}{\sigma} E_i\right)^{-\frac{1}{\xi}}]^{1-\delta_i}$$

$$log\, L(\xi, \sigma) = \sum_{i=1}^{K} \delta_i \left[\log \frac{1}{\sigma} - \left(\frac{1}{\xi} + 1\right) \log\left(1 + \frac{\xi}{\sigma} E_i\right)\right] - \sum_{i=1}^{K} \frac{1}{\xi} (1 - \delta_i) \log\left(1 + \frac{\xi}{\sigma} E_i\right)]$$

En dérivant cette vraisemblance par rapport à ses deux paramètres, nous obtenons le système suivant de deux équations à deux inconnues :

$$\begin{cases} L_1^{'} = \dfrac{\partial log L(\xi, \sigma)}{\partial \xi} = \dfrac{1}{\xi^2} \displaystyle\sum_{i=1}^{K} \log\left(1 + \frac{\xi}{\sigma} E_i\right) - \frac{1}{\xi} \sum_{i=1}^{K} \left(\frac{1}{\xi} + \delta_i\right) \frac{\xi E_i / \sigma}{1 + \frac{\xi}{\sigma} E_i} \\ L_2^{'} = \dfrac{\partial log L(\xi, \sigma)}{\partial \sigma} = -\dfrac{1}{\sigma} \displaystyle\sum_{i=1}^{K} \delta_i + \frac{1}{\sigma} \sum_{i=1}^{K} \left(\frac{1}{\xi} + \delta_i\right) \frac{\xi E_i / \sigma}{1 + \frac{\xi}{\sigma} E_i} \end{cases}$$

Dans le cas où $\xi = 0$, ces dérivées s'écrivent comme suit :

$$\begin{cases} L_1^{'}(0, \sigma) = \dfrac{\partial log L(0, \sigma)}{\partial \gamma} = -\dfrac{1}{2} \displaystyle\sum_{i=1}^{K} \frac{{E_i}^2}{\sigma^2} - \sum_{i=1}^{K} \delta_i \frac{E_i}{\sigma} \\ L_2^{'}(0, \sigma) = \dfrac{\partial log L(0, \sigma)}{\partial \sigma} = -\dfrac{1}{\sigma} \displaystyle\sum_{i=1}^{K} \delta_i + \frac{1}{\sigma^2} \sum_{i=1}^{K} E_i \end{cases}$$

Afin de trouver des solutions à ce système non linéaire, on a fait recours à une simulation en utilisant le logiciel MATLAB.

5. Comparaison de la performance des différents estimateurs de l'indice des valeurs extrêmes :

Dans cette section, nous comparons la performance des estimateurs de l'indice des valeurs extrêmes de Hill (H), des Moments (M) et celui que nous avons proposé dans le cas des données censurées aléatoirement à droite (NR) en termes de convergence pour une loi donnée. A cet effet, nous avons simulé 100 échantillons de taille $N = 10000$, issues d'une distribution de Pareto de paramètre $(1,\ 0.5)$, pour deux niveaux de censure 1% et 3%.

Pour chaque niveau de censure, nous avons comparé la moyenne empirique de ces trois estimateurs pour les n échantillons en fonction de la taille de l'échantillon K/N , avec la vraie valeur du paramètre que nous avons généré à travers notre échantillon, pour faire sortir l'erreur d'estimation équivalente pour chaque K/N comme le montre les graphiques ci-après.

❖ **Comparaison graphique des estimateurs :**

Le comportement en termes de précision des différents estimateurs présentés et illustrés dans les graphiques suivants qui mettent en évidence l'erreur d'estimation de trois estimateurs en fonction de la taille d'échantillon pour deux niveaux de censure 1% et 3%.

Figure N° 3 : Comparaison de l'erreur d'estimation en fonction de la méthode (1%et 3% respectivement)

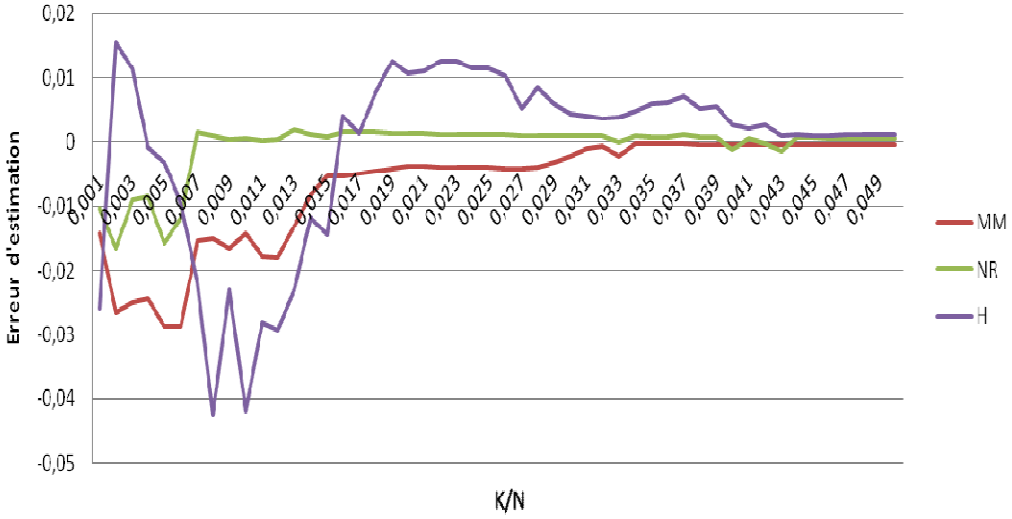
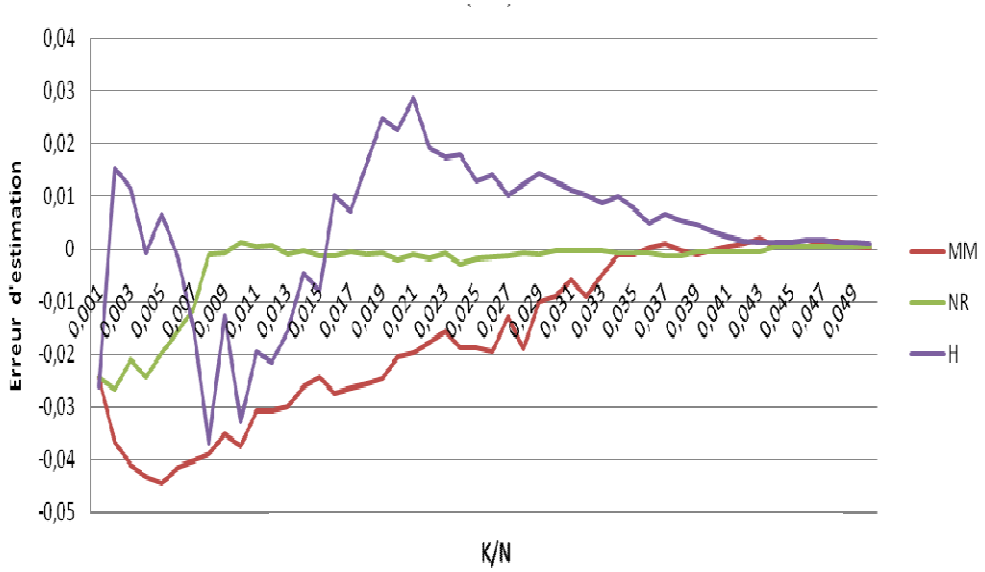


Figure N° 3 : Comparaison de l'erreur d'estimation en fonction de la méthode (1% respectivement)



Source : Réalisé par les auteurs

D'après les graphes précédents, nous remarquons que :

- ✓ L'estimateur de Hill est relativement plus volatil et moins efficace par rapport aux deux autres estimateurs notamment pour $k < 0{,}02 \times n$;

- ✓ L'estimateur des Moments et l'estimateur proposé NR se comportent presque de la même manière. Ils démarrent avec un écart significatif, mais ils s'approchent au fur et à mesure avec l'augmentation de la taille de l'échantillon ;
- ✓ Toutefois, il y a lieu de noter que les valeurs de l'estimateur de ces deux derniers sont un peu plus faibles (proche du zéro) que celle du Hill qui varie entre $-0{,}0425$ et $0{,}0154$ pour un niveau de censure qui ne dépasse pas le 1%, et entre $-0{,}0368$ et $0{,}0286$ pour un niveau de censure de 3%. Pour l'estimateur des Moments, il varie entre $-0{,}0257$ et $0{,}0003$ pour un niveau de censure de 1% et entre $-0{,}0446$ et $0{,}0019$ pour un niveau de censure de 3% . Quant à l'estimateur proposé NR, il varie entre $-0{,}0166$ et $0{,}0020$ pour un niveau de censure de 1%, et entre $-0{,}0266$ et $0{,}0013$ pour un niveau de censure de 3%.

Ainsi, nous avons pu constater, d'une part qu'au fur et à mesure qu'on diminue le niveau de censure, les courbes se rapprochent plus ; et d'autre part, les courbes deviennent plus rapidement linéaires et alignées avec le 0, ce qui veut dire que l'estimateur converge plus rapidement vers la vraie valeur. De plus, on voit clairement la suprématie de l'estimateur proposé de NR sur ceux de Hill et des Moments.

6. Application sur des données réelles :

Comme illustration numérique, nous avons proposé une application sur les durées de chômage en Algérie. Un domaine relativement nouveau pour ce type d'application, vu que l'ensemble des illustrations sur des données réelles faites dans cette théorie pour estimer les valeurs extrêmes en présence des censures concerne la médecine, la finance notamment les assurances et les phénomènes météorologiques.

Le choix de ce domaine d'application a l'avantage de nous permettre d'une part d'avoir une idée sur les durées de chômage maximales en Algérie ; et d'autre part de connaitre si vraiment les dispositifs déployés par l'État ont aidé à réduire cette longueur ou bien au contraire, si les demandeurs d'emploi de longue durée sont, toujours, perçus par les recruteurs comme manque de motivation, une raison principale pour laquelle leurs candidatures intéressent rarement les employeurs ; ils sont considérés comme moins intelligents, moins au courant des nouvelles technologies et plus difficiles à former, et par conséquent leurs demandes sont souvent rejetées.

À cet effet, notre application a nécessité une longue série de données pour établir cette modélisation, d'où le besoin d'un échantillon de taille assez importante. Dans notre cas, nous avons récolté une base de données exhaustive des **primo** demandeur , couvrant les 48 wilayas de l'Algérie, fournies par la Direction générale de l'Agence Nationale de l'Emploi (DG/ANEM) pour la période allant de **janvier 2012 à décembre 2016** avec **32 186** observations (durées) contenant les informations suivantes : l'immatriculation, la date de la première inscription dans le dispositif DAIP, la date de sortie (pour ceux qui ont trouvé un emploi), sexe,⋯ .

Ce choix est justifié par l'importance, la différence et l'hétérogénéité qui peut se trouver entre les différentes wilayas de l'Algérie, qui est due principalement à la spécificité de chaque région et de son offre d'emploi.

6.1. Estimation des paramètres du modèle GPD :

L'estimation des paramètres du modèle GPD à l'aide du logiciel Matlab donne les résultats suivants :

Tableau N° 1 : Estimation du modèle GPD

Séries/ Paramètres	Durées de chômage	
	Estimation	**Std. Err.**
σ_0	389.001	/
ε_0	0.5528547	/
σ	464.3017555	8.691059e-02
ε	0.8116431	2.001803e-06
Log vraisemblance	1799.61205359736	

Source : Réalisé par les auteurs

On remarque que le paramètre de forme ξ est positif ainsi que son intervalle de confiance ce qui signifie que les excédents de notre série suivent une loi de Pareto usuelle.

6.2. Le niveau de retour (NR) :

Après avoir ajusté nos données avec la distribution adéquate qui les représentent le mieux, nous allons établir le niveau de retour qui susceptible d'être dépassé au moins une fois pendant une période donnée, étant donné une année, 2 ans, voire même 3 ans. Nous nous intéressons à connaître la grandeur qui sera probablement excédée au moins une fois durant cet intervalle de temps futur.

Les quantiles estimés pour trois périodes de retour ainsi que leurs intervalles de confiance à 95% trouvés en utilisant le logiciel GenStatV.7 sont présentés dans le tableau suivant:

Tableau N° 2 : Les niveaux de retour estimés pour les trois périodes (1an, 2 ans, 3 ans) et leurs intervalles de confiance

Période de Retour (PR)	**NR**	**IC**
T= 1 année	1600.7806	[1600.780640 ; 1600.780642]
T= 2 ans	1695.9863	[1695.98633 ; 1695.98633]
T= 3ans	1731.3381	[1731.07317 ; 1731.60294]

Source : Réalisé par les auteurs

L'analyse du tableau nous montre l'existence d'une forte relation (relation croissante) entre les périodes de retour et les niveaux de retour associés à notre modèle. En effet, à chaque fois que la période de retour augmente, le niveau de retour associé subira une augmentation aussi ainsi qu'un élargissement au niveau de son intervalle de confiance.

Toutefois, il y a lieu de noter également que lorsque nous augmentons les périodes de retour nous perdons avec elles l'information sur les niveaux de retour.

De plus, d'après les quantiles estimés, à partir des paramètres calculés précédemment, pour les trois périodes de retour, nous remarquons que notre série prévoit des niveaux (durées de chômage) assez importants, citons comme exemple la période de retour de 1 an, où nous devons attendre environ 1 année en moyenne pour avoir une durée de chômage maximale de 1601 jours (l'équivalant de 4 années et 4 mois) avant de trouver un emploi, ou encore 2 ans en moyenne pour observer un niveau de retour de 1696 jours (l'équivalant de 4 années et 8 mois).

7. Conclusion :

À travers notre étude, nous avons tenté de proposer un nouvel estimateur de l'indice des valeurs extrêmes dans le cas des données censurées aléatoirement à droite en appliquant l'algorithme de Newton-Raphson sur une vraisemblance adaptée aux censures. Nous avons pu constater que l'estimateur de Hill est relativement plus volatil et moins efficace par rapport aux estimateurs des Moments et celui que nous avons proposé. Quant à l'estimateur des Moments et l'estimateur proposé, ils se comportent presque de la même manière.

Néanmoins, le domaine des valeurs extrêmes et de survie est un domaine très vaste et actuellement un champ de recherche en plein essor, notamment ce nouvel axe de recherche des valeurs extrêmes en présence des données censurées, qui reste relativement nouveau et intéressent avec de multiples applications, ce qui nous a amené à proposer les recommandations suivantes :

- ✓ D'un point de vue pratique, il serait intéressant de pouvoir inclure dans le modèle des covariables, en se basant sur la théorie des valeurs extrêmes bivariées (les coupoles) ou multivariées. Il est légitime de penser qu'il existe pas mal de covariables qui interviennent et qui impactent généralement le phénomène étudié. Il serait donc très utile d'en tenir compte, afin de préserver le contexte général de l'événement ;
- ✓ Le modèle que nous avons développé peut être étendu au cadre spatial. En effet, en sciences du climat par exemple, nous disposons souvent de mesures localisées en plusieurs endroits et il serait donc fort intéressant de pouvoir prendre en compte l'aspect spatial dans l'analyse de nos données ;
- ✓ S'intéresser aux nouveautés de ce domaine. En effet, la théorie probabiliste traditionnelle des valeurs extrêmes concerne la distribution asymptotique des maxima (des minima) d'une suite de variables aléatoires indépendantes et identiquement distribuées. Un des récents développements de cette théorie consiste à lever l'hypothèse d'indépendance des observations, certains problèmes restent ouverts et seraient intéressants de les aborder.

8. Liste Bibliographique:

Beirlant, J., Guillou, A., Dierckx, G.,& Fils-Villetard, A. (2007). Estimation of the extreme value and extreme quantiles under random censoring. *Extremes, 10*(3), 151-174.

Beirlant, J.,& Guillou,A. (2001). Pareto Index Estimation Under Moderate Right Censoring. *Scandinavian Actuarial Journal*, 111-125.

Beirlant,J., Teugels, J., & Vynckier,P. (1996). *Practical analysis of extreme values.*

Csörgö, S.,& Mason, D.M. . (1985). Central limit theorems for sums of extreme values. *Mathematical Proceedings of the Cambridge Philosophical Society, 98*(3), 547-558.

Davis, R.,& Resnick, S. (1984). Tail estimates motivated by extreme value theory.*The Annals of Statistics, 12*(4), 1467-1487.

Deheuvels, P., Häeusler, E.,& Mason, D.M. (1988). Almost sure convergence of the Hill estimator. *Mathematical Proceedings of the Cambridge Philosophical Society, 104*(02), 371-381.

Dekkers, A.L., Einmahl, J.H., & De Haan, L. (1989). A Moment estimator for the index of an extreme value distribution. *The Annals of Statistics, 17*(4), 1833-1855.

Dekkers, A.L.,& De Haan, L. (1989). On the estimation of the extreme-value index and large quantile estimation. *The Annals of Statistics, 17*(4), 1795-1832.

Drees, H., De Haan, L.,& Li D. (2006). Approximations to the tail empirical distribution function with application to testing extreme value conditions, 136(10),. *Journal of Statistical Planning and Inference, 136*(10), 3498-3538.

Einmahl, J.H., Fils-Villetard,A.,& Guillou, A. (2008). Statistics of extremes under random censoring. *Bernoulli, 14*(1), 207-227.

Embrechts, P., Klüppelberg, C.,& Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance.* Berlin: Spinger-Verlag.

Häeusler, E.,& Teugels, J.L. (1985). On asymptotic normality of Hill's estimator for the exponent of regular variation. *The Annals of Statistics, 13*(2), 743-756.

Hill, B. (1975). A Simple general approach to inference about the tail of a distribution. *The Annals of Statistics, 3*(5), 1163-1174.

Mason, D. (1982). Laws of large numbers for sums of extreme values. *The Annals of Probability, 10*(3), 754-764.

Necir, A. (2006). A Functional law of the iterated algorithm for Kernel-type estimators of the tail index. *Journal of Statistical Planning and Inference, 136*(3), 780-802.

Pickands, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics, 3*(1), 119-131.

Raggad, B. (2009). fondements de la théorie des valeurs extrêmes, ses principales applications et son apport a la gestion des risques du marché pétrolier. *Mathematics and Social Sciences*, p. 39.

Reiss, R.D.,& Thomas, M. (2007). *Statistical Analysis of Extreme Values from Insurance, Finance, Hydrology and other Fields.* Basel: Birkhäuser.

Smith, R. L. (1987). Estimating Tails of Probability Distributions. *The Annals of Statistics, 15*(3), 1174-1207.

9. Annexes :

Validation du modèle choisi : « Goodness of Fit »

Hypothèse des tests :

$$\begin{cases} H0 : \text{l'échantillon suit une loi de Pareto} \\ H1 : \text{l'échantillon ne suit pas une loi de Pareto} \end{cases}$$

Tableau N° 3 : Récapitulatif des résultats des tests non paramétriques

Gen. Pareto [#24]					
Kolmogorov-Smirnov					
Taille de l'échantillon Statistique Valeur de P Rang	285 0,05246 0,39928 4				
α	0,2	0,1	0,05	0,02	0,01
Valeur critique	0,06356	0,07244	0,08044	0,08992	0,09649
Rejeter?	Non	Non	Non	Non	Non
Anderson-Darling					
Taille de l'échantillon Statistique Rang	285 1,101 2				
α	0,2	0,1	0,05	0,02	0,01
Valeur critique	1,3749	1,9286	2,5018	3,2892	3,9074
Rejeter?	Non	Non	Non	Non	Non
Khi-Carré					
Degrés de liberté Statistique Valeur de P Rang	8 4,8715 0,77122 3				
α	0,2	0,1	0,05	0,02	0,01
Valeur critique	11,03	13,362	15,507	18,168	20,09
Rejeter?	Non	Non	Non	Non	Non

Source : élaboré par les auteurs à l'aide du logiciel Easy Fit 5.5