INSTITUT NATIONAL DE LA STATISTIQUE ET DES ETUDES ECONOMIQUES
Série des Documents de Travail du CREST
(Centre de Recherche en Economie et Statistique)

n° 2003-13

Evaluation of Food Risk Exposure based on Extreme Value Theory. Application to Heavy Metals from Sea Products

P. BERTAIL[1] **A. CREPET**[2]
M. FEINBERG[3] **J. TRESSOU**[4]

[1] CREST-LSA.
[2] INRA-CORELA.
[3] INRA-SIAB.
[4] INRA-CORELA.
Corresponding author : Jessica Tressou, INRA-Corela, 64-65 Boulevard de Brandebourg, 94204 Ivry sur Seine. France.
Email : jessica.tressou@ivry.inra.fr

Evaluation of food risk exposure based on Extreme Value Theory. Application to heavy metals from sea products.

Jessica Tressou*, INRA-CORELA Patrice Bertail, CREST, LSA

Max Feinberg, INRA-SIAB Amélie Crepet, INRA-CORELA

Jean Charles Leblanc, INRA

2003

*Corresponding author : Jessica Tressou, INRA Corela, 64-65, Bd de Brandebourg, 94205 Ivry sur Seine. email : Jessica.Tressou@ivry.inra.fr

Abstract

Each food may contain various amounts of some contaminants. These will not cause damage to health if levels of contamination are low and the food are not excessively consumed. This paper presents statistical methods based on extreme value theory, for evaluating risk exposures. We focus on the estimation of the probability for the exposure to exceed a fixed safe level such as Provisional Tolerable Weekly Intake (PTWI), when both consumption data and contamination data are independently available. Different calculations of risk exposure are proposed and compared. Indeed, as exposure is the product of contamination and consumption values, assumptions about the aggregation of data have a crucial role in the risk evaluation For many contaminants, PTWI belongs to the exposure tail distribution, which suggests the use of Extreme Value Theory to evaluate the risk. Our approach consists in modeling the exposure tail by a Pareto type distribution characterized by a Pareto index which may be seen as a measure of risk. Using propositions by Hall and Feuverger and Beirlant and al., we correct the bias of the usual Hill estimator to accurately estimate the risk index. We compare the results with an empirical plug-in method and show that the Pareto adjustment is relevant and efficient for low risk evaluation while the plug-in method should be used for risky contaminants. To illustrate our approach, we present some evaluations of risk exposure to heavy metals (lead, cadmium, mercury) via sea product consumption.

Résumé

Les aliments peuvent contenir, dans certaines proportions, des contaminants qui peuvent causer des problèmes de santé si l'exposition globale est trop élevée. Ce travail présente des méthodes statistiques d'évaluation des risques d'exposition basées sur la théorie des valeurs extrêmes. Nous nous intéressons à l'évaluation de la probabilité de dépasser un seuil fixé: la Dose Journalière Admissible (DJA), lorsque qu'on dispose indépendamment de données de consommation et de données de contamination. Différents type de calcul du risque sont proposés et comparés. Pour de nombreux contaminant, la DJA appartient à la queue de distribution de l'exposition suggérant l'utilisation d'outils issus de la théorie des valeurs extrêmes. Notre approche consiste à modéliser la queue de distribution par une loi de Pareto (perturbée par une fonction à variation lente) dont l'index s'interprète comme un indice de risque. Utilisant des propositions de Hall et Feuerverger, Beirlandt et al., nous corrigeons du biais de l'estimateur de Hill pour obtenir une évaluation précise du risque. Nous comparons cette méthode avec des méthodes de type empirique. Cette approche est illustrée par l'évaluation de l'exposition à des métaux lourds (plomb, Cadmium, mercure), due aux produits de la mer.

Key Words : Food Risk assessment, Extreme Value Theory, Pareto index, Heavy metals, Sea product consumption.

1 Introduction

Humans are exposed to heavy metals through out different pathways: air inhalation, drinking water, contaminated soils and contaminated food. Food sources such as fish and shellfish can become contaminated by trophic bioaccumulation. Metals are particularly toxic to children who may receive higher doses of metals from food than adults, since they consume more food relatively to their body weight than adults[1]. Some of the heavy metals like lead (Pb), mercury (Hg) and cadmium (Cd) are more dangerous for human health because of their accumulative properties. In order to describe the risk related to exposure to these heavy metals via sea products, it is necessary to separately consider lead and cadmium which are present in many other products and methylmercury (MeHg), another toxic form of mercury, which is almost exclusively present in sea products. Furthermore, for exposure to methylmercury, it will be interesting to compare children exposure to adults exposure since long term health effects could be more important for this more sensitive population.

A traditional method for the quantitative evaluation of the food consumer exposure, either for pollutants or nutrients, consists in using average composition of food items and average consumption values for a given food item or a group of items. This approach is clearly explained in reference guidelines published by FAO/WHO work group[2]. However if individual data are available, exposure calculation at individual level is recommended as it gives more accurate risk assessment, whereas hazard concerns extreme food consumers. It seems evident that risk increases when a consumer eats larger amount of a more polluted food. Moreover, the individual approach is relevant in understanding the individual behavior and the intimate structure of the food basket. In this work most attention is paid to the quantitative evaluation of the risk of exposure to contaminants, it is obvious that a similar reasoning can be used to evaluate nutrient deficiencies or, at the opposite, overexposures.

Exposure can be defined as the product of contamination and consumption data for given food items and contaminants. Global exposure is a summation of several exposure values. Due to the various data collection methods, many exposure measurements can be proposed. In this paper, the proposed probabilistic approach takes into account the whole structure of the recorded data, that is the marginal distributions of contamination and consumption data. The parameter of interest is the probability that a level of exposure, due to several food items, exceeds a given risk level. This level may be fixed a priori, for instance it can be the Provisional Tolerable Weekly Intake (PTWI) established by the Joint Expert Committee on Food Additives (JECFA) of FAO/WHO, or any adequate toxicological reference level. When dealing with risk assessment, an important issue also consists in underscoring consumer target groups, exposed to high values, due either to higher consumption or higher food contamination. Therefore, estimating the whole tail of the exposure distribution allows a better discrimination of these target groups.

Extreme Value Theory (EVT) has encountered a great success in many appli-

cation fields, such as flood or stock exchange prediction. The originality of EVT is to fully take into account the very high observed values. One criticism which is often made to this theory is that it only consists in modeling a part of the distribution, where there are a few or even no observations. Actually, this criticism may be addressed to any statistical modeling technique, since any model always brings some piece of information where there is no data. This study will demonstrate the interest and the feasibility of EVT for the consumer exposure quantitative evaluation. The principle is to model the tail of the exposure distribution by a Pareto type distribution, characterized by a Pareto index which can be interpreted as a measurement of risk. The well-known instability of the classical Hill estimator of the Pareto index may be greatly improved by using bias correction techniques introduced by Hall and Feuerverger (1999)[5]. Furthermore, this approach will be demonstrated to yield good quantification of risk of exposure. Results will be compared to a more empirical approach based on Monte-Carlo estimators of the distribution[4].

As an application, the risk exposure for lead, cadmium and methylmercury contained in sea products - fish, farmed fish, mollusk and shellfish - will be evaluated. The purpose here is not to evaluate the global food risk exposure but rather to study the risks linked to the exposure to heavy metals from sea products. These contaminants were chosen for both methodological and practical reasons. The exposure to lead and cadmium due to sea product consumption is expected to be low in comparison to the one due to all food consumption. In particular, empirical methods even tends to predict a null risk; the proposed EVT techniques allows a better extrapolation. Methylmercury is a toxic naturally occurring in fish after ingesting mercury polluted feeds. The associated risk is thus completely specific to sea product consumption: a precise evaluation of risk exposure is thus of great interest.

Section 2 describes the general framework for risk exposure assessment: definition and notation, calculus assumptions and characterization of risk. Section 3 presents the methodology based on EVT and tail estimation. Contents of section 4 is the evaluation of the risk exposure for lead, cadmium and mercury via sea product consumption and a discussion about the different methods of quantification used.

2 Exposure level calculation and risk modelling

2.1 Characterization of the risk

Chemical food risks to human health are assessed by comparing the dietary exposure with an adequate safe exposure level, such as Provisional Tolerable Weekly Intake (PTWI) proposed by the Joint FAO/WHO Expert Committee on Food Additives (JECFA). PTWI itself is defined as the estimated toxicological value of the weekly amount of a contaminant that can be ingested without appreciable risk during the lifetime. Our goal is to estimate the probability for exposure to exceed the PTWI.

One underlying hypothesis is that individuals are facing a constant level of contamination and keep the same consumption behavior over their lifetime. Moreover, it is assumed that occasional short-term excursions above the PTWI would have no major health consequences, provided that the average intake over long periods is not exceeding the PTWI; but the estimation of this long period intake is not actually possible using the available data.

If K_i is defined as the exposure value to a given contaminant for an individual i $(i = 1, ..., n)$ and assuming that exposure values are available for all individuals and expressed in the same unit as the PTWI, a simple way to estimate the risk is to use the Plug-In (PI) or empirical estimator of the probability to exceed the PTWI, defined as:

$$\frac{\#\left(K_i > PTWI\right)}{n}$$

where $\#\left(K_i > PTWI\right)$ denotes the number of exposure values that exceed the PTWI.

The results obtained with the PI estimator can be compared to those issued with the Tail Estimation (TE) method extensively described in section 3. One clear drawback of the PI estimate is that risk can not be evaluated if PTWI is too large when compared to the higher observed values (extreme tail of the empirical distribution). Thus, when risk or sample size are small, precise quantification is not possible with this method.

2.2 Assumptions for exposure calculation

Various strategies for exposure calculation can be achieved depending on the nature of the available data. A quick review will help in understanding the various assumptions made in this work.

Individual food consumption data and consumer body weight are available.

While PTWI is expressed as contaminant unit per kilogram of body weight it is interesting to know the consumer body weights. However, in many consumption survey, no such data is available and consumption recorded at the household level. Although, the information is available for this study, it is interesting to evaluate the influence of the usual technique applied to get around. In this application, ABW will denote the use of an approximated body weight of 60 kg. TBW will denote the use of the true body weight. The impact of such an approximation will be discussed.

No underlying probability distribution on consumption or contamination data is necessary.

When dealing with extreme values, any adjustment to a probability distribution, such as log-normal or exponential, is rather efficient in measuring mean

behavior but irrelevant for these data. Moreover, adjustment tests, such as Kolmogorov or χ^2, give more importance to the central tendency than to extreme values. In addition, it is not sufficient to model the distribution of each contaminated food item consumption to understand the wholesome phenomenon because it essentially depends on the correlation structure of these consumptions as some products may be complementary or substitute. Modeling the distribution of the whole vector of consumptions is generally impossible as it lies in a space of large dimension. Two kinds of calculus will be considered:

- Deterministic calculus. The contaminant concentration for each food will be expressed according to three way: (i) D-AVE the average of all available contamination data for this food; (ii) D-97.5 for the 97.5^{th} percentile and (iii) D-MAX for the maximum. In this notation, D stands for deterministic because no randomization is assumed concerning contamination data.
- Double random sampling. This other exposure evaluation method is a Monte-Carlo method[4]. It consists in randomly drawing, on one hand a consumer that is a basket of food consumption values and his associated body weight, on the other hand as many contamination values as food items in the basket. This method is denoted 2R while both consumption and contamination distributions are randomly used.

Different ways for data aggregation.

When coupling contamination and food consumption data, different levels of aggregation are possible depending on the calculus mode and the size of the data set. For small contamination data sets, it is useless to consider a large number of food items in consumption data. On the contrary, the calculation will be more accurate if each food consumption may be weighed by the correct composition data. In order to evaluate the impact of aggregation or disaggregation, two levels noted AL and DL ranging from the most and the less aggregated are considered. For example, if data are available for each fish species, the aggregated level (AL) will consists in using one value for all species; on the contrary for the disaggregated level (DL) each species value . is a food item. AL is necessary for random samplings so that composition data set is large enough. Only two aggregation levels are applied but it would be possible to defined more than that.

How censored data can be treated ?

Due to the detection or quantification limits of analytical methods, contamination data are very often left-censored. This rounding effect is related to the physical chemical phenomena involved in any analytical measurement. According to the proportion of censored data, these are usually replaced by,

either the limit of detection, or by the half of this limit or by zero. This last assumption is the less conservative. Some details on the consequences of these different assumptions are available in a study on Ochratoxin A[3]. In this application, the first assumption will be used since almost no censored data are present.

As a summary, for each exposure computation, the calculation is performed under the following assumptions:

- the aggregation level (AL or DL),
- the calculus mode (D-AVE, D-97.5, D-MAX or 2R),
- the assumption about the body weight (ABW or TBW).

Furthermore, individual consumptions are assumed to be independent and identically distributed as well as contamination data.

3 Risk exposure estimation from distribution tails

3.1 Extreme Value Theory (EVT) for risk assessment

EVT is well developed in finance and hydrology: stock exchange variation, portfolio selection; flood occurrences[6]. In these fields, extreme values are more interesting than averages because "extraordinary" events are more interesting than "ordinary". Contamination and consumption data present the same properties i.e. risk mainly concerns high consumers or highly polluted food items, which are extreme values. In order to study these values, it is necessary to understand the asymptotic behavior of the sample maximum or minimum. At the opposite lowest nutrient values are the most relevant when dealing with malnutrition. A few basic facts about EVT are now recalled.

Let $X_1, ..., X_n$ be a n-sample with cumulative distribution function (cdf) F, i.e. $F(x) = \Pr(X \leq x)$. In the following,

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

denotes the associated ordered sample so that $X_{n,n}$ is the sample maximum.

Under regularity conditions, the Fisher Tippett Theorem shows that there exists a sequence of normalization terms a_n et b_n such that

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

where W is a random variable (r.v.) with non degenerated law G.

There are only 3 possibilities for G, Gumbel, Fréchet or Weibull distributions. They can be written according to the Jenkinson representation as:

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

where limit case $\gamma \longrightarrow 0$ is Gumbel law, case $\gamma > 0$ corresponds to Fréchet law and case $\gamma < 0$ is Weibull law.

These laws are called extreme value distributions and each one corresponds to a special tail behavior: Gumbel law is related to light tailed-distribution such as normal or exponential distributions; Fréchet law to heavy-tailed distributions such as Pareto, Cauchy or Student distributions and Weibull law to finite support distributions that is for instance uniform distribution.

In the case of food risk exposure, the Fréchet type is the best adapted because exposure values can often reach very high levels and it is observable that the tails of distribution may be heavy. This approach is conservative since it is assumed that very high values are not very rare. It was decided to model the distribution tail of the exposure as a Pareto law, related to Fréchet type. In this context γ is interpreted as a risk index and many estimation methods were described. For sufficiently large x, one generally assumes that

$$1 - F(x) = Cx^{-1/\gamma}$$

where F denotes the cdf of the exposure and γ is the risk index.

As far as the PTWI is sufficiently large, the probability for exposure to exceed it would be defined by

$$C\left[PTWI\right]^{-1/\gamma}$$

Figure 1 clearly illustrates the influence of γ on thickness of the distribution tail and consequently on the risk as defined earlier.

Figure 1

3.2 Estimation of parameters

The question of fitting the distribution tail to a Pareto law consists in estimating the parameters C and γ. The first hypothesis is that for sufficiently large x, $F(x) = 1 - Cx^{-1/\gamma}$.

This notion of "sufficiently large" is quantified by selecting a fraction of the sample - i.e. the k largest observed values - and by supposing that these data are distributed according to a Pareto law. If $(X_i)_{i=1,\ldots,n}$ are independent and identically distributed (iid), conditionally to k, maximum likelihood technique allows to

estimate γ and C by:

$$\begin{cases} \gamma_{MV}(k) = H_{k,n} = \frac{1}{k}\sum_{i=1}^{k} \log \frac{X_{n-i+1,n}}{X_{n-k,n}} \\ \\ C_{MV}(k) = \frac{k}{n}\left(X_{n-k,n}\right)^{\frac{1}{H_{k,n}}} \end{cases}$$

where $H_{k,n}$ denotes Hill estimator, which is strongly biased when there is a small deviation from the exact Pareto case.

The Hill estimator is very sensitive to the choice of k as shown in Figure 2.

Figure 2

If $n \to \infty$ and $k/n \to 0$ then $H_{k,n}$ is supposed to be asymptotically gaussian with mean 0 and variance γ^2/k so that it should be expected that the Hill estimator reaches a plateau; but it is scarcely observed as already stated[6]. This behavior can be explained by the following facts: for small k the variance of the estimator is big; for large k the tail distribution is not strictly of Pareto type but rather of the form:

$$F(x) = 1 - Cx^{-1/\gamma}L(x)$$

where $L(x)$ denotes a slowing varying function satisfying for all $t > 0$, $\frac{L(tx)}{L(x)} \to 1$ as $x \to \infty$, which takes into account small deviations from the exact Pareto case. It should be stressed that any distribution with $X_{n,n}$ asymptotically Fréchet can be modelled by a distribution with a tail of this form.

One example of slowly varying function is $L(x) = 1 + Dx^{-\beta}$, with $\beta > 0$ and $D \in \mathbb{R}^*$. This form can be justified by the fact that a population (with risk exposure X) may be a mixture of two different populations with risk exposures X_1 and X_2 corresponding to risk indexes γ_1 and γ_2. This is clearly the case when considering food exposure to some contaminants within an heterogeneous population. Define

$$X = \begin{cases} X_1 \text{ with probability } p \text{ ; } X_1 \sim Pareto(C_1, \gamma_1) \\ \\ X_2 \text{ with probability } 1-p \text{ ; } X_2 \sim Pareto(C_2, \gamma_2) \end{cases} , \gamma_1 > \gamma_2$$

In that case, the resulting population has a tail of the form:

$$\Pr(X > x) = p\Pr(X_1 > x) + (1-p)\Pr(X_2 > x)$$

$$= pC_1x^{-1/\gamma_1} + (1-p)C_2x^{-1/\gamma_2}$$

which can be seen as a small deviation from the $Pareto(C, \gamma)$ distribution:

$$\Pr(X > x) = Cx^{-1/\gamma}\left[1 + Dx^{-\beta}\right]$$

with $C = pC_1$, $\gamma = \gamma_1$, $D = \frac{(1-p)C_2}{pC_1}$ and $\beta = 1/\gamma_2 - 1/\gamma_1 > 0$.

Population mixture can therefore justify the introduction of slowly varying functions. This slowly varying function induces a bias on the estimator and may strongly reduce the rate of convergence of the Hill estimator. Bias correction methods have been introduced by Hall and Feuverger[5] and Beirlant and al.[8]. The principle of the bias correction method is to interpret the Hill estimator as an estimator of the QQ plot slope perturbed by a small deviation induced by the slowly varying function. Taking the weighted average of several slopes allows to reduce the bias showing that this average behaves like an exponential r.v. with mean depending on the parameters. The technical principles about the bias correction method and about the estimation of parameters are described in Appendix A.

These estimations can be done for different values of k ($\widehat{\gamma}_k$ is the current estimator of γ) and the optimal sample fraction k^* is chosen as the solution of the program:

$$\min_{k;k>10} \frac{\widehat{\gamma}_k^2}{k} + \left(H_{k,n} - \widehat{\gamma}_k\right)^2$$

which consists in minimizing the asymptotic mean squared error ($AMSE$) of the Hill estimator. Figure 3 gives an example of bias correction.

Figure 3

As explained above, the exposure risk indicator is the probability for risk exposure to exceed the PTWI; that is according to our model:

$$C\left[PTWI\right]^{-1/\gamma}$$

We will thus use

$$\widehat{\gamma^*} = \widehat{\gamma}_{k^*}$$

that is the "unbiased" Hill estimator taken at the optimal sample fraction k^*, and

$$\widehat{C^*} = \frac{k^*}{n}\left(X_{n-k^*,n}\right)^{\frac{1}{\widehat{\gamma}_{k^*}}}$$

the resulting estimator of constant C.

Thus the estimated risk will be:

$$\widehat{C^*}\left[PTWI\right]^{-1/\widehat{\gamma^*}}$$

4 Exposure to heavy metals for sea product consumers

4.1 Data description

Food consumption data

Consumption data came from the French survey INCA[11] which concerns the food consumption of 3003 individuals of 3 years and more. This survey concerns all consumptions at home or outside, during one week. Besides of a detailed food nomenclature of about 900 food items clustered in 45 groups, individual socio-demographic data are available, including the individual body weight and age.

Among this food list, 92 food items containing fish or sea products were found in the groups "Fish", "Shellfish and Mollusk", "Mixed dishes", "Meat products" and "Soups". For some of these items, such as breaded fish, consumption data were weighed by a recipe factor. The operational study file contained the properly weighed consumption values for 92 products and $n = 2513$ sea product consumers, including socio-demographic information. As contamination data were clustered into three categories ("Fish", "Farmed fish" and "Mollusks and shellfish"), each of the 92 food items was linked to one of these categories. This leads to two levels of aggregation which are noted as:

- DL: Disaggregated Level, C_j^i is the consumption of product j for sea product consumer i, with j varying from 1 to 92. .
- AL: Agregated Level, $C_{(j)}^i$ is the consumption of product from category (j) for consumer i, with (j) being "Fish", "Farmed fish" or "Mollusks and shellfish".

So that a consumer is more generally defined by C^i, a 92-dimensional or a 3-dimensional vector and his body weight w^i for i varying from 1 to n.

Contamination data

Sea product contamination data were collected through different analytical surveys performed by several French institutions (MAAPAR[12], IFREMER[13]) during the period 1994-2002. For each of the three studied contaminants (Pb, Cd and Hg), there were respectively 2995, 2641, and 3194 contamination values expressed in mg per kg of fresh weight. These values were clustered into three categories ("Fish", "Farmed fish" and "Mollusks and shellfish") according to their contamination level. Thus, calculus of exposure was possible for the AL level, deterministic method and 2R. Concerning the DL level, it was

necessary to look up all analytical data in order to associate a value to each 92 food items. For instance, for "Fried sole" or "Vapor sole" all the contamination data concerning sole were used to calculate average or maximum, while for vaguer named items, such as "Fish soup" or "Fried fish", all data from clusters "Fish" and "Farmed fish" were taken.

According to Claisse and al.[14], methylmercury in sea products can be extrapolated from mercury contents. Therefore, conversion factors were applied to analytical data in order to get the corresponding methylmercury (MeHg) concentration in food: 0.84 for fishes, 0.43 for mollusks and 0.36 for shellfish.

International toxicological references (PTWI)

The toxicological limit to be used were established and revised by the JECFA. The most recent references were used for this study and were:

- Lead, 25 μg/week/kg b.w. (revision 1999[15]),
- Cadmium, 7 μg/week/kg b.w. (revision 2000[16]),
- Methylmercury, 3.3 μg/week/kg b.w. (revision 1999[15]).

4.2 Results and discussion

Results for food risk exposure to lead (Pb), cadmium (Cd) and methylmercury (MeHg) are given in table 1. Each line of this table corresponds to a different calculation of exposure for a given contaminant according to the proposed assumptions leading to 18 scenarios. For example, for scenario 1, the exposure to lead from sea products is described by its average, its 97.5^{th} percentile and its maximum over the sea product consumers. This first scenario corresponds to a calculation with a deterministic calculus at disaggregated level (DL) using average of contamination (D-AVE) and true body weight (TBW). The last columns give the associated risks calculated with our new method based on tail estimation (TE) and the Plug-In method (PI). For 2R calculus mode, size of random samplings was 10,000 and the presented results correspond to the mean results obtained after 100 repetitions of the same calculus.

Table 1

This table does not present the results with approximated body weight (ABW): this approximation leads to a systematic under-evaluation of the exposure. Exposure is about 1,2 times lower in the case ABW. For example, average exposure to Pb (scenario 1) varies from 0.325 μg/week/kg b.w. for TBW to 0.304 μg/week/kg b.w. for ABW. Indeed assuming that all individuals have a 60kg body weight is very imprecise since there are children in the sample. However, the calculus was made in

order to show the impact of such vague approximations. In household consumption surveys, two approximations have to be made: first to get the individual consumptions from the total household consumptions and then to approximate the weighed consumptions. If body weight is not available, it is convenient to use different body weights according to the sex and age of individuals.

An important remark concerns the significance of all these results. This risk evaluation of exposure to heavy metals was made on effective sea product consumers in the INCA data. A multiplicative coefficient of $2513/3003 = 84\%$ may be applied to risk calculated with PI to take into account the non consumers in order to extrapolate to the whole population. However, because of the short period of the survey, the bias due to the observed zeros is well known: individuals with null consumptions in INCA may be true non-consumers of sea products or may infrequently consume sea products, maybe in large quantity, but not during the survey week. This bias that we can evaluate with other sources on household consumptions (Secodip) is not significant in the case of sea products.

Our main observations are :

- the aggregation level assumption has a high impact on the results. DL gives lower exposure levels and lower risks than AL for all contaminant at a given calculus mode. For example, for Pb, comparison of average exposures of scenarios 1 and 3 show the importance of aggregation. This can be explained by the fact that the mean contamination of DL is lower than the mean contamination of AL. Indeed, under AL assumption, averages are taken over a larger number of observations and high values boost the average of contamination. For example , average of contamination for tuna is higher than for any other fish but in AL, all fish are assumed to be contaminated at the average level of all fishes which is higher because of tuna. However, 2R calculus is not possible for DL assumption since there is not be enough data to sample in.

- At a given aggregation level, D-AVE et 2R give similar results in average but randomization of contamination of calculus 2R allows to reach higher exposure levels so that 97.5^{th} percentile and maximum are respectively higher for 2R than for D-AVE. Likewise, risk is higher for 2R than for D-AVE (see scenarios 9 and 12, similar averages but different maximum and risks). Indeed, if high consumptions are associated to high levels of contamination, some exposure may be very high and 2R allows to consider them without using an unrealistic assumption such as D-MAX or D-97.5. These two last methods are not realist but present the advantage to be conservative. Indeed if D-MAX or D97.5 gives null risks or negligible risks, there is no need to be more accurate in the process.

- PI methods gives null risk for D-AVE calculus for Cd and Pb (see lines 1, 3, 7 and 9). For Pb, D-97.5 also give null risk with PI method. Exposure to these two contaminants is not high enough to be risky, which is pleasant since these

metals are contained in other food. However, a null risk does not exist and the TE method allows a much more accurate quantification.

- TE mostly gives a higher risk evaluation than PI and differences are sometimes very important (for example, scenario 18, risk varies from about 25% to 10%). Moreover, TE sometimes does not allow to evaluate risk when it is important (see scenarios 14, 16 and 17). As it is shown in the graphic illustration of Figure 4, if the $PTWI$ is far from the distribution tail (that is too low compared to data), it is not in the support of Pareto distribution so that no risk evaluation is possible. Indeed, Pareto cdf is defined for $x \geq a$, where a is such that $F(a) = 0$, i.e. $Ca^{-1/\gamma} = 1 \Rightarrow a = C^{\gamma}$. Therefore, if $PTWI < a$, the probability to exceed the $PTWI$ is not defined and the PI method will be used. Furthermore, if the $PTWI$ is too close to a, risk evaluation may be too high (it may be the case for scenario 15). As a summary about the TE method, a good risk evaluation is obtained if the PTWI belongs to the distribution tail ($PTWI_3$ in the illustration); an irrelevant risk evaluation ($PTWI_2$ in the illustration) or no risk evaluation ($PTWI_1$ in the illustration) on the opposite case.

Figure 4

Results concerning MeHg according to the age of the population are presented in Table 3. Risk was calculated according to PI method since PTWI may not belong to the distribution tail, i.e. risk is too high to use TE. Three calculus scenarios are presented: DL D-AVE, AL D-AVE and AL 2R. For 2R calculus mode, size of random sampling was 5,000 and the presented results correspond to the mean results obtained after 100 repetitions of the same calculus.

Table 2

The role of the aggregation level is even more important in this case for all population groups and especially for 3-8 year old children, where risk varies from 4% (DL) to 25% (AL) for D-AVE calculus mode. However, it is clear that according to this data, exposure of children (aged 3 to 8) is systematically higher than the exposure of the rest of the population. As D-AVE calculus is concerned, contamination is the same for all individuals so that the observed differences are due to the consumption behaviors. Children eat more sea products relatively to their body weight than the rest of the population. To be more accurate, confidence intervals for PI risks are currently being constructed thanks to the use of incomplete U-Statistics. Our first results show that the observed differences according to the population age are significative. About the characterization of risky population, developments are needed as suggested in Bertail(2002)[10]

5 Conclusions

This paper leads to two types of conclusions: first, about method efficiency and then, about exposure to heavy metals for sea product consumers.

It is important to note that assumptions of calculus (such as levels of aggregation used to couple data, calculus mode, body weight approximation,...) have a strong impact on the values of exposure so that one must not use numerical results without detailing all the used assumptions. Indeed, aggregation of data and body weight approximation lead to under-evaluation of risk. Furthermore, our definition of risk is based on PTWI, the definition of which is based on a tolerable intake over lifetime. Thus, results should be nuanced as the data and methods do not take into account chronic consumptions or long-term contamination levels.

Concerning the feasibility of our method based on tail estimation, it is important to check whether the studied contaminant is risky or not. Indeed, if the PTWI does not belong to the distribution tail, Pareto tail adjustment is useless while, on the opposite case, it allows to accurately quantify very low risk. Developments are needed concerning confidence interval for such probabilities to exceed a given toxicological level.

As far as food risk is concerned, according to the data used and by comparison to the PTWI, methylmercury intake via the consumption of sea product seems important for a significant part of the population, above all children. Our evaluation method of the risks for lead and cadmium intakes are clearly more satisfying than the usual methods which tend to under-estimate the risk.

References

[1] Risk Assessment of Priority Chemicals, WHO (World Health Organisation); IPCS (International Program for Chemical Safety); EHCs (Environmental Health Criteria), Publications about all chemicals and contaminants available at *http://www.who.int/pcs/pubs/pub_ehc_alph.htm.*

[2] FAO/WHO. *Food consumption and exposure assessment of chemicals,* 1997, 10-14 February, Report of a FAO/WHO Consultation; Geneva, Switzerland.

[3] Tressou, J., Leblanc, J.C., Feinberg, M., and Bertail, P., (2002) *Evaluation du risque alimentaire lié à l'Ochratoxine A, Contribution du vin et des produits à base de vin*, Rapport interne INRA ONIVINS

[4] Gauchi, J.P. and Leblanc J.C., (2002), Quantitative Assessment of Exposure to the Mycotoxin Ochratoxin A in Food, *Risk Analysis, Vol. 22, No. 2.*

[5] Feuerverger, A. and Hall, P., (1999), Estimating a tail exponent by modelling departure from a Pareto Distribution, *Annals of Statistics, Vol. 27, No. 2.*

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

[7] Drees, H. and Kaufmann E., (1998), Selecting the optimal sample fraction in univariate extreme value estimation, *Stochastic Processes and their Applications 75 pp149-172*

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

[9] Pyke, R. (1965) Spacings, *Journal of the Royal Statistis Society, Series B (Methodological), Volume 27, Issue 3.*

[10] Bertail, P., (2002) *Evaluation des risques d'exposition à un contaminant: quelques outils statistiques*. Document de travail, CREST, $n°2002-39$.

[11] Enquête INCA (1999) CREDOC-AFFSA-DGAL, *Enquête nationale sur les consommations alimentaires*, Tech & Doc Lavoisier, Coordinateur: J.L Volatier.

[12] MAAPAR, Ministère de l'Agriculture, de l'Alimentation , de la Pêche et des Affaires Rurales, *Résultats des plans de surveillance pour les produits de la mer*, 1998-2002.

[13] IFREMER, *Résultat du réseau national d'observation de la qualité du milieu marin pour les mollusques (RNO)* 1994-1998.

[14] Claisse, D.Cossa, D., Bretaudeau-Sanjuan, G., Touchard, G., and Bombled, B. (2001), *Methylmercury in molluscs along the French coast*, Marine pollution bulletin, Vol. 42, No. 4, pp. 329-332.

[15] FAO/WHO, *Evaluation of certain food additives and contaminants* _for lead and methylmercury_, Fifty third report of the Joint FAO/WHO Expert Committee on Food Additives, WHO technical report series n°896, Geneva, 1999.

[16] FAO/WHO, *Evaluation of certain food additives and contaminants* _for cadmium_, Fifty fifth report of the Joint FAO/WHO Expert Committee on Food Additives, WHO technical report series n°901, Geneva, 2000.

[17] Tressou, J. (2002). Méthodes statististiques pour l'évaluation des risques, le cas de l'Ochratoxine A. Preprint, INRA.

A Appendix: Bias correction method

Suppose that data are distributed according to the cdf

$$F(x) = 1 - Cx^{-1/\gamma}\left(1 + Dx^{-\beta}\right)$$

Then, the generalized inverse cdf[1] is defined, for small y, by

$$F^{\leftarrow}(1-y) = \left(\frac{y}{C}\right)^{-\gamma}(1+\delta_2(y)) \approx \left(\frac{y}{C}\right)^{-\gamma}\exp(\delta_2(y)) \tag{1}$$

with $\delta_2(y) = \gamma C^{-\beta_1} D y^{\beta_1}$, where $\beta_1 = \gamma\beta$.

General results on order statistics give:

$$\log X_{n-i+1,n} = \log\left(F^{\leftarrow}(U_{n-i+1,n})\right) = \log\left(F^{\leftarrow}(1-U_{i,n})\right) \tag{2}$$

where $U_{i,n}$ is the order statistic of a uniform r.v. upon $[0,1]$, which can approximated by $\frac{i}{n+1}$ (its expectation).

For any small y, applying log to (1) gives:

$$\log\left(F^{\leftarrow}(1-y)\right) = -\gamma\log y + C_1 + \delta_2(y) \qquad \text{where } C_1 = \gamma\log C. \tag{3}$$

Using (2) and (3), we get for small values of i (that is $i = 1, ..., k$) :

$$\log X_{n-i+1,n} = -\gamma\log U_{i,n} + C_1 + \delta_2(U_{i,n}) \tag{4}$$

Let us recall Renyi representation[9] introducing the r.v.:

$$\forall i = 1, ..., n, \qquad T_{n-i+1,n} = \sum_{j=1}^{n-i+1}\frac{E_j}{n-j+1}$$

where $(E_j)_{j=1,...,n}$ is an exponential r.v. with mean 1 and $E_{i,n}$ denote the associated i^{th} order statistics.

Thanks to Renyi representation, we have the two following results:

$$T_{n-i+1,n} = \sum_{j=1}^{n-i+1}\left(E_{j,n} - E_{j-1,n}\right) \overset{\mathcal{L}}{=} E_{n-i+1,n} \tag{5}$$

and

$$-\log U_{i,n} = -\log\left(1 - U_{n-i+1,n}\right) = T_{n-i+1,n} \tag{6}$$

Then let us introduce $Z_i = i(\log(X_{n-i+1,n}) - \log(X_{n-i,n}))$.

Using (4) and Renyi representation (equations (5) and (6)), we have:

$$Z_i = i\gamma(T_{n-i+1} - T_{n-i}) + i\left[\delta_2\left(\exp(-T_{n-i+1,n})\right) - \delta_2\left(\exp(-T_{n-i,n})\right)\right] \tag{7}$$

[1] The strict definition of which is $F^{\leftarrow}(x) = \inf\{y \in R, F(y) \geqslant x\}$.

Noting $\delta_3(z) = \delta_2(\exp(-z))$, a Taylor expansion leads to:

$$\delta_3(T_{n-i+1,n}) - \delta_3(T_{n-i,n}) \simeq (T_{n-i+1,n} - T_{n-i,n})\delta_3'(T_{n-i,n}) \tag{8}$$

But we have $T_{n-i,n} \overset{\mathcal{L}}{=} -\log(U_{i+1,n}) \simeq log\frac{n+1}{i+1} \simeq log\frac{n}{i}$ and $\delta_3'(logx) = -(x)\,\delta_2'(x)$, so that:

$$\delta_3'(T_{n-i,n}) = \delta_3'\left(log\frac{n}{i}\right) = -\left(\frac{i}{n}\right)\delta_2'\left(\frac{i}{n}\right) \tag{9}$$

$$= \gamma\beta_1 C^{-\beta_1} D\left(\frac{i}{n}\right)^{\beta_1} = \gamma\delta_1\left(\frac{i}{n}\right) \tag{10}$$

where $\delta_1(x) = D_1 x^{\beta_1}$ with $D_1 = -\beta_1 C^{-\beta_1} D$.

Combining equations (7), (8) and (9) and the fact that $T_{n-i+1,n} - T_{n-i,n} = \frac{E_{n-i+1}}{i}$ (according to the definition of r.v. $T_{n-i+1,n}$), we finally deduce that for $i = 1, ..., k$:

$$Z_i \approx E_{n-i+1}\gamma\left[1 + \delta_1\left(\frac{i}{n}\right)\right] \approx E_{n-i+1}\gamma\exp\left[\delta_1\left(\frac{i}{n}\right)\right]$$

This means that the weighted difference of log spacings $Z_i = i(\log(X_{n-i+1,n}) - \log(X_{n-i,n}))$ behaves like as an exponential r.v. with mean $\gamma\exp\left[\delta_1\left(\frac{i}{n}\right)\right]$ depending on the parameters γ, β_1 and D_1.

These parameters can be estimated by maximum likelihood (MV) or by least square (LS) considering as a dependent variable $\log Z_i$. These estimations can be done for different values of k (we thus get $\widehat{\gamma}_k$, $\widehat{\beta_{1_k}}$ and $\widehat{D_{1_k}}$ estimators of γ, β_1 and D_1) and the optimal sample fraction k^* is chosen as the solution of the program:

$$\min_{k;k>10} \frac{\widehat{\gamma}_k^2}{k} + (H_{k,n} - \widehat{\gamma}_k)^2$$

which consists in minimizing the mean squared error of the Hill estimator.

Practically, as advised in many papers dealing with this kind of models, we fix the value of $\beta_1 = \gamma\beta$ to 1 (see Drees and Kaufmann (1998)[7]) and used MV resolution to get estimation of γ and $D_1 = -D/C$. LS resolution is not possible if some of the Z_i are null which is the case as soon as two individuals have the same exposure (frequent case in our deterministic framework). These techniques were tested on simulated data in Tressou (2002)[17].

As explained before our risk indicator is the probability for risk exposure to exceed the PTWI that is according to our model:

$$C[PTWI]^{-1/\gamma} \tag{11}$$

or if slowly varying function is considered

$$C[PTWI]^{-1/\gamma}(1 + D_1[PTWI]^{\beta_1}) \sim C[PTWI]^{-1/\gamma} - D[PTWI]^{\frac{\gamma-1}{\gamma}} \tag{12}$$

We will thus use

$$\widehat{\gamma^*} = \widehat{\gamma}_{k^*} \text{ for } \gamma,$$

$$\widehat{C^*} = \frac{k^*}{n} \left(X_{n-k^*,n}\right)^{\frac{1}{\widehat{\gamma}_{k^*}}} \text{ for } C,$$

$$\text{and } \widehat{D^*} = -\widehat{D_{1_{k^*}}} \times \widehat{C^*} \text{ for } D$$

Comparing estimation (11) and (12) we remark that the two approximations were equivalent.

List of Figures :

List of Tables :

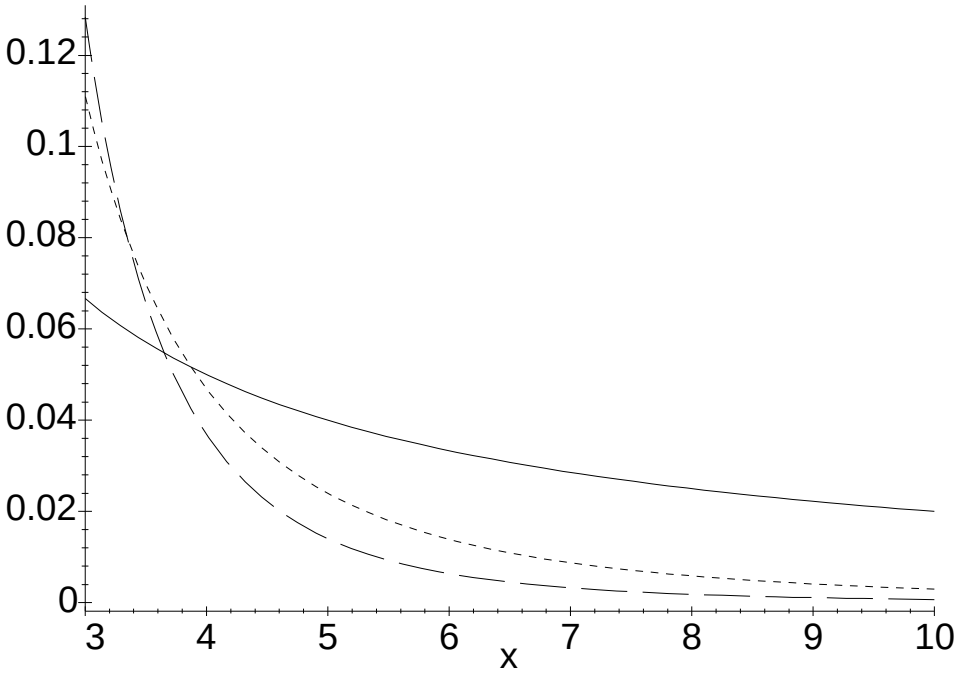

Figure 1: Pareto Distribution tail for different values of $\gamma : \gamma = 1$ (Solid), $\gamma = 0.5$ (Dots) and $\gamma = 0.3$ (Dash)

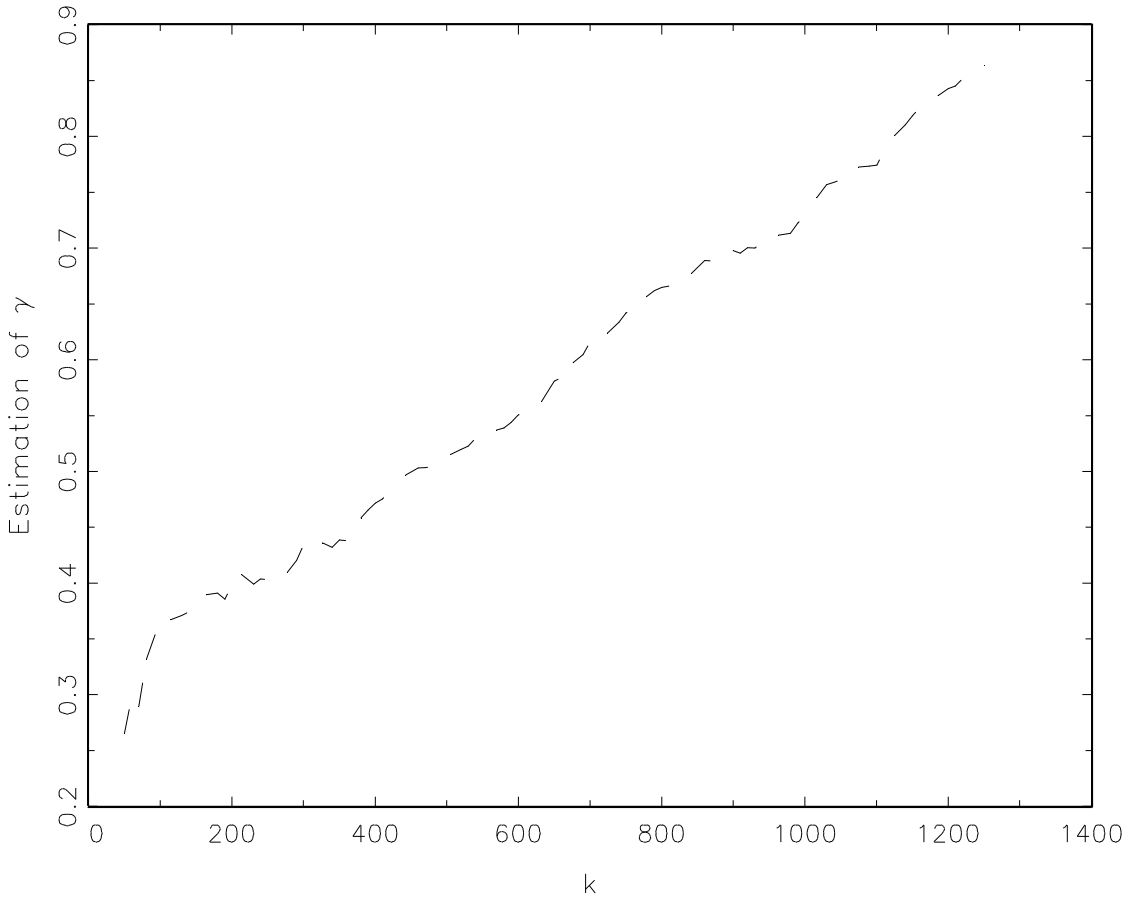

Figure 2: Hill estimator of the risk index γ for different values of k.
Case of the exposure to lead, Disaggregated data, Average contamination.

Scenario	Contaminant	Assumptions		Exposure (in µg / week / kg b.w.)			Associated Risk	
		Aggregation level	Calculus mode	Average	97,5th percentile	Maximum	TE	PI
1	Pb	DL	**D-AVE**	**0.325**	1.406	5.143	3.17E-07	**0**
2			**D-MAX**	3.847	15.239	36.239	3.76E-03	4.78E-03
3		AL	**D-AVE**	**0.387**	1.774	7.735	2.90E-06	**0**
4			**D-97.5**	1.290	6.176	26.776	2.20E-04	**0**
5			**D-MAX**	6.392	23.095	93.934	1.67%	1.07%
6			**2R**	**0.386**	2.096	21.725	1.03E-04	2.60E-05
7	Cd	DL	**D-AVE**	0.199	1.061	3.537	7.41E-05	0
8			**D-MAX**	2.592	13.200	32.080	10.94%	9.15%
9		AL	**D-AVE**	**0.235**	1.211	**5.434**	7.54E-05	0
10			**D-97.5**	0.780	4.054	18.132	4.10E-03	1.99E-03
11			**D-MAX**	4.694	20.763	90.021	---	16.95%
12			**2R**	**0.234**	1.422	**19.391**	7.92E-04	7.97E-04
13	MeHg	DL	**D-AVE**	0.628	2.712	17.213	**1.46%**	**1.71%**
14			**D-MAX**	9.167	39.989	110.486	**---**	64.27%
15		AL	**D-AVE**	1.337	5.031	12.852	23.61%	2.98%
16			**D-97.5**	5.757	21.783	55.666	**---**	50.78%
17			**D-MAX**	19.242	72.519	185.514	**---**	87.43%
18			**2R**	1.339	7.513	74.644	**24.96%**	**9.52%**

Table 1: Food risk exposure to Lead (Pb), Cadmium (Cd) and Methylmercury (MeHg) for sea product consumers

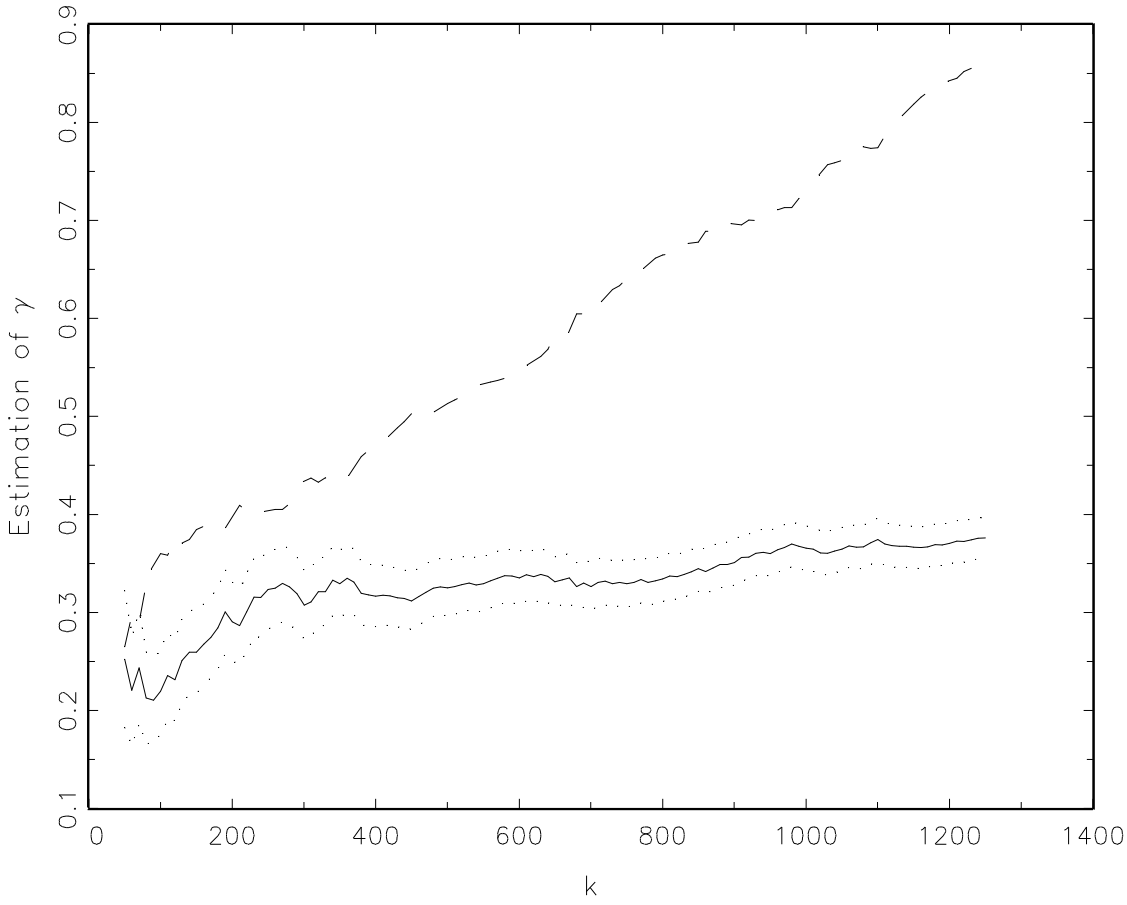

Figure 3: Example of bias correction for the risk index γ: Hill estimator (dashed line), debiased Hill estimator (solid line) and confidence interval for the Debiased estimator (dots);

The minimization of $AMSE$ gives $k^* = 50$, $\widehat{\gamma}^* = 0,252$ and $H_{k^*,n} = 0,265$.

Case of the exposure to lead, Disaggregated data, Average contamination.

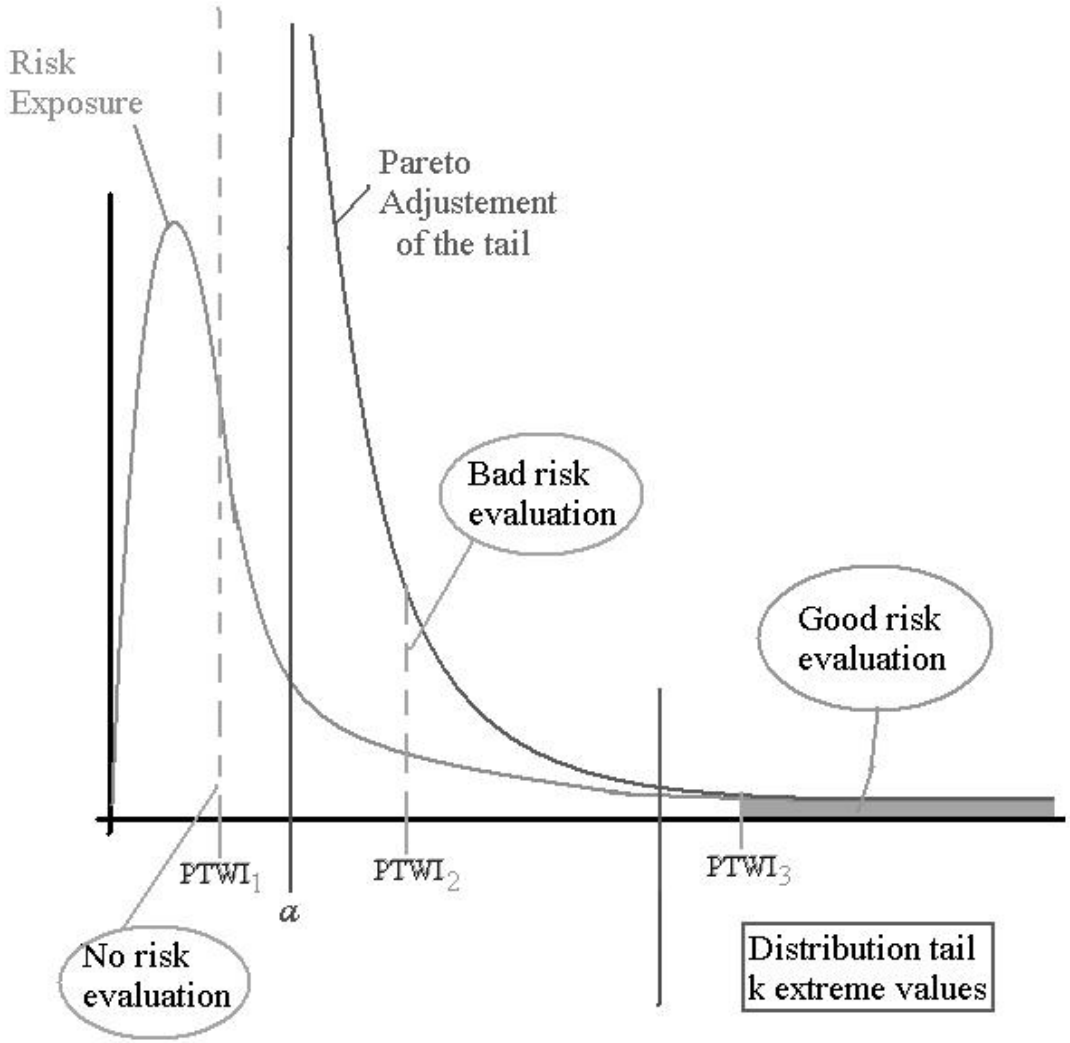

Figure 4: Pareto Adjustement and risk evaluation.

Assumptions		Risk associated to Exposure (TBW)				
Aggregation level	**Calculus mode**	**3-8 years old**	**9-15 years old**	**16-60 years old**	**over 60 years old**	**All sea product consumers**
DL	**D-AVE**	**4.09%**	1.60%	1.25%	0.56%	1.71%
AL	**D-AVE**	**25.45%**	6.18%	2.89%	4.21%	2.98%
	2R	**19.51%**	10.43%	6.39%	7.39%	9.52%

Table 2: Risk exposure to Methylmercury for sea product consumers according to age (Method of risk evaluation: PI)