EVALUATION DES RISQUES D'EXPOSITION À UN CONTAMINANT ALIMENTAIRE : QUELQUES OUTILS STATISTIQUES.

PATRICE BERTAIL, LABORATOIRE DE STATISTIQUE, CREST

ABSTRACT. [Résumé] L'objet de cette étude est de présenter et de développer dans le cadre de l'évaluation des risques alimentaires un certain nombre de méthodes probabilistes et/ou statistiques permettant dans une perspective statique, d'une part d'évaluer les risques liés à l'exposition à certains contaminants et d'autre part de caractériser les populations exposées à ces risques. Nous montrons dans ce travail comment il est possible de modéliser l'exposition à certains contaminants alimentaires grâce à la théorie des valeurs extrêmes. La probabilité de dépasser un certain seuil de toxicité est modélisée par une loi de probabilité de type Pareto, dont l'index s'interprète comme un indice de risque. Nous discutons brièvement les problèmes liés à l'estimation de ce paramètre. Afin de caractériser les populations dites à risque, l'indice des lois de Pareto n'est pas fixé, mais est lui-même une fonction de variables explicatives. Cette approche permet de tester et de mettre en évidence les caractères ayant un impact sur les risques d'exposition. Ces propos sont illustrés par la mise en oeuvre de ces méthodes sur l'évaluation des risques liés à la contamination par le mercure dans l'alimentation.

[Abstract] The purpose of this paper is to introduce (to non statistician) and develop some statistical methods for 1) evaluating risk to exposure of some contaminants present in food 2) caracterizing risky populations. We show how it is possible to model high exposure thanks to Pareto distribution, which index may be interpreted as as a risk index. We briefly discuss the estimation problem. To caracterize the risky populations, we allow the index to depend on some socio-economical covariates through some specified link function. We illustrate the proposed methods on risk assessment to exposure to mercury in France.

Date: Preliminary version : September 2001. This version: Aout 2002 .

1991 *Mathematics Subject Classification.* Primary 05C38, 15A15; Secondary 05A15, 15A18.

Key words and phrases. Valeurs extrêmes, loi de Pareto généralisée, intervalles de confiance, index de Pareto, population à risque, risque alimentaire.

Mes remerciements à tous les participants du séminaire : "Approches économiques et statistiques du risque : applications à l'environnement et la consommation alimentaire" et en particulier à J.C Bureau et S. Marette, mes coorganisateurs. Ce projet a été en partie financé par l'AIP 2000 INRA "Comportement Alimentaire" coordonnée par Ph. Verger.

1. Introduction

L'objet de cette étude est de présenter et de développer dans le cadre de l'évaluation des risques alimentaires un certain nombre de méthodes probabilistes et/ou statistiques permettant dans une perspective statique, d'une part d'évaluer les risques liés à l'exposition à certains contaminants et d'autre part de caractériser les populations exposées à ces risques. Certains aliments contiennent en effet dans des teneurs plus ou moins grandes des contaminants, qui par effets d'accumulation peuvent créer des problèmes de santé. Pour citer quelques exemples, le mercure, le plomb, les dioxines, les micotoxines etc...présents dans de nombreux aliments peuvent avoir une influence néfaste sur la santé. Curieusement si les dangers biologiques, physicochimiques de ces produits sont bien connus grâce aux progrès de la chimie analytique, il n'existe que très peu de travaux statistiques sur l'évaluation globale des risques d'exposition des individus (cf. Gauchi (2000)). L'absence quasi total d'évaluation quantitative fiable ne fait qu'entretenir la confusion entre risque réel et risque perçu. Pour des raisons de santé publique, de répercution psychologique de l'information ou encore pour des raisons économiques ou politiques (faut-il interdire tel produit, empêcher les importations de tel autre), raisons qui posent la question de définitions de normes, il est important de quantifier cette notion de risque pour en apprécier ou en relativiser l'importance (voir à ce propos le Pogramme Cadre de la Communauté Européenne: 6ème PCRD).

D'un point de vue statistique, on peut s'intéresser à l'estimation de la probabilité de dépasser dans une certaine population un certain seuil de toxicité (seuil issu d'études médicales au delà duquel il peut y avoir danger pour la santé). En statistique classique, lorsqu'on dispose d'observations (l'échantillon) indépendantes ou non, on s'intéresse le plus souvent aux phénomènes de tendances centrales, la moyenne, la médiane voire le mode. Les théorèmes asymptotiques (lois des grands nombres, théorèmes centraux limites) assurent alors que les estimateurs usuels donnent une bonne idée du phénomène dans sa globalité. Cependant dès que l'on cherche à modéliser la notion de risque pour laquelle les phénomènes extrêmes sont fondamentaux, le problème se trouve déplacé : ce sont alors les grandes et les petites valeurs, i.e. les observations éloignées de la tendance centrale, qui deviennent les plus intéressantes. Ainsi l'exposition globale d'un individu à un contaminant donné est fonction de son panier de consommation et il est clair que ce sont les populations fortement consommatrices des produits les plus exposés qui sont les plus touchées et qui vont nous permettre de quantifier le risque d'exposition. D'autres problèmes sont également intéressants. On peut s'interroger sur les caractéristiques de ces populations, que nous appellerons population à risque, notamment lorsqu'on dispose d'informations socio-économiques sur les individus concernés. Dans la perspective de fixation de normes, en supposant que plusieurs produits comportent un risque sanitaire, on peut aussi essayer d'estimer les normes relatives devant être imposées sur chaque produit pour que seulement une tranche (petite) de la population soit exposée. Les méthodes statistiques adaptées à ces types de problèmes relèvent essentiellement de la théorie des valeurs extrêmes, théorie bien connue des statisticiens ou économètres de la finance et de l'assurance (voir par exemple les ouvrages de Resnik(1987) ou Embrechts, Klüppelberg, Mikosch (1999)) mais nécessitent un certain nombre d'adaptations dans le cadre de la consommation alimentaire, un des problèmes majeurs restant entre autres la caractérisation des populations à risque lorsqu'on dispose de covariables.

Dans une première partie, nous rappelons brièvement quelques éléments théoriques essentiels de la théorie des valeurs extrêmes en insistant plus particulièrement sur leur interprétation en termes de risque sanitaire. L'introduction de modèles de type Pareto pour modéliser les queues de distribution de la contamination globale permet de quantifier la notion de risque d'exposition et de calculer la probabilité de dépasser un certain seuil de toxicité. L'indice de Pareto, intervenant dans ces mo- dèles, s'interprète en effet comme un indice de risque. Nous évoquerons rapidement les méthodes d'estimation concurrentes de ce paramètre. Nous illustrons ces résultats par l'étude de la contamination par le mercure, contaminant présent dans très peu de produits consommés, essentiellement les poissons et fruits de mer, ce qui rend les résultats obtenus au moyen des techniques présentées plus facilement interprétables et permet de mieux saisir l'intérêt de ces méthodes.

Dans une seconde partie, nous développons des modèles simples pour caractériser les populations à risque. Ces modèles reposent essentiellement sur la modélisation paramétrique de la queue de la distribution sous la forme d'une distribution de type Pareto dans laquelle l'indice de Pareto est supposé dépendre de covariables socio-économiques. La forme de la dépendance pose un certain nombre de problèmes statistiques liés aux phénomènes d'agrégation, qui seront discutés en annexe.

2. Valeurs extrêmes et estimation de l'index de Pareto

2.1. **Valeurs extrêmes**

. L'ensemble des résultats exposés ici sont bien connus des statisticiens et on pourra se référer par exemple aux ouvrages de Embrechts, Klüppelberg et Mikosch (1999) ou de Reiss et Thomas (2001). Bien qu'elle soit de plus en plus utilisée dans les sciences environnementales, ce type d'analyse est peu, voire pas du tout utilisée en toxicologie alors que ces techniques peuvent sans doute aider à l'étude quantitative des risques : elles permettent par exemple de simuler l'impact du choix de normes de seuil de toxicité sur le risque global encouru par les populations considérées. L'objet de ce paragraphe est donc de rappeler et de donner les résultats essentiels de cette théorie sans entrer dans des détails techniques. Nous essaierons de donner une interprétation simple des quantités introduites en termes de risque sanitaire.

Pour étudier le comportement des grandes valeurs de l'échantillon, il est important de comprendre le comportement asymptotique du maximum. Les résultats de ce paragraphe nous permettront de justifier le choix de certaines formes fonctionnelles qui seront faites ensuite dans la modélisation du risque de contamination. Dans toute cette partie, on suppose que l'on dispose d'observations $X_1, X_2, ...X_n$ indépendantes de même fonction de répartition $F(x) = Prob(X < x)$. On notera dans la suite respectivement l'inverse généralisée de F

$$F^{-1}(x) = \inf(y \in R,\ F(y) \geq x).$$

Le point terminal de F (i.e. la plus grande valeur possible pour X_i pouvant prendre la valeur $+\infty$)est donné par

$$s(F) = \sup(x, F(x) < 1)$$

et la fonction de survie par

$$\overline{F}(x) = \Pr(X > x) = 1 - F(x)$$

Ainsi pour $\delta \in]0, 1[$, on note $x_\delta = F^{-1}(\delta)$ le quantile d'ordre δ de la distribution.

En terme de risque alimentaire, les X_i représenteront dans la suite le niveau d'exposition global de chaque individu i à un certain contaminant (due à l'alimentation). Pour illustrer notre propos nous considérerons dans le suite le cas du mercure, métal lourd, présent dans peu d'aliments essentiellement les produits de la mer. Si l'on connaît par exemple un niveau d_0 au delà duquel ce contaminant peut être dangereux, appelé dans la suite seuil de toxicité, $\overline{F}(d_o)$ représente donc la "proportion" de personnes exposées à un risque sanitaire dans la population. Dans le cas du mercure, l'Organisation Mondiale de la Santé (OMS) fixe la dose journalière admissible (DJA) en mercure à 0.71 $\mu g\ /jour$ et par kilo de poids corporel ce qui correspond à un seuil de toxicité de l'ordre de $18,14mg$ sur une année pour un individu de 70kg. On sera ansi amener à évaluer la probabilité de dépasser ce seuil mais aussi la probabilité de dépasser la DJA/10 et plus généralement la proba de dépasser un seuil d_0 fixé. Inversement dans une optique de calibrage, si α est un seuil petit par exemple 10^{-6}, si l'on pose $\delta = 1-\alpha$, $x_\delta = F^{-1}(\delta)$ est donc le seuil à partir duquel "seulement" 1 personne sur 1 million sera touchée par le risque sanitaire, cette quantité est l'analogue de la "Value at Risk" ou VAR en finance. Ainsi si cette quantité est grande par rapport au seuil de toxicité, il y a lieu de s'inquiéter sur les risques d'exposition.Si elle est beaucoup plus petite, on peut éventuellement songer à réviser les normes en vigueur.

Soit $X_1, ...X_n$ un échantillon de taille n. On note en général

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

l'échantillon ordonné, de sorte que $X_{n,n}$ est la valeur maximale de l'échantillon. Il est facile de voir que $X_{n,n}$ converge lorsque $n \to \infty$ vers le point terminal de l'échantillon (fini si la distribution a un support fini à droite, infini sinon). Dans l'optique d'un théorème limite et de la construction d'intervalles de confiance ou de prédiction, on peut alors s'intéresser aux renormalisations de cet estimateur qui conduisent à une loi limite. On dit que G est une loi des extrêmes, s'il existe des suites a_n et b_n telles que

$$\frac{X_{n,n} - a_n}{b_n} \underset{n\to\infty}{\to} W$$

où W est une v.a. de distribution non dégénérée G. Compte tenu du fait que l'on peut toujours normaliser a_n *et* b_n de manière à prendre en compte les paramètres de taille et d'échelle, il n'existe d'après le théorème de Fisher Tippett que trois lois possibles pour G selon la forme de la queue de la distribution F des X_i :

Loi de type I: Gumbel $G_0(x) = exp(-exp(-x))$ avec

$$a_n = F^{-1}(1-1/n)$$

et

$$b_n = \overline{F}(a_n)^{-1} \int_{a_n}^{\infty} \overline{F}(u)du$$

Loi de type II : Fréchet pour $\gamma > 0$, $F_\gamma(x) = \left\{ \begin{array}{c} \exp(-x^{-1/\gamma}),\ \text{si } x > 0 \\ 0 \quad ,\ \text{sinon} \end{array} \right.$ avec $a_n = 0$ et $b_n = F^{-1}(1-\frac{1}{n})$

Loi de type III : Weibull pour $\gamma < 0$, $W_\gamma(x) = \left\{ \begin{array}{c} \exp(-(-x)^{-\gamma}),\ \text{si } x < 0 \\ 1 \quad ,\ \text{sinon} \end{array} \right.$ avec $a_n = s(F)$ et $b_n = a_n - F^{-1}(1-\frac{1}{n})$

que l'on peut représenter (par passage à la limite de γ en 0 et à une normalisation près) sous la forme suivante dite représentation de Jenkinson.

$$G_\gamma(x) = \exp(-(1+\gamma x)^{-1/\gamma}) \text{ si } \ 1+\gamma x > 0$$

Le cas limite $\gamma \longrightarrow 0$ correspond à la loi de Gumbel, le cas $\gamma > 0$ à la loi de Fréchet et $\gamma < 0$ à la loi de Weibull. Si la loi du maximum de n v.a. aléatoires i.i.d. de loi F est G_γ alors on dit que le maximum est attiré par G_γ et par extension que F appartient au domaine d'attraction de G_γ, ce qui est noté $F \in D(G_\gamma)$. On peut par exemple montrer que la loi normale, la loi exponentielle et la loi log-normale appartiennent au domaine d'attraction de la loi de Gumbel. Cette propriété est due au fait que toutes ces lois possèdent des queues de distribution peu épaisses.

Les lois de Pareto, de Cauchy, de Student appartiennent au domaine d'attraction des lois de Fréchet. Ces lois se caractérisent par la présence de queues de distribution lourde (non-exponentielle) ayant tendance à générer de grandes valeurs. L'indice γ comme nous le verrons dans la partie suivante est alors un indicateur de risque.

La loi uniforme et les lois qui ont un support fini mais avec une asymptote en leur point terminal (par exemple les lois béta) appartiennent au domaine d'attraction de la loi de Weibull. Le coefficient γ modélise le comportement de la loi des observations près du point terminal. Ce type de loi peut être utile pour modéliser des comportements à seuil. Par exemple dans une optique inverse de celle que nous adoptons ici, on peut s'intéresser aux personnes qui sont peu exposées à certains contaminants ou qui ont des déficiences en certains nutriments. Dans ce cas, on sera amené à étudier le comportement du minimum et de la loi au voisinage de 0 (par exemple s'il y a beaucoup de non consommateurs ou de personnes consommant peu d'un produit). Il peut alors être intéressant d'estimer le paramètre γ au voisinage de 0.

On dispose de caractérisations très précises du domaine d'attraction de chaque loi F en fonction du comportement de ces queues de courbes (voir Bingham, Goldie et Teugels (1987)). Ces caractérisations sont souvent très techniques et difficilement vérifiables par le praticien, aussi nous n'entrerons pas ici dans ses considérations techniques. Un article récent de Bertail, Haeffke, Politis, White (2000) montre qu'il est possible de proposer des estimations des constantes de normalisations et de la distribution asymptotique en s'affranchissant presque complètement des hypothèses faites usuellement sur la queue de courbe de F.

En terme de risque sanitaire, l'obtention des lois précédentes et en particulier l'estimation du coefficient γ que nous aborderons dans le paragraphe suivant sont importantes par exemple pour évaluer la probabilité que l'ensemble de la population soit en deçà d'un certain seuil d_0. On peut alors évaluer entre autre la quantité

$$P(\max_{1\leq i\leq n} X_i < d_0) \approx \exp(-(1+\gamma(d_0-a_n)/b_n)^{-1/\gamma}),$$

ce qui signifie que l'on doit non seulement estimer le coefficient γ mais également déterminer voire estimer les paramètres de renormalisation a_n et b_n. Si l'échantillon est de taille petite, on peut également s'intéresser au comportement du maximum sur une population de taille beaucoup plus grande N (par exemple à l'échelle nationale), auquel cas il est important de connaître la forme fonctionnelle des paramètres de renormalisation en fonction de n. L'objet de l'article de Bertail, Politis, White (2000)) est de proposer des estimateurs de ces quantités à partir de l'utilisation de méthodes de sous-échantillonnage.

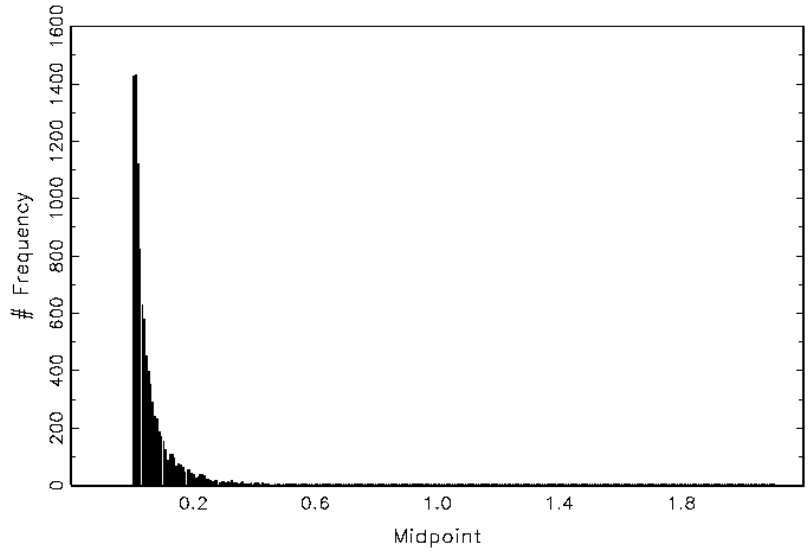

FIGURE 1. Distribution de la contamination totale en mercure en mg/an

2.2. **Loi de Pareto et Pareto généralisée**

. L'une des méthodes les plus fréquentes pour modéliser le comportement extrême des distributions et caractériser les quantiles extrêmes (voir par exemple quelques travaux empiriques appliqués à l'hydrologie, à la finance et à l'assurance dans Reiss et Thomas (2001)) est de modéliser les queues de distributions par des lois de type Pareto.

La figure 1 donne la forme de la queue de distribution empirique de la contamination globale en mercure obtenue à partir des données de panel Secodip (données par ménage ramenées à un individu, observées sur une année soit 3214 relevés, redressées par des pondérations obtenues par calage sur marge) et de données de contamination en mercure (essentiellement sur les produits de la mer frais, en conserve ou surgelés). Ces données (très incomplètes car ne tenant pas compte des repas hors domicile et construites en supposant une consommation identique de chaque membre du ménage) sont discutables : elles nous serviront plus à illustrer notre propos et à montrer comment on peut mettre en oeuvre les mé- thodes proposées, qu'à tirer des conclusions définitives. Dans le cas particulier du mercure, aucun individu ne se situe dans la zone à risque i.e. n'a de valeur supérieure à 18mg/an, dose annuelle admissible (ce qui est loin d'être le cas pour d'autres contaminants comme les dioxines ou l'ochratoxyne). Un estimateur empirique classique (nombre d'individus au dessus du seuil /nombre total d'individus) donnerait une probabilité de dépasser le seuil de 0, ce qui conduit à sous-estimer considérablement le risque. C'est pour cette raison que la modélisation de la queue de la distribution est indispensable. On notera que, de manière générale, sur ce type de données, la distribution a une queue très épaisse (la valeur maximum est de l'ordre de 2mg/an) ce qui justifie empiriquement l'utilisation de modèle de type Pareto.

Les avantages de ce type de modélisation par rapport à d'autres plus globales où l'on modélise le comportement d'ensemble de la distribution, par exemple au moyen de tests d'adéquation (voir par exemple Gauchi (2000)) sont doubles :

-on ne prend en compte ici que la partie intéressante de la distribution en termes de risque. On sait en effet que les tests usuels d'adéquation à des distributions connues (exponentielles, log-normales, gamma etc...) privilégient le centre de la distribution.

- l'approche est conservative dans la mesure où l'on aura toujours tendance à surévaluer les risques (les probabilités de dépasser un certain seuil), ce qui n'est pas le cas si l'on utilise des lois classiques avec queues de courbes exponentielles.

Pour x suffisamment grand, nous supposerons donc que la queue de courbe a la forme

$$F(x) = 1 - C/x^{\alpha} \tag{2.1}$$

où C est est une constante ou encore de manière plus robuste et plus générale

$$F(x) = 1 - L(x)/x^{\alpha} \tag{2.2}$$

où $L(.)$ est une fonction dite à variation lente (typiquement un paramètre d'échelle, un log ou des produits de log itérés) satisfaisant

$$\text{pour tout } t > 0,\ \frac{L(tx)}{L(x)} \to 1 \text{ quand } x \to \infty.$$

Ce type de fonction permet de rendre plus flexible la modélisation de la queue de la distribution et permet par exemple de tenir compte du fait que la population résultante est l'agrégation de plusieurs populations ayant des queues de courbes différentes.

On peut aisément montrer à partir des caractérisations de Von Mises que ces lois appartiennent au domaine d'attraction de la loi de Fréchet. On a dans ce cas $a_n = 0$ et $b_n = F^{-1}(1 - \frac{1}{n})$ et $\gamma = \alpha^{-1}$.

Il est aisé de montrer que l'on a respectivement pour (2.1) et (2.2),

$$\begin{aligned} F^{-1}(x) &= \left((1-x)/C\right)^{-1/\alpha} \\ b_n &= n^{1/\alpha} = n^{\gamma} \end{aligned}$$

et

$$\begin{aligned} F^{-1}(x) &= (1-x)^{-\gamma} l((1-x)^{-1}) \\ b_n &= n^{\gamma} l(n) \end{aligned}$$

où $l(.)$ est également une fonction à variation lente en ∞. La probabilité de dépasser un seuil d_0 est simplement donnée dans chacun des deux cas respectivement par

$$\overline{F}(d_0) = C d_0^{-\alpha}$$

$$\overline{F}(d_0) = d_0^{-\alpha} L(x_0)$$

qui sont des fonctions décroissantes de α.

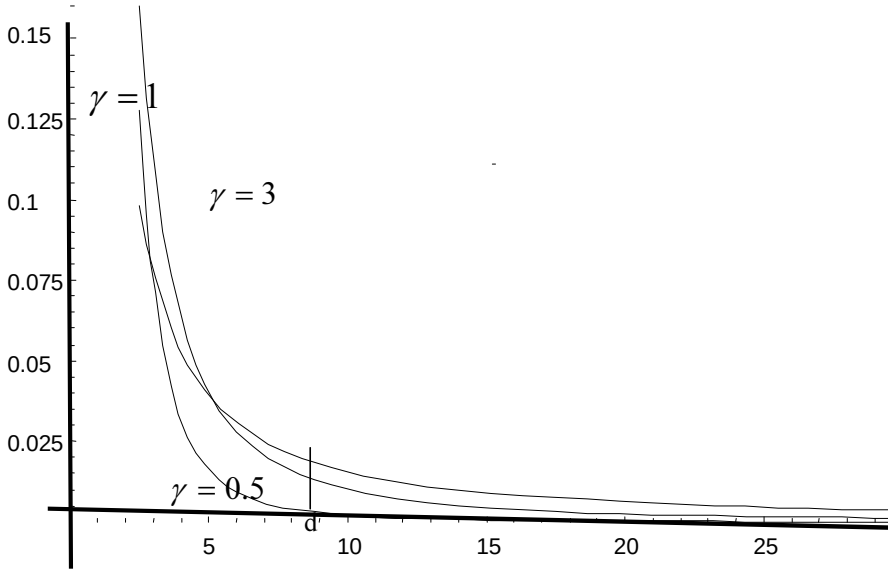

FIGURE 2. Comparaison des queues de courbes de type Pareto pour divers γ

On préfère généralement reparamétrer la loi de Pareto en introduisant l'indice $\gamma = 1/\alpha$, qui s'interprète directement comme un indice de risque. Plus γ est petit, moins la population extrême (représenteée par les queues de courbes) peut prendre de grandes valeurs (voir la figure 2). L'indice $\gamma = \infty$ correspond à une situation de risque maximal. Un des problèmes de la théorie statistique des valeurs extrêmes est de fournir une estimation adéquate de α ou γ, ce qui est clairement plus aisé dans le modèle (2.1) que dans le modèle général semi-paramétrique (2.2) dans lequel la fonction à variation lente joue le rôle d'un paramètre de nuisance de dimension infinie. L'introduction de la fonction à variation lente n'est pas simplement un jouet mathématique comme ont tendance à le penser certains statisticiens appliqués. Des fonctions à variations lentes peuvent apparaître très naturellement lorsqu'on modélise par exemple des phénomènes agrégés ou que l'on considère des mélanges de populations ayant des risques différents. Nous montrons en particulier dans l'annexe 1 sur quelques exemples précis comment celles-ci peuvent apparaître.

Ayant observé un échantillon (statique) de consommation de taille n, l'estimation de α permet alors d'évaluer les probabilités de dépasser un certain seuil déterministe de toxicité ou dans une approche inverse de caractériser les forts consommateurs en déterminant les quantiles extrêmes de la distribution, typiquement $F^{-1}(1-\zeta)$ pour ζ très petit parfois inférieur à $1/n$.

Une paramétrisation en terme d'indice de risque γ permet d'introduire une forme plus générale de loi de Pareto qui joue un rôle important dans la mé- thode d'estimation dite P.O.T. (peak over threshold : "pic au dessus d'un seuil") et la caractérisation des populations dites à risques. Celle-ci a la forme suivante

$$W_\gamma(x) = \begin{cases} 1-(1+\gamma x)^{-1/\gamma} \text{ pour } \begin{cases} 0<x \text{ et } \gamma>0 \\ 0<x<1/|\gamma| \text{ et } \gamma<0 \end{cases} \\ \\ \exp(-x) \quad \text{pour } x>0 \qquad \text{et} \quad \gamma=0 \end{cases}$$

Lorsque X est de loi Pareto, c'est la loi conditionnelle de $X > x + d_0$ sachant que $X > d_0$ (pour $d_0 = 1/\gamma$) d'où son nom de loi des excès. Il est clair que W_γ est de type Pareto pour $\gamma > 0$ (elle appartient donc au domaine d'attraction de la loi de Fréchet). W_0, la limite de W_γ lorsque $\gamma \to 0$, est une loi exponentielle (dans le domaine d'attraction de la loi de Gumbel) . Pour $\gamma < 0$, W_γ est à support borné et de type béta (dans le domaine d'attraction de la loi de Weibull). De manière générale, on a donc

$$W_\gamma \in D(G_\gamma)$$

En terme de risque d'exposition à un certain contaminant au dela d'un certain seuil, cette distribution peut permettre de modéliser des comportements très différents et est particulièrement adaptée pour mettre en évidence des populations plus ou moins exposées à des risques. En effet, si γ est grand alors la queue de courbe de la distribution est très épaisse et la probabilité que la consommation dépasse un certain seuil d_0 donné est grande. Si $\gamma = 0$, cette probabilité est faible. Enfin si $\gamma < 0$ (par exemple pour des sous-populations de non- ou faibles consommateurs des produits contaminés), la probabilité est très faible si $d_0 < 1/|\gamma|$ et nulle pour $d_0 \geq 1/|\gamma|$. Ainsi dans ces conditions, $1/|\gamma|$ s'interprète comme le seuil de risque nul. Pour mettre une plus grande flexibilité d'estimation et tenir de phénomène d'echelle, il sera utile d'introduire des paramètres μ et $\sigma > 0$ et de considérer

$$W_{\gamma,\mu,\sigma}(x) = W_\gamma((x-\mu)/\sigma).$$

Dans ces conditions μ s'interprète comme l'infimum du support et σ est un paramètre d'échelle. On notera que dans le cas $\gamma < 0$ le support de la loi est $[\mu, \mu + \sigma/|\gamma|]$.

2.3. **L'estimateur de Hill.** L'estimateur de Hill (1975) de γ est sans doute le plus utilisé de la théorie des valeurs extrêmes, même si de nombreux travaux récents remettent en cause sa suprématie (voir par exemple l'ensemble des travaux récents de Beirlant, KUL, Belgique). L'estimateur de Hill pour un k fixé dans $\{1, ..., n-1\}$ ne fonctionne que pour $\gamma > 0$ et est donné par

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

Il s'interprète comme l'estimateur du maximum de vraisemblance de γ dans le modèle (2.1), lorsqu'on ne conserve que les k plus grandes valeurs ou plus simplement comme un estimateur de la pente d'un QQ (quantile–quantile) plot. Rappelons que la méthode du QQ-plot est une méthode graphique empirique très simple pour tester l'adéquation d'une distribution empirique à une loi F donnée se basant simplement sur la constatation que les $F^{-1}(X_{i,n})$ suivent la même loi que n variables uniformes ordonnées d'espérances respectives $\frac{j}{n+1}$ de sorte que les points $(X_{i,n}, F^{-1}(\frac{i}{n+1}))$ pour i grand doivent être quasiment alignés sur une droite La figure 3 donne ce graphique dans le cas de la distribution de la contamination globale par le mercure.

L'estimateur de Hill est un estimateur trivial de la pente à l'infini. Cependant il est clair que l'estimateur de Hill est très sensible au choix du nombre de points

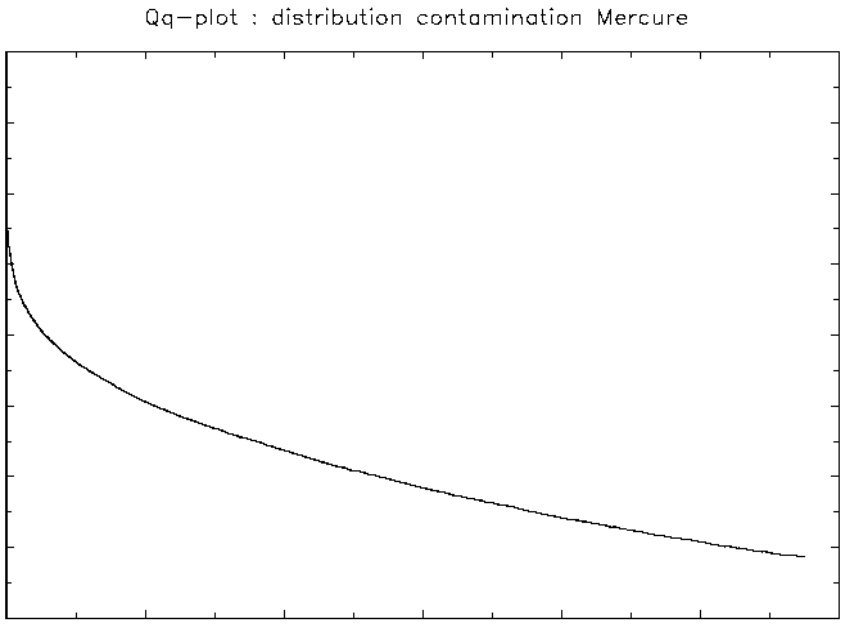

FIGURE 3. QQ-plot : abscisse contamination

retenus dans la queue de distributions k permettant de le calculer, comme le montre le graphique suivant donnant $H_{k,n}$ en fonction de k. Ce type de graphique est connu sous le nom de "Hill-Horror Plot" dans la littérature financière à cause du mauvais comportement de l'estimateur que l'on constate ici aussi. Théoriquement si k est petit devant n, cet estimateur est convergent de γ et l'on devrait donc observer une certaine stabilité de l'estimateur ce qui est loin d'être le cas en pratique.

Ce comportement s'explique par le fait que pour des tailles de k petites, la va- riance de l'estimateur est forte (forte variabilité du graphique près de l'origine) tandis que pour des tailles de k élevées, la queue de distribution n'est plus strictement de type Pareto (2.1) mais plutôt de type (2.2). La fonction à variation lente (qui peut s'expliquer par le fait que la distribution dans le cas de la contamination est un mélange de plusieurs Pareto) induit un biais fort sur l'estimateur. Par ailleurs, elle peut induire une vitesse de convergence de l'estimateur très lente vers la gaussienne (en $\log(n)$ si la fonction a variation lente est elle même en log). Des méthodes d'élimination systématique du biais et de choix optimal de k (en terme d'écart-quadratique moyen) ont été proposées par Hall et Feuerverger (1999), Beirlant et al (1999). Le graphique 5 donne l'estimateur de Hill débiaisé par la méthode de Beirlant et al.(1999), ainsi que les valeurs d'autres estimateurs classiques de α (estimateur par la méthode des moments, l'estimateur de Pickands et Drees Pickands). Il apparaît clairement que c'est l'estimateur de Hill débiaisé (et dans une moindre mesure l'estimateur par la méthode des moments) qui possède la plus grande stabilité. Le choix optimal de k obtenu par la méthode de Beirlant et al (1999) est $k_{opt} = 260$ conduisant à une estimation de valeur de l'index

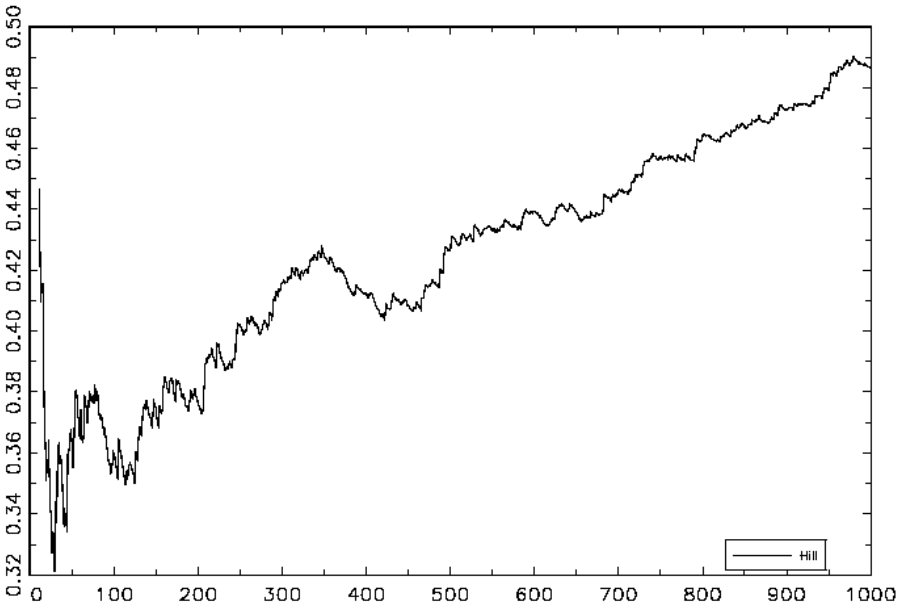

FIGURE 4. Estimateur de Hill $\widehat{H}_{k,n}$ en fonction de k

de $\widehat{H}_{k,n} = 0.394$. La constante C intervenant dans (2.1) peut être estimer par $X_{n,k}\,(\frac{k}{n})^{1/\widehat{H}_{k,n}}$. Ceci permet d'estimer le risque de dépasser la DJA du mercure de l'ordre de 1.1e-4. Compte tenu de la nature des données, on peut penser que cette évaluation est plutôt optimiste. Une évaluation plus précise basée sur des consommations individuelles (incluant les consommations hors domicile) est actuellement en cours.

Ainsi si l'on dispose d'un estimateur $\widehat{\gamma}_{k,n}$ de γ (par exemple $H_{k,n}$ ou sa version débiaisée), il est alors possible de prédire simplement par extrapolation sous l'hypothèse (2.1) la valeur des quantiles extrêmes de la manière suivante

$$\widehat{\overline{F}}^{-1}_{n,k}(\delta) = X_{n-k,n}(\frac{k+1}{(n+1)}/\delta)^{\widehat{\gamma}_{k,n}} \tag{2.3}$$

Un des points fréquemment omis dans la littérature appliquée sur les extrêmes est l'estimation de la fonction à variation lente et la construction d'intervalles de confiance pour une transformation non-linéaire du paramètre γ et notamment de la VaR (en théorie (2.3) n'est valide que sous (2.1)). Des travaux tenant compte de ce problème avec applications à des données financières ont été récemment réalisés par Bertail, Haeffke, Politis et White (2000). Les auteurs y proposent de nouvelles méthodes d'estimation de l'index α, en présence du paramètre de nuisance L. L'idée est de généraliser et d'utiliser les propriétés universelles des méthodes de sous-échantillonnages (voir Politis et Romano (1994)) et d'estimer la vitesse de convergence du maximum pour obtenir simultanément un estimateur de α et de la fonction à variation lente. On peut alors montrer que l'estimateur de la vitesse

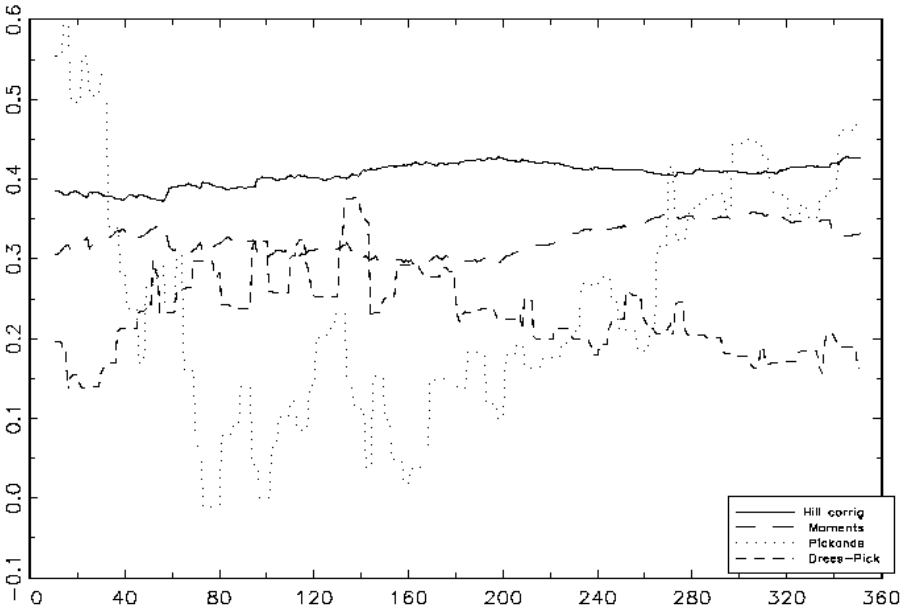

FIGURE 5. Comparaison d'estimateurs de l'index de Paréto, contamination en mercure.

de convergence peut présenter des changements de régime qui rendent plus facile le choix du k optimal. L'application de ces méthodes au cas de la contamination en mercure donne un estimateur dont le comportement en fonction de k_n est très caractéristique : une forte variabilité, un palier de stabilité (correspondant à la valeur de l'index) puis un fort biais (dû à un changement de régime): voir la figure 6.

Le choix optimal de k dans ce cadre est $k_{opt} = 244$ et conduit à une estimation de l'ordre de 0.387 très proche de celle obtenue avec l'estimateur de Hill débiaisé par la méthode de Beirlant et al.(1999).

2.4. **L'estimation indirecte : méthode P.O.T..** La méthode la plus ancienne pour estimer l'index α ou γ consiste à utiliser directement la forme de la loi des extrêmes et à ajuster une loi de type extrême généralisée à la loi du maximum. Cette méthode a été très largement critiquée du fait de la perte d'information, évidente lorsqu'on ne dispose que d'un échantillon (et donc d'un seul maximum). La méthode P.O.T. (Peak Over Threshold) (développée dans les année 70 en hydrologie puis abondamment étudiée en statistique, voir par exemple Pickands (1975), Smith (1987), Davison et Smith (1990), ou Reiss et Thomas (2001) pour de plus amples références) est une méthode qui repose sur le comportement des valeurs observées au delà d'un seuil d. Si on observe $X_1, X_2, ..., X_n$ on appelle $Y_1 - d, Y_2 - d, ..., Y_{K(n)} - d$, excès d'ordre d (les pics au dessus du seuil d), les valeurs des X_i (recentré ultérieurement par d) qui dépassent le seuil d. Le nombre $K = K(n)$ de telles

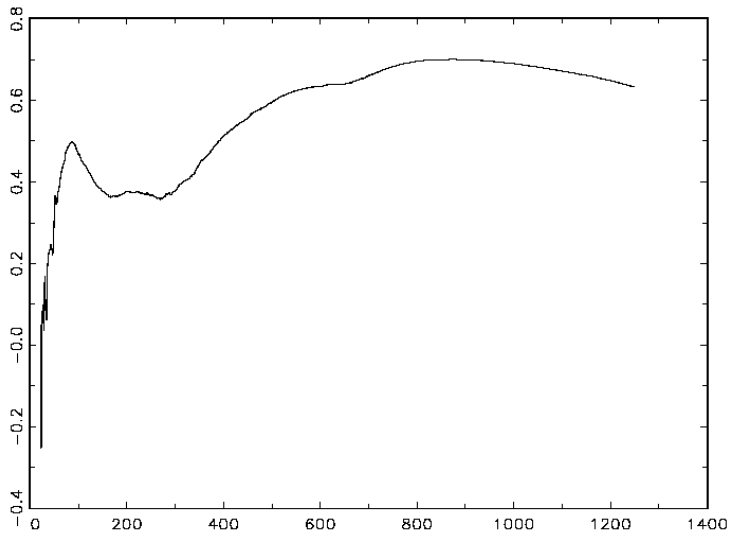

FIGURE 6. Estimateur de γ basé sur la méthode de Bertail at al.(2000)

variables est aléatoire de loi binomiale $B(n, \overline{F}(d))$ (en effet $K = \sum_{i=1}^{n} 1_{\{Xi>d\}}$)

$$\Pr(K = k) = C_n^k \overline{F}(d)^k (1 - \overline{F}(d))^{n-k}.$$

Conditionnellement à K, les Y_i ont pour distribution

$$\begin{aligned} F_d(x) &= \Pr(X \leq x + d \,|X > d) \\ &= (F(x + d) - F(d))/(1 - F(d)), \text{ pour } x \geq d \end{aligned}$$

La théorie des processus ponctuels permet de montrer qu'il y a en fait totale séparation (indépendance) entre les valeurs des Y_i et le nombre de telles valeurs. On peut aisément constater que les lois de Pareto généralisées $W_{\gamma,\mu,\sigma}(x)$ sont les seules lois qui assurent une stabilité de la loi des excès au delà d'un certain seuil dans la mesure ou il existe des paramètres σ_d et μ_d tels que $F_d(x) = F((x - \mu_d)/\sigma_d)$ pour $F = W_{\gamma,\mu,\sigma}$.

On peut alors montrer que si F est dans le domaine d'attraction d'une loi des extrêmes alors on a

$$\lim_{d \to s(F)} \sup_{0 \leq x \leq s(F)-d} |F_d(x) - W_{\gamma,0,\sigma(d)}(x)| = 0$$

i.e. que l'on peut approcher la loi des excès pour un seuil grand (proche du point terminal) par une loi de Pareto généralisée de variance inconnue (dépendant de d).

Une des méthodes les plus utilisées pour déterminer un estimateur de γ et de la VAR est de ne considérer que les valeurs dépassant un certain seuil d assez grand et d'y ajuster une loi de type Pareto généralisée puis d'estimer les paramètres par

la méthode du maximum de vraisemblance (E.M.V.). Smith (1987) a montré que pourvu que $\gamma < 1/2$, l'estimateur du maximum de vraisemblance existe et est asymptotiquement gaussien. En effet pour $\gamma < 1/2$, les moments d'ordre 2 existent et la matrice d'information de Fisher est finie. D'autres méthodes basées sur le calcul de moments ont été également proposées. Cette approche est très utilisée en finance (Teugels (1985)) ou en hydrologie (Hosking et Wallis (1987)). La question la plus problématique d'un point du vue tant théorique que pratique est le choix du seuil d (équivalent en fait dans l'approche directe au choix k du nombre de maxima à retenir pour le calcul de l'estimateur de Hill). Dans notre cadre, ce type d'estimation de γ conduit à des résultats très proches de ceux déjà obtenus mais s'avère plus intéressants dans l'optique de la partie suivante.

3. Modélisation et caractérisation des populations à risque.

Mettre en évidence des populations à risque revient implicitement à supposer que, conditionnellement à certaines variables exogènes $Z_1, ..., Z_n$ (qui vont définir des sous-populations), le risque d'exposition à certains contaminants est différent. On peut dans un premier temps pour mettre en évidence cette hétérogénéité essayer de comparer pour différentes catégories les estimateurs des indices de risques sur des sous-populations. Le graphique 7 donne par exemple les estimateurs de Hill obtenus pour des C.S.P. différentes.

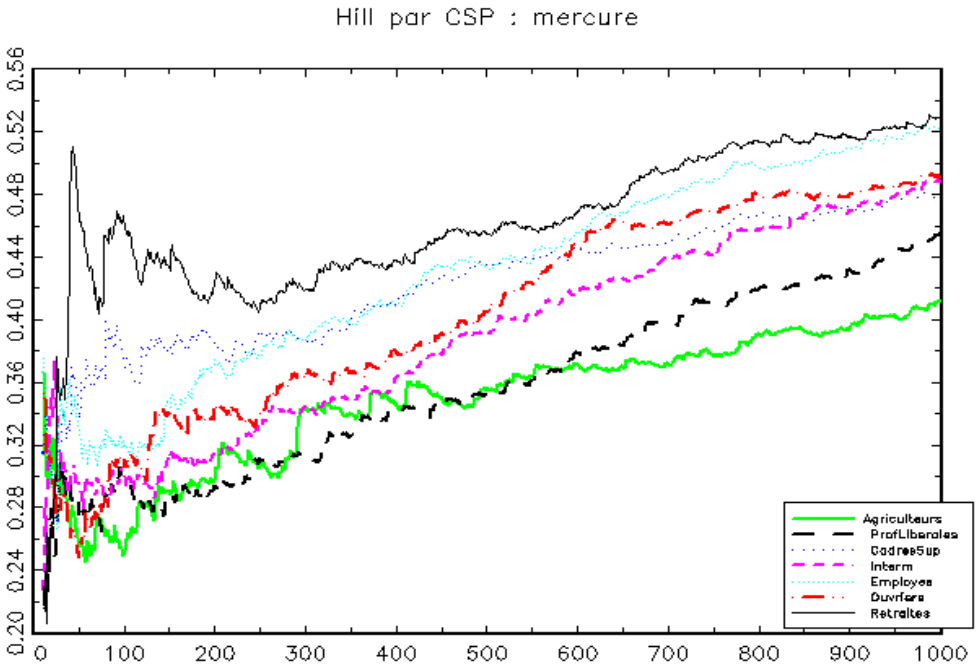

Figure 7. Hill par CSP

Bien que l'on se heurte là encore au problème du biais et du choix optimal de k, on constate cependant une certaine hiérarchie dans les niveaux de risque (avec

un indice très fort pour les retraités et les cadres supérieurs et beaucoup plus faible pour les professions libérales et les agriculteurs). C'est ce phénomène que l'on aimerait pouvoir confirmer par des méthodes plus précises. Il faut en effet se méfier d'une interprétation directe de ce graphique: l'effet taille des sous-populations peut fortement affecter la précision des estimateurs, mais aussi le choix du k optimal qui a priori est différent pour chacune de ces sous-populations. Une solution possible qui permet d'estimer l'impact des variables socio-démographiques simultanément est de considérer un modèle du type Pareto ou Pareto généralisé dans lequel l'indice de risque est, conditionnellement aux variables socio-démographiques Z, une fonction de ces variables,

$$\gamma = h(Z).$$

De manière à pouvoir tester l'impact de certaines variables sur le niveau du risque, il est plus intéressant de faire des hypothèses sur la forme du lien. En effet un modèle totalement non-paramétrique ne serait pas identifiable. Une spécification possible et simple (pour les besoins de l'exposé) est de retenir une formulation de type "single-index" pour l'index γ, c'est à dire une fonction de lien h de la forme

$$\gamma = \Gamma(Z'\beta)$$

et une forme de type Pareto généralisé pour la queue de distribution. Dans la formulation la plus générale du modèle, on peut supposer la fonction Γ inconnue. Ce type de modèle fera l'objet de travaux ultérieurs. Nous supposerons ici que Γ est connu spécifié typiquement linéaire (si les variables explicatives sont toutes des variables dichotomiques) ou est une fonction bornée Dans cette approche, l'estimation du modèle permet de quantifier l'impact des variables explicatives sur le niveau de risque d'exposition encouru. Ce modèle ne permet néanmoins pas de séparer les populations à faibles risques (celles qui contribuent à la distribution pour $X < d$) des autres.

3.1. **Caractérisation des populations à risques à partir de la loi des excès.** Un modèle possible est de considérer qu'au delà d'un certain seuil d conditionnellement aux vecteurs $Z = (Z_i)_{0\leq i\leq q}$ où $Z_0 = 1$, la distribution des excès (distribution de $X - d$ conditionnellement à $X > d$ et à Z) est du type

$$W_{Y|Z}(x) = 1 - (1 + \Gamma(Z'\beta)y/\sigma)^{-1/\Gamma(Z'\beta)}, \tag{3.1}$$

où Γ est une fonction croissante bornée (la borne supérieure étant $1/2$) nulle en 0. L'index $\gamma = \Gamma(Z'\beta)$ est donc à la transformation non-linéaire Γ près, une fonction linéaire des observations (en effet Γ^{-1} existe et $\Gamma^{-1}(\gamma) = Z'\beta$) . L'hypothèse de croissance de la fonction Γ permet d'interpréter directement le signe et la valeur des coefficients $(\beta_i)_{0\leq i\leq q}$. Nous reviendrons dans la suite sur l'hypothèse de bornitude de la fonction Γ.

Ce type de spécification dans lequel l'index dépend de variables explicatives avec une forme fonctionnelle linéaire pour Γ, a été introduit par Davison et Smith (1990). Le fait que la fonction de lien soit non bornée induit néanmoins une structure très forte sur la loi non conditionnelle de Y. En effet (voir annexe 1), si la loi de Z charge tout $\mathbb{R}^+$, la loi agrégée de Y est de type Pareto avec un indice de risque $\gamma = \infty$, situation qui est rarement réaliste en pratique.

Par ailleurs, si $\Gamma(Z\prime\beta) > 1$, l'EMV n'est même pas convergent (voir Smith (1987)). L'introduction d'une fonctionnelle Γ bornée par $1/2$ (pour assurer la normalité asymptotique de l'estimateur du maximum de vraisemblance) permet

d'introduire une plus grande flexibilité dans le modèle : par ailleurs la forme de Γ peut également donner des renseignements sur d'éventuels phénomènes de seuil ou de saturation.

Dans ce cadre, la log vraisemblance du modèle (basées sur les K valeurs $Y_i = X_i - d > 0$ et leurs covariables associées $Z_{[i]}$) est donnée par

$$l_W(y_1, ...y_K, \sigma, \beta) = -\sum_{i=1}^{K} \log(\sigma) + (1 + \frac{1}{\Gamma(z'_{[i]}\beta)}) \log(1 + \Gamma(z'_{[i]}\beta) y_i/\sigma)$$

Les estimateurs du maximum de vraisemblance de β et σ sont solutions des équations

$$-K\sigma + \sum_{i=1}^{K} \frac{\Gamma(z'_{[i]}\widehat{\beta}) + 1}{1 + \frac{\Gamma(z^{i\prime}_{[i]}\widehat{\beta})}{\widehat{\sigma}} y_i} y_i = 0$$

$$\sum_{i=1}^{K} \frac{z'_{[i]}\Gamma^{(1)}(z'_{[i]}\widehat{\beta})}{\Gamma(z'_{[i]}\widehat{\beta})} \left[\frac{1}{\Gamma(z'_{[i]}\widehat{\beta})} \log(1 + \Gamma(z'_{[i]}\widehat{\beta}) y_i/\widehat{\sigma}) - (1 + \Gamma(z'_{[i]}\widehat{\beta})) \frac{y_i/\widehat{\sigma}}{1 + \Gamma(z'_{[i]}\widehat{\beta}) y_i/\widehat{\sigma}} \right] = 0$$

L'information de Fisher du modèle vaut

$$I_1(\beta, \sigma) = \begin{pmatrix} \sum_{i=1}^{K} z'_{[i]} z_{[i]} \frac{2\Gamma^{(1)}(z'_{[i]}\beta)^2}{(1+\Gamma(z'_{[i]}\beta))(1+2\Gamma(z'_{[i]}\beta))} & I_{\beta,\sigma} = -\sum_{i=1}^{K} \frac{z_{[i]}\Gamma^{(1)}(z'_{[i]}\beta)}{(1+\Gamma(z'_{[i]}\beta))(1+2\Gamma(z'_{[i]}\beta))} \\ I_{\beta,\sigma} = -\sum_{i=1}^{K} \frac{z_{[i]}\Gamma^{(1)}(z'_{[i]}\beta)}{(1+\Gamma(z'_{[i]}\beta))(1+2\Gamma(z'_{[i]}\beta))} & \frac{1}{\sigma^2} \sum \frac{1}{1+2\Gamma(z'_{[i]}\beta)} \end{pmatrix}$$

Ce modèle est intéressant dans la mesure où il permet à partir de techniques classiques d'estimation (EMV) d'obtenir des informations sur l'impact des variables exogènes Z sur la forme des queues de distributions et donc sur l'indice de risque. Néanmoins on constatera, qu'une grande partie de l'information est perdue dans la mesure où seules les K plus grandes valeurs de l'échantillon sont retenues. Or il apparaît intéressant de comprendre non seulement l'impact des Z sur les consommateurs avec un fort risque de contamination mais aussi de savoir quelles sont les variables qui influent sur l'appartenance ou non à la queue de la distribution.

Une solution est de proposer un modèle de type probit sur cette appartenance ou non, i.e. de modéliser $P(X > d)$ sous la forme

$$P(X > d|Z) = h(Z'\gamma).$$

Ce type de modèle est à rapprocher des modèles de type double Hurdle i.e. des modèles en deux étapes utilisés en économie du consommateur (voir Bertail, Caillavet, Nichèle (1999)) et peut se justifier dans le cadre de l'estimation des risques liés à certains contaminants par le fait que le risque peut provenir de deux sources: le fait de consommer ou non un produit contaminé (l'information pouvant jouer un role non-négligeable sur cette décision) puis dans un second temps du niveau de cette consommation. Les effets des variables explicatives sur la première étape (consommation ou non) peuvent être très différents de ceux sur le niveau. On peut très bien concevoir que le fait d'avoir des enfants a un impact positif sur le fait d'acheter des céréales et donc sur le risque d'exposition à l'Ochratoxyne A, mais que cette variable ait un effet nul (voire négatif) sur la probabilité que le niveau d'exposition (i.e. dans cette modélisation que γ soit très élevé).

Comme aucune information sur la distribution de la loi de Y sachant $Y < q$ n'est supposée, les estimateurs du maximum de vraisemblance de γ et β s'obtiennent en estimant respectivement le modèle probit dans la première étape, que ce soit par des techniques paramétriques usuelles (maximum de vraisemblance si h est spécifié) soit par des techniques non-paramétriques puis en estimant comme nous venons de le faire précédemment β par l'estimateur du maximum de vraisemblance.

On notera que l'un des inconvénients de ce modèle est que le seuil au delà duquel la loi est de type Pareto est supposé fixé. Une autre possibilité qui ne distingue pas entre les deux étapes est de modéliser directement le comportement de la queue de la distribution de la variable X et non plus de la distribution des excès Y.

Les résultats suivants ont été obtenus à partir d' informations socio-démographiques restreintes (catégories socio-professionnelles, diplômes, structure familiale, variables géographiques) issues de l'enquête Secodip associées aux données de contamination par le mercure. Le mercure est un produit relativement simple à analyser vu sa présence dans un faible nombre de produits. Les résultats suivants montrent l'intérêt d'une approche en deux étapes. L'étape probit (sous l'hypothèse usuelle de normalité des résidus du modèle latent) et le modèle (3.1).ont été estimés par la méthode du maximum de vraisemblance.

La figure 8 permet de comparer les estimateurs du maximum de vraisemblance dans le modèle probit (appartenance ou non à la queue de distribution) obtenus lorsque l'on fait varier le nombre d'individus retenus dans la queue de distribution à partir d'un seuil d_1 suffisamment grand (ici de l'ordre 1.7mg). Ceci permet d'éviter l'écueil du choix de d et donc de voir dans quelle mesure les estimateurs obtenus sont robustes à ce choix. Les intervalles de confiance étant très serrés autour de la valeur estimée, ils n'ont pas été représentés sur le graphique : seules quelques variables (les variables de diplôme) ne sont pas toujours significatives.

On note sur ce graphique la très grande stabilité des coefficients. La variable de référence pour les CSP est la catégorie "profession intermédiaire". Toutes les autres catégories ont un impact négatif (par rapport à la référence) sur l'appartenance à la région à risque : l'impact est particulièrement marqué pour les agriculteurs et les inactifs (chef de famille inactif), ce qui s'interprète facilement par la part très faible des produits de la mer dans la consommation de ces catégories. Le fait d'avoir des enfants a aussi un impact négatif fort sur l'appartenance à la région à risque.

Dans les graphiques suivants nous analysons l'impact des variables retenues sur le risque, c'est à dire la potentialité de l'individu à se trouver dans les région extrêmes en fonction des variables retenues. On notera que cet impact peut être complètement différent de celui observé dans le probit, un phénomène très fréquemment observé dans les modèles de consommation (cf. Bertail, Caillavet et Nichèle (1999)) Nous présentons dans les figures 9 à 11 les estimateurs ainsi que les intervalles de confiance dans le modèle (3.1) associés aux variables de CSP, diplôme et avec enfant/sans enfant. Les variables de référence sont respectivement pour la CSP "profession intermédiaire", BEPC pour les diplômes et sans enfant .

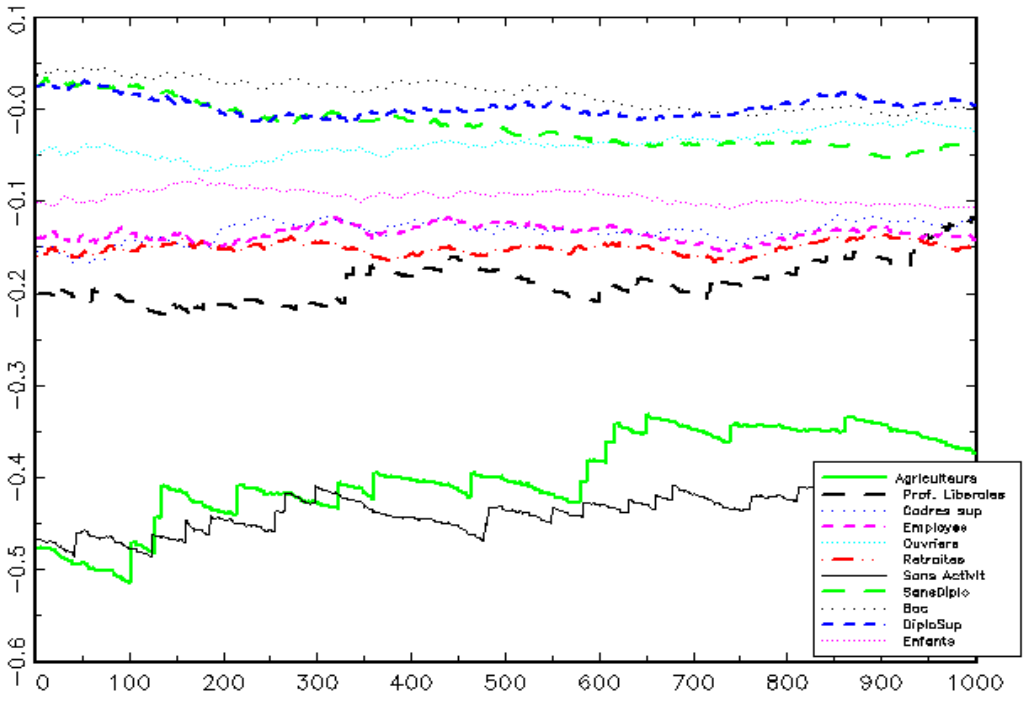

FIGURE 8. Coefficients estimés du modèle probit

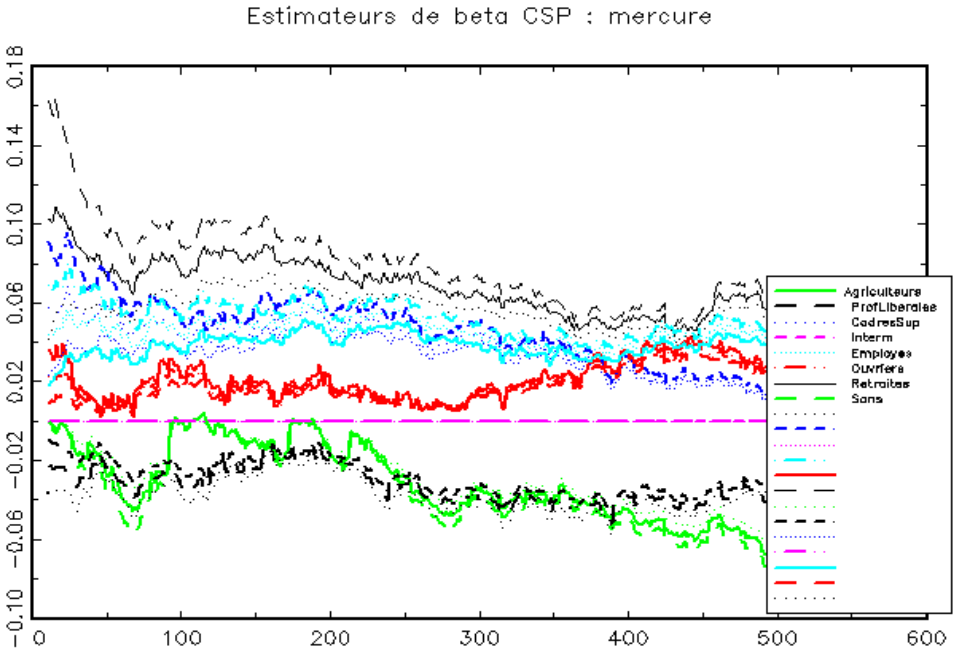

Figure 9 : Estimation simultannée de l'impact des variables CSP sur le risque.

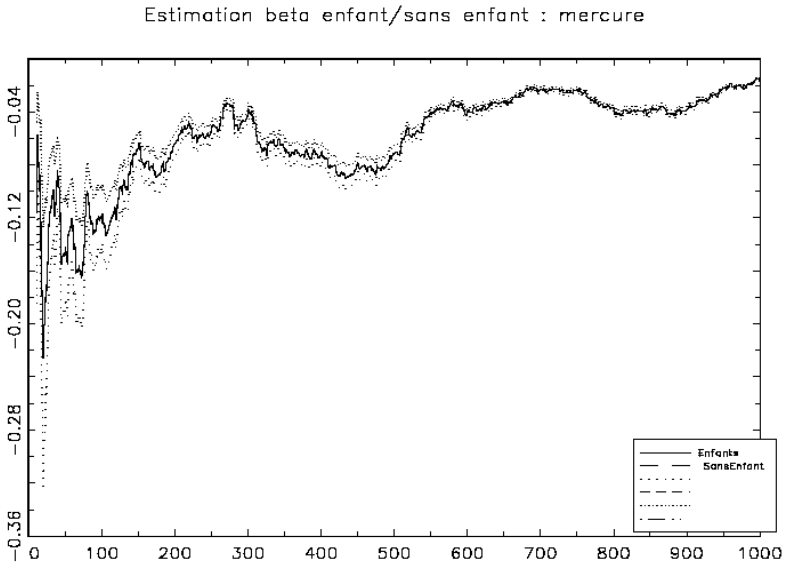

Figure 10 : Impact de la variable sans Enfant sur le niveau du risque d'exposition au mercure

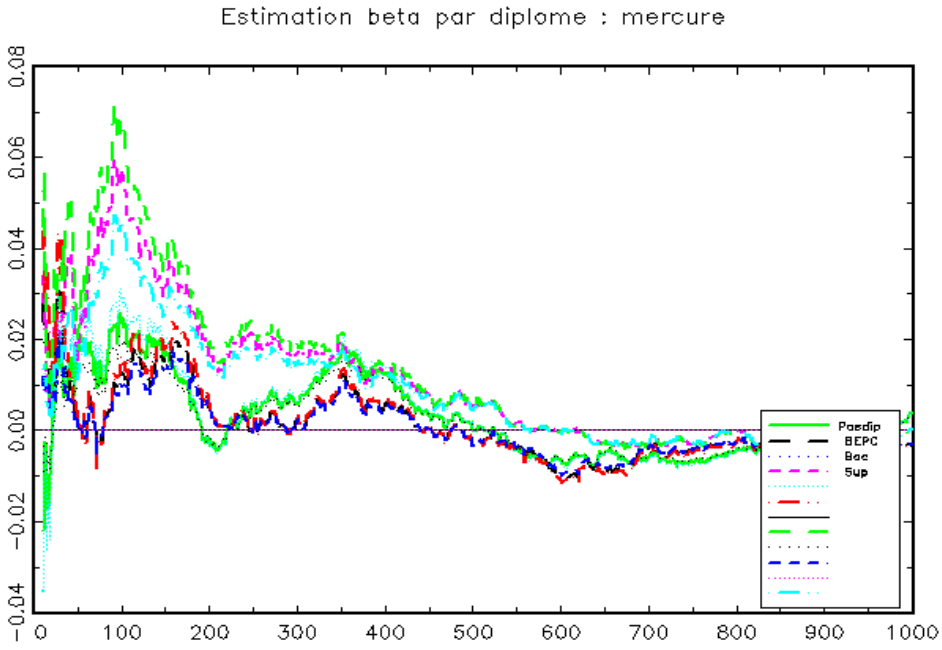

Figure 11 : Impact du diplome sur le niveau du risque d'exposition

On constate que l'appartenance aux CSP agriculteurs et professions libérales a un impact négatif sur le risque de contamination. Plusieurs explications (que viennent confirmer d'autres études en cours) peuvent expliquer ce phénomène :

-il reflète des pratiques alimentaires propres aux CSP (les agriculteurs mangent effectivement peu de produits de la mer)

-l'effet de l'information sur la contamination potentielle des produits peut avoir un effet plus grand chez les professions libérales que chez d'autres CSP.

Par ailleurs, l'appartenance aux CSP "Cadre Sup", "Employé" et "Retraité" a un impact positif significatif (quel que soit le seuil choisi) sur le risque. Pour les premiers, on peut penser que cet effet est lié au revenu, les produits contenant du mercure étant des produits chers... Pour les retraités on peut penser qu'il s'agit à la fois d'un effet géographique "bord de mer" (nous n'avons pas pu inclure de variable géographique) et des préférences alimentaires marquées (poisson plus consommé que la viande pour ses supposées valeurs nutritionnelles et ses qualités masticatrices...)

Le graphique 10 étudie l'impact des variables "avec enfants", "sans enfants". Il montre que le fait d'avoir des enfants (variables de référence "sans enfant") a un impact négatif significatif sur le risque. On notera que le coefficient est toujours significativement différent de zero mais que la valeur parait assez instable suivant le nombre d'observations retenues.

D'autres variables introduites dans le modèle semble plus difficile à interpréter c'est par exemple le cas du diplôme du chef de famille. Selon le choix de k l'impact des variables peut être positif ou négative par rapport à la variable de référence (ici Bac) par ailleurs les intervalles de confiance sont très larges . Il ne nous parait pas possible d'interpréter les résultats dans ce cas.

3.2. **Conclusions et améliorations.** L'évaluation des risques alimentaires liés à certains contaminants se heurte actuellement à deux écueils : l'absence quasi totale de données sur l'évaluation globale des consommations individuelles sur longue période et l'absence de fiabilité ou plutôt de cohérence des données de contaminations. La mise en correspondance de ces données est en soit un problème important que nous n'avons pas abordé ici mais fait actuellement l'objet de travaux. En effet , pour un contaminant touchant de très nombreux aliments, tel l'ochratoxine présent dans les fruits secs, les céréales et tous les produits dérivés (pain, pâtes, plats cuisinés), certains jus de fruits, le café, le vin etc...établir une correspondance entre nomenclatures de consommation et de contamination pour pouvoir construire des distributions de contaminations globales, via des coefficients d'équivalence, utiliser des données de sources de consommation différentes pour pouvoir couvrir l'ensemble de produits relèvent d'une véritable gageure. Il n'est même plus clair que les méthodes décrites ici puissent s'appliquer directement sur des données construites. Des travaux ultérieurs validant ces méthodes sont nécessaires si l'on veut avoir une bonne appréciation des risques, estimer des probabilités de dépasser des seuils déterministes ou tout simplement fixer des normes. Les problèmes également posés par la structure des données observées (couplage de bases, censures (à gauche) des contaminations dues aux limites de détection des matériels utilisés) posent de nombreuses questions auxquelles les statisticiens se doivent de répondre.

Les méthodes basées sur la théorie des valeurs extrêmes nous paraissent clairement plus adaptées à cette problématique que les techniques actuellement utilisées (essentiellement des moyennes de consommations pondérées par des moyennes de

contamination). Le but de cet article est de montrer en quoi elles peuvent apporter des débuts de réponse à l'évaluation quantitative des risques liés à certains contaminants alimentaires et ainsi contribuer à la mise au point de protocoles d'évaluation fiable. Les hypothèses de type Pareto sur les queues de distribution semblent parfaitement adaptées à la modélisation des contaminations fortes. Cependant, les estimateurs usuels (estimateur de Hill) peuvent s'avérer très mauvais dans ce cadre. Nos propos sont illustrés pour des questions de simplicité et de disponibilité des données sur le cas du mercure présent dans peu de produits consommés, mais les phénomènes décrits ici sont aussi valides pour de très nombreux contaminants. Des techniques plus avancées (débiaisage des estimateurs, estimation des fonctions à variation lente) permettent d'avoir une idée plus précise des risques en jeu et doivent être développées.

La comparaison des populations à risque et leur mise en évidence restent par ailleurs un problème délicat. Nous donnons ici un exemple de modélisation du niveau du risque en considérant un modèle de type Pareto dans lequel l'indice est lui-même une fonction spécifiée de certaines variables exogènes. Nous avons supposé la fonction de liens connue mais des travaux de types semiparamétriques plus généraux où la fonction n'est pas spécifiée peuvent s'avérer interessants. Dans le cas étudié, on met en évidence des risques plus grands de contamination chez les retraités, les catégories supérieures et les couples sans enfant, même si le risque de contamination par cette molécule semble relativement faible.

4. Annexe 1

Phénomènes d'agrégation pour des lois de Pareto.

Les phénomènes d'agrégation sont en général des phénomènes très complexes qui peuvent conduire à des lois agrégées parfois inhabituelles. Nous allons donner ici quelques exemples précis qui vont nous permettre de mieux comprendre dans quelles conditions les types de modèles que nous considérons ici peuvent être adaptés à la caractérisation des populations à risque.

Considérons un modèle dans lequel tous les individus sont caractérisés par une distribution de Pareto de la forme

$$P[X > x|Z = z] = x^{-1/(zb)} \text{ pour } x > 0$$

où X est une variable aléatoire réelle positive et b un coefficient positif.

De manière générale, si Z a pour densité f_Z de support $\Omega \subset R^{+*}$, la densité non-conditionnelle de la variable X (i.e. celle que l'on étudierait sans tenir compte de l'hétérogénéité des comportements) est donnée par

$$\begin{aligned} f_X(x) &= \int_\Omega \frac{1}{zb} x^{-1-1/zb} f_Z(z)dz \\ &= \int_\Omega \frac{1}{zb} \exp((-1-1/(zb))\log(x))\ f_Z(z)dz \end{aligned}$$

a) Z **uniforme sur un intervalle**$[a, a+1]$

Pour chaque z donné , la variable X d'intérêt se comporte comme une loi de Pareto d'indice de risque $\gamma = zb$ dans l'intervalle $[ab, a(b+1)]$. a s'interprète donc comme un paramètre de position et b comme un paramètre d'échelle.

La densité non-conditionnelle de la variable X est donnée par

$$\begin{aligned} f_X(x) &= \int_a^{a+1} \frac{1}{zb} x^{-1-1/zb} dz \\ &= \frac{1}{bx} \int_{\frac{\log(x)}{b(a+1)}}^{\frac{\log(x)}{ab}} \frac{e^{-t}}{t} dt \end{aligned} \tag{4.1}$$

Un équivalent asymptotique de (4.1) (pour x $\to \infty$) est donné par

$$f_X(x) \sim b^{-1} x^{-1-1/a(b+1)}$$

c'est à dire que la population globale se comporte comme une loi de Pareto d'indice de risque $\gamma = a(b+1)$, qui correspond en fait à l'indice de risque maximal dans la population indexée par z. De sorte que l'on aura toujours tendance dans ce cas à surestimer les risques en ne prenant pas en compte la variable Z.

b) **Z de type Pareto**

Si on suppose maintenant que la variable Z suit une loi de type Pareto de densité

$$f_Z(z) = \frac{\log(z)}{z^2}$$

i.e. de fonction de répartition

$$F_Z(z) = 1 - \frac{1}{z} - \frac{\log(z)}{z} \text{ pour } z > 1$$

(ce qui signifie qu'une partie de la population possède un indice de risque très grand avec une probabilité non-négligeable), alors la densité non conditionnelle de Y est donnée par

$$f_X(x) = \frac{1}{b}\int_0^\infty x^{-1-1/zb}\frac{\log(z)}{z^3}dz$$

Un calcul d'intégrale standard donne (γ désigne la constante d'Euler)

$$\begin{aligned} f_X(x) &= \frac{b}{2\log(x)^2}x^{-\frac{1+b}{b}}\Big(2-2x^{1/b}+2\gamma x^{1/b}+ \\ &\quad 2x^{1/b}\int_{\log(x)/b}^{+\infty}\frac{e^{-t}}{t}dt+2x^{1/b}\log(\log(x)/b)\Big) \end{aligned}$$

Si l'on prend la partie principale de cette expression pour x grand on a (C_b est une constante ne dépendant que de b)

$$f_X(x) \sim C_b x^{-1}\frac{\log(\log(x))}{\log(x)^2}$$

i.e. X se comporte donc comme une loi de Pareto d'indice $\gamma = \infty$ et a approximativement pour fonction de répartition pour des x grands $1-\frac{1}{Log[x]}-\frac{\log(\log(x)}{\log(x)}$. Là encore l'agrégation conduit à ne considérer que le pire des cas sans tenir compte de la structure de la population des Z. Un tel comportement est difficilement observable en pratique ce qui explique l'indroduction de la fonction Γ bornée introduite dans (3.1).

c) **Z avec des queues de courbes légères**

Si on suppose que Z a pour densité

$$f_Z(z) = 2z\exp(-z^2),\ z>0$$

de fonction de répartition

$$F_Z(z) = 1-\exp(-z^2)$$

alors

$$f_X(x) = \frac{2}{xb}\int_0^\infty \exp(-1/(zb)\log(x))\exp(-z^2)dz$$

Le théorème de Laplace permet d'obtenir un équivalent de cette expression de la forme (ou C_b^1 et C_b^2 sont des constantes ne dépendant que de b)

$$f_X(x) \sim \frac{C_b^1\log(x)^{1/3}}{x}\exp\left(-C_b^2\log(x)^{2/3}\right)$$

On en déduit que X se comporte encore dans ce cas comme une Pareto de type $\gamma=\infty$. Il est aisé d'obtenir un résultat similaire lorsque la queue de courbe est en $exp(-z^\alpha)$ pour $\alpha > 1$.(seule la fonction à variation lente change). On notera que dans ce cas très simple l'agrégation de lois de Pareto simples conduit à une distribution de type Pareto avec une fonction à variation lente dont la structure est très complexe.

Ce résultat montre que, en présence de variables explicatives de type gaussien., le modèle retenu avec $\Gamma(z) = z$ pour caractériser les populations à risque ne s'avère pas bon dans la mesure où les distributions d'exposition à des contaminants agrégées peuvent rarement, voire jamais, être modélisées par des Pareto d'indice $\gamma=\infty$ (correspondant au cas de risque maximal).

Le modèle ave Γ linéaire est donc plus intéressant lorsqu'on dispose de variables de type catégorielles ou plus généralement de variables explicatives bornées comme dans le cas a). De façon générale, ces exemples montrent clairement qu'il n'est pas très réaliste de considérer des fonctions Γ non bornées. Lorsque Γ est positive bornée avec pour borne supérieur γ_S alors la distribution agrégée de X est au pire une distribution de type Pareto d'indice de risque $\gamma = \gamma_S$, même si les variables explicatives sous-jacentes ont des supports infinis.

5. Annexe 2

On a dans le modèle (3.1),

$$\begin{aligned}\frac{\partial^2 l_W(y_1,...,y_K)}{\partial\beta\partial\beta'} &= \sum_{i=1}^{K} z_{[i]}^{'} z_{[i]} \left(\frac{2y_i\Gamma^{(1)}(z_i'\beta)^2}{\sigma\Gamma(z_i'\beta)^2(1+\frac{y_i}{\sigma}\Gamma(z_{[i]}'\beta))} \right. \\ &\quad + \log[1+\frac{y_i}{\sigma}\Gamma(z_{[i]}'\beta)] \left(-\frac{2\Gamma^{(1)}(z_i'\beta)^2}{\Gamma(z_i'\beta)^3} + \frac{\Gamma^{(2)}(z_i'\beta)}{\Gamma(z_i'\beta)^2}\right) \\ &\quad \left. -(1+\frac{1}{\Gamma(z_i'\beta)})(-\frac{y_i^2\Gamma^{(1)}(z_i'\beta)^2}{\sigma^2(1+\frac{y_i}{\sigma}\Gamma(z_{[i]}'\beta))^2} + \frac{y_i\Gamma^{(2)}(z_i'\beta)}{\sigma(1+\frac{y_i}{\sigma}\Gamma(z_{[i]}'\beta))}\right)\end{aligned}$$

$$\begin{aligned}&\frac{\partial^2 l_W(y_1,...,y_K)}{\partial^2\sigma} \\ &= \frac{K}{\sigma^2} - \sum_{i=1}^{K}(1+\frac{1}{\Gamma(z_i'\beta)})\left(-\frac{y_i^2\Gamma(z_{[i]}'\beta)^2}{\sigma^4(1+\frac{y_i}{\sigma}\Gamma(z_{[i]}'\beta))^2} + \frac{2y_i\Gamma(z_{[i]}'\beta)}{\sigma^3(1+\frac{y_i}{\sigma}\Gamma(z_{[i]}'\beta))}\right)\end{aligned}$$

$$\frac{\partial^2 l_W(y_1,...,y_K)}{\partial\beta\partial\sigma'} = \sum_{i=1}^{K} \frac{y_i\ z_{[i]}\Gamma^{(1)}(z_{[i]}'\beta)}{\sigma^2(1+\frac{y_i}{\sigma}\Gamma(z_{[i]}'\beta))}\left(-\frac{y_i(\Gamma(z_{[i]}'\beta)+1)}{\sigma(1+\frac{y_i}{\sigma}\Gamma(z_{[i]}'\beta))}+1\right)$$

On en déduit l'expression de la matrice d'information de Fischer

$$I(\beta,\sigma) = \begin{pmatrix} I_{\beta,\beta} & I_{\beta,\sigma} \\ I_{\beta,\sigma}' & I_{\sigma,\sigma}\end{pmatrix} = -E\begin{pmatrix} \frac{\partial^2 l_W(y_1,...,y_K)}{\partial\beta\partial\beta'} & \frac{\partial^2 l_W(y_1,...,y_K)}{\partial\beta\partial\sigma'} \\ \frac{\partial^2 l_W(y_1,...,y_K)}{\partial\beta\partial\sigma'}' & \frac{\partial^2 l_W(y_1,...,y_K)}{\partial^2\sigma}\end{pmatrix}$$

où

$$I_{\beta,\sigma} = -\sum_{i=1}^{K}\frac{z_{[i]}\Gamma^{(1)}(z_{[i]}'\beta)}{\sigma(1+\Gamma(z_{[i]}'\beta))(1+2\Gamma(z_{[i]}'\beta))}$$

$$I_{\beta,\beta} = 2\sum_{i=1}^{K} z_{[i]}z_{[i]}'\frac{\Gamma^{(1)}(z_{[i]}'\beta)^2}{(1+\Gamma(z_{[i]}'\beta))(1+2\Gamma(z_{[i]}'\beta))}$$

$$\begin{aligned}I_{\sigma,\sigma} &= -\frac{K}{\sigma^2} + \frac{2}{\sigma^2}\sum_{i=1}^{K}\frac{1+\Gamma(z_{[i]}'\beta)}{1+2\Gamma(z_{[i]}'\beta)} \\ &= \frac{1}{\sigma^2}\sum\frac{1}{1+2\Gamma(z_{[i]}'\beta)}\end{aligned}$$

References

[1] Beirlant, J. , Dierckx, G., Goegebeur, Y. , Matthys, G. (1999). Tail index estimation and an exponential regression model, Extremes 2(2), 177-200.

[2] Bertail, P. Haefke, C. , Politis, D. N., White, A. (2000). A subsampling approach to estimating the distribution of diverging statistics with applications to assessing financial market risk, en révision pour Journal of Econometrics.

[3] Bertail P. , Caillavet, F. , Nichèle, V. (1999). Consumption of Home-produced food : double hurdle analysis of french households decisions, Applied Economics, 31,1631-1640.

[4] Bingham, N.H. , Goldie, C.M. et Teugels, J. L. (1987). Regular Variation, Encyclopedia of Mathematics and its applications, Cambridge Univ. Press.

[5] Davison, A. C. and Smith, R. L. (1990). Models for Exceedances over high Thresholds, 52, 3, 393-442.

[6] Embrechts, P. , Klüppelberg, C. et Mikosch, T. (1999). Modelling Extremal Events for Insurance and Finance, Applications of Mathematics, Springer.

[7] Fischer, R. A. et Tippett, L. H. C. (1928). Limiting forms of the frequency distributions of the largest or smallest member of a sample, Proc. Camb. Phil. Soc., 24, 180-190.

[8] Gauchi, J. P. (2000). Quantitative Risk assessment of exposure to Ochratoxin A in food , Document de travail Inra Biometrie, à paraitre Risk Analysis.

[9] Feuerverger, A., Hall, P. (1999). Estimating a tail exponent by modelling departure from a Pareto distribution , Ann. Statist., 27, p. 760-781.

[10] Hill, B. M.(1975). A simple general approach to inference about the tail of a distribution . Ann. Statist. , 3, 1163-1174.

[11] Hosking, J. R. M. and Wallis , J. R. (1987). Parameter and quantile estimation for the generalized Pareto distribution, Technometrics, 29, 339-349.

[12] Pickands, J. (1975). Statistical Inference using extreme order statistics, Ann. Statist. , 3, 119-131.

[13] Politis, D.N. and Romano, J.P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *Ann. Statist.,* **22**, pp 2031-2050.

[14] Resnik, S. I. (1987). Extreme Values, Regular Variation and Point Process, Applied Probability Series, Springer.

[15] Reiss, R.D. et Thomas, M. (2001). Statistical Analysis of Extreme Values, with applications to Insurance, Finance, Hydrology and Other Fields, Birkhäuser.

[16] Smith, R.L. (1987). Estimating tails of probability distributions, Ann. Statist., 15, 1174-1207.

[17] Teugels, J. L.(1985). Extreme values in insurance mathematics, in Tiago de Oliveira, J (ED.) Statistical Extremes and Applications., 252-259. Reidel, Dordrecht.

CREST, Laboratoire de Statistiques, Timbre J340, 3, Ave Pierre Larousse,, 92245 Malakoff Cedex, France

E-mail address: Patrice.Bertail@ensae.fr

URL: http://www.crest.fr/pageperso/ls/bertail/bertail.htm