

---

## EXPOSITION AUX RISQUES ALIMENTAIRES ET PROCESSUS STOCHASTIQUES : LE CAS DES CONTAMINANTS CHIMIQUES

Stéphan Clémenton & Jessica Tressou

---

**Résumé.** — A l'instar d'autres domaines tels que la sécurité nucléaire, la finance ou l'assurance pour lesquels l'analyse des risques est un enjeu essentiel, la sécurité alimentaire est aujourd'hui un champ d'application naturel des outils de la modélisation probabiliste et des méthodes statistiques. La première partie de cet article décrit ainsi les premiers travaux, conduits principalement au sein de l'unité "Methodologies d'analyse de risque alimentaire" de l'INRA, visant à formuler en termes mathématiques les approches statiques utilisées au niveau international pour l'analyse des risques alimentaires, domaine où le recours à la modélisation est relativement récent. L'exposition au risque alimentaire chimique y est décrite par une variable aléatoire, représentant la dose de contaminant ingérée par un individu représentatif d'une population donnée au cours d'un repas, d'une journée ou d'une semaine. L'analyse statistique du risque afférent à ce type de contamination consiste alors à inférer la distribution de cette variable aléatoire, ou certaines de ses caractéristiques telles que sa queue, décrivant la probabilité d'occurrence d'événements particulièrement dangereux. Pour certains contaminants, dits "à cinétique longue", l'effet d'accumulation au sein de l'organisme résultant des prises alimentaires successives, conduit à introduire le temps comme paramètre descriptif essentiel du phénomène d'exposition, lequel est alors décrit non plus par une variable aléatoire, mais par un *processus stochastique*. Nous présentons ici un modèle markovien parcimonieux permettant de décrire le processus d'exposition au risque de contamination au méthyle-mercure par voie alimentaire, évoluant par sauts au moment des prises alimentaires et de façon déterministe entre les prises selon la pharmacocinétique propre au contaminant, et d'évaluer les risques toxicologiques afférents au moyen de procédures statistiques adaptées, fondées sur les données disponibles en pratique. Nous nous attacherons à expliquer pourquoi la dimension temporelle de ce type de modèle nous oblige à revisiter le problème de la construction d'indicateurs de risque pertinents. En particulier, la description du comportement extrême de l'exposition fait appel à des concepts tout à fait spécifiques aux séries temporelles.

---

**Mots clefs.** — Risque alimentaire, dose hebdomadaire tolérable, valeurs extrêmes,  $U$ -statistique, simulation Monte-Carlo, processus markovien, renouvellement, méthyle-mercure.

**Abstract.** — Similarly to other domains such as nuclear safety, finance or insurance for which risk analysis is crucial, food safety is now a natural application field for probabilistic tools and statistical methods. The first part of the article is devoted to the description of early works, mainly conducted at INRA-Met@risk Methodologies for food risk analysis, aiming at a mathematical formulation of static approaches used at the international level for food risk analyses. Food chemical exposure is described as a random variable representing the contaminant intake of an individual from a given population over a meal, a day or a week. The statistical analysis of the associated risk consists of inferring on the distribution of this random variable, with special interest on the tail of the distribution where the most hazardous situations occur. For contaminants with slow kinetics, successive intakes result in an accumulation so that the integration of time as a parameter of the exposure phenomenon becomes crucial. The random variable is then replaced with a *stochastic process*. A parsimonious markovian model describing the exposure process is introduced. In this modeling, the temporal evolution of the contamination exposure is entirely determined by the accumulation phenomenon due to successive dietary intakes and the pharmacokinetics governing the elimination process in between intakes. Statistical techniques adapted to the available data and the dynamic model are described. We also explain why the introduction of a time dimension requires the construction of relevant risk indicators relying mainly on the analysis of the extremal behavior of the process.

**Keywords :** Food risk, provisional tolerable weekly intake, extreme values,  $U$ -statistic, Monte-Carlo simulation, markovian process, renewal, methylmercury

## Table des matières

1. Introduction.....	2
2. Evaluation du risque alimentaire : approches statiques.....	6
3. L'exposition vue comme un processus stochastique.....	11
4. Discussion et perspectives de recherche.....	24
Remerciements.....	25
Références.....	26

## 1. Introduction

L'évaluation des risques alimentaires est un domaine d'application relativement nouveau pour les statisticiens. Cette thématique trouve depuis peu sa place dans les congrès internationaux de Statistique, ainsi qu'en attestent les sessions organisées aux 38èmes Journées de la Société Française de Statistique (Clamart, 2006) et au 25ème Congrès Européen de Statistique (Oslo, 2005, session "Statistics in environmental and food sciences", <http://www.ems2005>).

no). Elle constitue également l'une des sept priorités du 7ème Programme Cadre de Recherche et Développement européen, <http://www.telecom.gouv.fr/programmes/7pcrd>. Plusieurs unités de recherches pluridisciplinaires entièrement dédiées à l'analyse de risque alimentaire, comme l'unité INRA Mét@risk en France (INRA, <http://www.paris.inra.fr/metarisk>), le centre de recherche Biometris aux Pays Bas ([www.biometris.nl](http://www.biometris.nl)) ou encore l'Institut de Sécurité Alimentaire et de Nutrition Appliquée aux Etats-Unis ([www.foodrisk.org](http://www.foodrisk.org)), ont récemment été créées.

Une analyse de risque alimentaire vise en premier lieu à déterminer si une substance donnée peut poser un problème de santé publique. Le cas échéant, il s'agit ensuite de caractériser les individus les plus exposés, puis d'élaborer les moyens de réduction du risque les plus efficaces et de mettre en oeuvre éventuellement certaines mesures de sécurité sanitaire, [26]. Au cours d'une telle analyse, les compétences du statisticien sont susceptibles d'intervenir à diverses occasions, afin de quantifier à partir de données expérimentales ou d'enquête une grande variété de phénomènes liés au risque alimentaire, qu'ils soient de nature biologique ou comportemental, et l'incertitude y afférant, [23]. De nombreux modèles statistiques de croissance bactérienne ont ainsi été développés dans le cadre de la *microbiologie prévisionnelle*, [39]. La *modélisation dose-réponse* vise à construire un modèle de régression afin d'expliquer l'amplitude d'une réponse biologique de l'organisme en fonction d'une dose d'exposition à une substance, un composé chimique en général, susceptible d'engendrer des effets néfastes sur la santé, [18]. Des modèles économétriques peuvent également être utilisés afin de décrire la demande en biens alimentaires, [21]. Enfin, l'épidémiologie mathématique permet de mettre en évidence le lien entre une exposition significative et le développement d'une maladie ou l'occurrence d'un effet spécifique, [14].

Plus formellement, l'analyse de risque, telle que les comités d'experts<sup>(1)</sup> et la FAO, <http://www.fao.org>, la définissent, se décompose en trois étapes.

1. **L'appréciation du risque.** Cette étape consiste à identifier les événements dangereux, estimer leur probabilité d'occurrence et mesurer l'importance des effets néfastes sur la santé humaine.

<sup>(1)</sup>Citons pour la France, l'Agence Française de Sécurité Sanitaire des Aliments (Afssa) ; pour l'Union Européenne, l'Autorité européenne de sécurité des aliments (EFSA pour European Food Safety Authority) et les comités internationaux d'experts appelés par la commission Codex Alimentarius, l'Organisation des Nations Unies pour l'Alimentation et l'Agriculture (FAO de l'anglais pour Food Agricultural Organization) et l'organisation mondiale de la santé (OMS ou WHO de l'anglais pour World Health Organization) : le JECFA (Joint FAO/WHO Expert Committee on Food Additives and contaminants) qui traite les risques liés aux additifs et aux contaminants chimiques, le JMPR (Joint FAO/WHO Meetings on Pesticide Residues) qui évalue le risque lié aux résidus de pesticides et le JEMRA (Joint FAO/WHO Meetings on Microbiological Risk Assessment) qui traite le risque microbiologique.

2. **La gestion du risque.** Il s'agit d'identifier les différentes mesures de diminution du risque préalablement apprécié et de quantifier, en incluant les incertitudes afférentes, la réduction de risque selon chaque scénario afin de déterminer des solutions jugées acceptables. Ces mesures peuvent prendre plusieurs formes : introduction de teneurs maximales en contaminant sur certains aliments, retrait du marché de certaines denrées, recommandations nutritionnelles... Dans ce cadre, les impacts économiques de telles mesures sont étudiés et mis en balance avec les réductions de risque attendues.
3. **La communication sur le risque.** Elle peut s'appliquer à tout moment de l'analyse de risque entre les responsables de l'estimation du risque, les responsables de la gestion du risque et les autres parties intéressées (milieux professionnels, consommateurs).

L'appréciation du risque, aussi appelée évaluation du risque, a fait l'objet d'un numéro spécial de *Food and Chemical Toxicology* (Vol. 40, n° 2 et 3, mars 2002) auquel le lecteur pourra se référer pour une description plus détaillée. Elle requiert de traiter les questions suivantes.

- **Identification du danger [4] et caractérisation du danger [22].** Il s'agit d'identifier les couples aliments-pathogènes pour lesquels existe un danger, *i.e.* pouvant provoquer des effets néfastes sur la santé, et d'étudier les mécanismes d'action du toxique ainsi que sa cinétique dans l'organisme (absorption, métabolisme et élimination). Ceci requiert des techniques de toxicologies *in vitro* ou *in vivo* chez l'animal. Il en résulte des relations dose-réponse entre la dose ingérée et le ou les effets néfastes considérés ou plus simplement des doses tolérables par l'organisme, d'abord pour l'animal puis pour l'homme.
- **Evaluation de l'exposition [35] et caractérisation du risque [45].** Le but est ici de quantifier l'exposition des individus d'une population donnée à l'agent pathogène étudié sur une période suffisamment longue en comparaison des effets étudiés. Il s'agit donc d'évaluer la consommation des aliments incriminés et leur contamination pour estimer l'exposition puis de comparer l'exposition aux doses tolérables ou relations "dose-réponse" obtenues dans l'étape de caractérisation du danger.

C'est à cette dernière étape que nous nous intéresserons principalement dans cet article en nous concentrant sur l'évaluation du risque lié à la présence de contaminants chimiques dont la toxicité est avérée et chronique. Le danger est dans ce cas beaucoup plus insidieux puisque c'est l'exposition chronique, *i.e.* sur une période très longue, qui peut avoir des effets néfastes sur la santé des individus. Plus précisément, pour chaque contaminant chimique susceptible d'avoir ce type d'effet, les médecins et toxicologues déterminent une dose

tolérable par l'organisme humain, le plus souvent à partir d'études expérimentales chez l'animal, [22] : si cette dose est dépassée tout au long de la vie ou du moins sur une longue période, l'individu est considéré comme un sujet à risque. Cette dose est appelée Dose Journalière Tolérable (DJT) ou Dose Hebdomadaire Tolérable (DHT) selon la période considérée et est exprimée relativement au poids corporel de l'individu.

Contrairement aux risques microbiologiques pour lesquels des modèles temporels de croissance bactérienne ont été développés, la plupart des modèles étudiés ces dernières années dans l'analyse de risques chroniques liés à la présence de contaminants chimiques n'inclue pas le temps comme paramètre descriptif du phénomène d'exposition, sous prétexte que le niveau de contamination des aliments est relativement stable au cours du temps. On cherche donc à déterminer la probabilité de dépassement de la dose de référence pour un comportement alimentaire habituel ("usual intake"), approché le plus souvent par la consommation moyenne observée dans des enquêtes de courte durée (une semaine). Ce type d'approche soulève déjà certaines questions importantes décrites dans la Section 2 mais est surtout problématique du fait que des excursions occasionnelles au delà de la dose de référence ne constituent pas selon les toxicologues un danger réel. De plus, les propriétés pharmacocinétiques des contaminants, e. g. leur vitesse d'élimination au sein de l'organisme, ne sont pas du tout prises en compte dans ces approches statiques, de même que la variabilité du processus de consommation.

Dans une première section, nous décrivons des méthodes statistiques développées dans la perspective d'évaluer les risques alimentaires chroniques dans un cadre statique, plus précisément les méthodologies élaborées autour d'une définition statique du risque comme la probabilité que l'exposition à un contaminant dépasse une dose tolérable par l'organisme déterminée de manière indépendante ; définition du risque très proche de celle donnée aujourd'hui par les réglementations nationale et internationale. Dans une seconde section, nous introduisons le concept de *processus d'exposition*, lequel intègre la dimension temporelle du phénomène de contamination par voie alimentaire. Les propriétés d'un modèle stochastique parcimonieux proposé dans [53] sont brièvement rappelées, les quantités susceptibles de décrire le risque dans ce cadre dynamique sont discutées et des techniques d'inférence statistique adaptées aux données disponibles en pratique, fondées sur les capacités de simulation des processeurs actuels, sont également décrites. Les résultats numériques obtenus pour l'exposition au méthyle-mercure alimentaire au sein de la population des femmes françaises en âge de procréer illustrent la méthodologie promue dans cet article. Enfin, dans une troisième section, nous indiquerons les perspectives de recherches pressenties pour les années à venir dans ce domaine d'application.

## 2. Evaluation du risque alimentaire : approches statiques

Comme nous venons de le mentionner dans la section précédente, les risques alimentaires font l'objet d'une analyse multi-facette, réalisée par des équipes pluridisciplinaires et discutée dans différents comités d'experts internationaux. Ils sont quantifiés à partir de données d'enquêtes sur la consommation alimentaire et de données analytiques sur la contamination des aliments dans la majorité des cas puisque que l'exposition n'est que très rarement observée directement. En France, la consommation alimentaire est le plus souvent mesurée à partir de l'enquête INCA de 1999 qui fournit le détail de l'ensemble des prises alimentaires de 3003 individus sur 7 jours, [15]. D'autre part, les toxicologues déterminent, en tous cas pour les contaminants que nous étudions, une dose hebdomadaire tolérable provisoire (DHTP, ou provisional tolerable weekly intake -PTWI- en anglais) : les premiers travaux ont donc consisté à formaliser le calcul du risque comme la probabilité de dépassement de cette dose hebdomadaire tolérable ; voir [49] et [9].

Une première approche empirique, entièrement non paramétrique, a été développée, *cf.* [10, 50], laquelle permet d'intégrer à la fois la variabilité observée au niveau des consommations et des contaminations des aliments en prenant en compte les caractéristiques particulières des deux types de données : la forte dépendance entre les consommations des aliments et la censure à gauche des données de contamination. Cependant, quand la dose hebdomadaire tolérable se situe assez loin dans la queue à droite de la distribution d'exposition, le dépassement de cette dose peut être vu comme un événement rare, au moins relativement à la taille des échantillons observés. Les techniques empiriques s'avèrent alors non performantes et tendent à sous-estimer le risque de ce type d'événement, correspondant néanmoins à un danger très important. La théorie des valeurs extrêmes peut dans ce cas compléter l'analyse, [52].

**2.1. Evaluation empirique du risque.** — L'idée de l'évaluation empirique du risque est simplement de combiner les distributions observées de consommation hebdomadaire (à valeurs dans l'ensemble des nombres réels positifs  $\mathbb{R}^+$ ) aux différentes valeurs de contaminations (également à valeurs dans  $\mathbb{R}^+$ ) par une simulation de type Monte-Carlo puis de dénombrer les expositions obtenues dépassant la dose hebdomadaire tolérable.

**Formalisation mathématique.** Notons  $P$  le nombre d'aliments ou groupe d'aliments contenant le contaminant étudiés, et considérons  $C = (C^{(1)}, \dots, C^{(P)})$  les consommations "usuelles", exprimées relativement au poids corporel de l'individu, de ces  $P$  aliments, sur une semaine par exemple, et  $Q^{[p]}$  le niveau de contamination de l'aliment  $p$ , pour  $p = 1, \dots, P$ . On dispose d'un  $n$ -échantillon  $(c_1, \dots, c_n)$  de consommation de loi notée  $F_C$ ,

de  $L_p$ -échantillons  $q_1^{[p]}, \dots, q_{L_p}^{[p]}$  de contamination de loi  $F_{Q^{[p]}}$ , mutuellement indépendantes. Alors un estimateur empirique de la probabilité que l'exposition au contaminant dépasse une dose  $d > 0$  donnée s'écrit

$$\hat{\theta}_d = \frac{1}{\Lambda} \sum_{i=1}^n \sum_{j_1=1}^{L_1} \dots \sum_{j_P=1}^{L_P} \mathbb{1} \left( \sum_{p=1}^P q_{j_p}^{[p]} c_i^{(p)} > d \right),$$

où  $\Lambda = n \times \prod_{p=1, \dots, P} L_p$ , la loi forte des grands nombres assurant que  $\hat{\theta}_d$  tend presque-sûrement vers  $\theta_d = \mathbb{P}(U > d)$ ,  $U = \sum_{p=1}^P Q^{[p]} C_p$  désignant l'exposition au contaminant via la consommation des  $P$  aliments étudiés. Il est noté dans [10] que  $\hat{\theta}_d$  est une *U-statistique généralisée d'ordre  $P + 1$  et de degrés  $1, \dots, 1$* . En pratique, le nombre  $\Lambda$  est souvent trop grand pour pouvoir appliquer des stratégies de calcul naïves. Ainsi, dans une simulation Monte-Carlo non paramétrique avec  $M$  réplifications, les observations de consommations et contaminations peuvent être tirées aléatoirement avec remise pour former des expositions, sans chercher à considérer toutes les combinaisons consommation-contamination possibles mais seulement un nombre  $M \ll \Lambda$  parmi celles-ci, ce qui correspond à une version "incomplète" de la *U-statistique généralisée*. Ce constat permet l'utilisation des outils relatifs aux *U-statistiques*, décrits dans [37] par exemple, pour étudier le comportement asymptotique de l'estimateur ainsi obtenu (quand  $n, L_1, \dots, L_P$  deviennent grands), proposer une décomposition de sa variance, via la décomposition de Hoeffding, ainsi que montrer la validité de la construction d'intervalles de confiance par bootstrap, [10].

**Traitement de la censure des données analytiques.** L'estimateur précédent, fonction des lois empiriques  $F_{C,n}, F_{Q^{[1]}, L_1}, \dots, F_{Q^{[P]}, L_P}$  associées aux lois  $F_C, F_{Q^{[1]}}, \dots, F_{Q^{[P]}}$ , peut être adapté au cas de données censurées en remplaçant les fonctions de répartition empiriques par les estimateurs non paramétriques de Kaplan-Meier de la fonction de répartition pour chaque variable censurée. Une censure (à gauche) des variables décrivant la contamination des aliments apparaît en effet du fait de l'existence de limites de détection pour toute analyse biologique. Des arguments de différentiabilité au sens d'Hadamard et de méthode delta fonctionnelle ont permis de dériver des résultats similaires à ceux obtenus dans le cas non censuré : le comportement asymptotique, la décomposition de la variance et la validité de la construction d'intervalles de confiance par bootstrap ont été étudiés dans [50]. La Figure 1 illustre l'intérêt de ce type d'approche en comparaison avec d'autres traitements de la censure des données analytiques sur l'exemple de l'Ochratoxine A, mycotoxine présente dans les céréales et produits à base de grains, ayant potentiellement des effets sur le système rénal et urinaire à long terme.

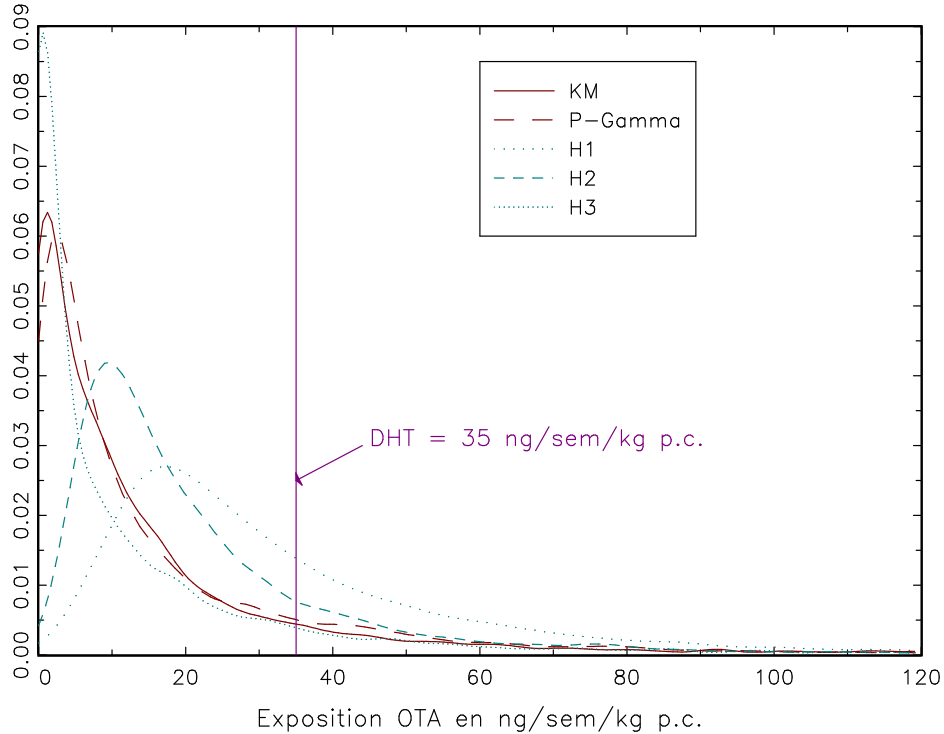


FIGURE 1. Estimation de l'exposition avec prise en compte de la censure des données. Cas de l'ochratoxine A.

Remplacement des données censurées par les valeurs de censure (H1), par les valeurs de censure divisées par 2 (H2), par zéro (H3), modèle paramétrique de loi Gamma (P-Gamma), et estimateur de Kaplan Meier (KM).

**Combinaison de sources de consommation.** Les méthodologies décrites dans les paragraphes précédents ont été appliquées en utilisant les données individuelles de consommation INCA et sont également implémentées dans le logiciel CARAT (*Chronic and acute risk assessment*), chapitre 12 de [9]. Toutefois, il est clair que les données de panel Secodip recensant les achats alimentaires hebdomadaires des ménages sur longue période, présente un avantage certain en termes de durée de suivi des individus, en particulier lorsqu'on s'intéresse à un risque chronique, [1]. Dans ce cadre statique, la combinaison des deux sources de données de consommation permet déjà d'affiner l'estimation proposée. La méthodologie retenue est la vraisemblance empirique, méthode semi-paramétrique pour laquelle on cherche une pondération optimale des observations pour l'estimation d'un paramètre donné, voir [43] par exemple. Dans

notre problème, le paramètre est la probabilité de dépasser la dose tolérable, les observations sont les deux jeux de données de consommation et les données de contamination. Les techniques dédiées à l'étude des  $U$ -statistiques présentées précédemment permettent ici de proposer une solution approchée au problème d'optimisation, même dans le cas où le nombre d'aliments pris en compte dans l'analyse est grand, *cf.* [16].

## 2.2. Evaluation des " petits risques " : théorie des valeurs extrêmes.

— Pour certains contaminants et selon la qualité des données disponibles, les méthodes empiriques ne sont pas satisfaisantes car elles peuvent conduire à une estimation de risque nulle (les expositions construites ne dépassent que peu ou pas la dose de référence). Des hypothèses sur la forme de la queue de distribution de l'exposition permettent alors de quantifier ces risques dits "faibles", mais correspondant néanmoins à de "grands" dangers. Le choix des queues de courbe épaisses, ou plus formellement d'appartenance au domaine d'attraction du maximum de Fréchet, se justifie de plusieurs manières. D'abord, l'exposition est obtenue comme le produit croisé de la consommation et de la contamination et peut donc prendre des valeurs arbitrairement grandes avec une probabilité importante. Ensuite, il revêt un caractère conservateur au sens où ce choix tend davantage à surestimer le risque que celui des queues de courbe fines : ce caractère conservateur est souvent recherché en analyse de risque, le principe de précaution restant une règle communément admise.

**Indice de Pareto et mesure de risque "extrême".** Sous hypothèse de queues de courbe épaisses, le principal paramètre à estimer est l'indice de Pareto dont l'inverse peut être interprété comme un indice de risque. Supposons que l'exposition  $U$  à un contaminant appartienne au domaine d'attraction du maximum de Fréchet, que nous noterons  $MDA(\Phi_\alpha)$ , *cf.* [25], alors sa distribution  $F_U$  satisfait la condition suivante pour  $x$  suffisamment grand,

$$(1) \quad \mathbb{P}\{U > x\} = x^{-\alpha}L(x),$$

où  $\alpha$  est un paramètre strictement positif et  $L$  est une fonction à variation lente, *i.e.* satisfaisant pour tout  $t > 0$  :

$$(2) \quad \lim_{x \rightarrow \infty} \frac{L(tx)}{L(x)} = 1.$$

Un estimateur classique de  $\gamma = 1/\alpha$  dans le cas d'une fonction à variation lente constante est l'estimateur de Hill, [33], obtenu comme maximum de vraisemblance conditionnel à la sélection d'un certain nombre  $k$  de valeurs extrêmes (ici les  $k$  expositions les plus élevées) et dont on sait qu'il est biaisé dès que  $L$  n'est pas constante. Si  $U_1, \dots, U_n$  sont les expositions à un contaminant

de  $n$  individus indépendants alors l'estimateur de Hill s'écrit

$$(3) \quad H_{k,n} = \frac{1}{k} \sum_{i=1}^k \log(U_{n-i+1,n}) - \log(U_{n-k,n}),$$

où  $k$  désigne le nombre de valeurs extrêmes à retenir, et  $U_{i,n}$  est la statistique d'ordre  $i$  associée au  $n$ -échantillon. En précisant la forme de la fonction à variation lente (voir par exemple [5, 28]), on peut corriger le biais de cet estimateur. Ces fonctions permettent dans le cas des risques alimentaires de tenir compte du mélange éventuel de populations et de déterminer le nombre  $k$  de valeurs extrêmes constituant effectivement la queue de distribution. Pour chaque valeur de  $k$ , on construit en effet un modèle de régression sur les écarts de statistiques d'ordre  $Z_i = i(\log U_{n-i+1,n} - \log U_{n-i,n})$ , modèle dont les résidus sont exponentiels. Deux types de fonctions à variation lente ont été envisagés dans ce cadre : le type puissance  $1 + Dx^{-\beta}$  avec  $D \in \mathbb{R}, \beta > 0$  ou logarithmique  $(\log x)^\theta$  avec  $\theta > 0$ ; cf. [49, 52]. La Figure 2 illustre l'effet de la correction sur un exemple, le nombre de valeurs extrêmes optimale étant alors déterminé de façon standard par un arbitrage biais-variance, voir [32]. L'estimateur corrigé est ensuite utilisé pour estimer le risque ou, par un raisonnement inverse de type " Value at Risk " (VaR), les niveaux d'expositions dépassés par une infime partie de la population. Il est aussi possible de caractériser les populations les plus à risque en écrivant l'indice de Pareto comme une fonction de variables socio-démographiques dont les paramètres sont estimés par maximum de vraisemblance, cf. [9], chapitre 9.

**Classification de valeurs extrêmes et populations "à risque".** Une autre approche fondée sur des techniques de statistique bayésienne non paramétrique a été mise en oeuvre sur un échantillon de données relatif à l'ochratoxine A, [51]. Elle consiste à supposer les observations distribuées selon un mélange de lois de Pareto et d'en estimer les différentes composantes en choisissant pour la loi a priori du mélange un processus de Dirichlet. Le caractère presque sûrement discret de ces processus permet d'identifier des clusters d'observations tirées selon la même composante du mélange et ainsi des partitions aléatoires générées par un processus dit "de restaurant chinois pondéré" ou encore "Weighed Chinese Restaurant Process". On parvient à déterminer à la fois le nombre de composantes du mélange et leurs paramètres par l'utilisation d'un algorithme de Monte-Carlo par Chaîne de Markov (MCMC) de type échantillonnage de Gibbs. Dans le cadre de l'application à la classification de valeurs extrêmes d'exposition, on ne parvient pas réellement à identifier des groupes d'individus à risque tant l'analyse des 11 clusters optimaux se révèle difficile. On obtient par contre un nouvel estimateur de l'indice de Pareto, prenant en compte les diverses composantes du mélange et une estimation non paramétrique de la

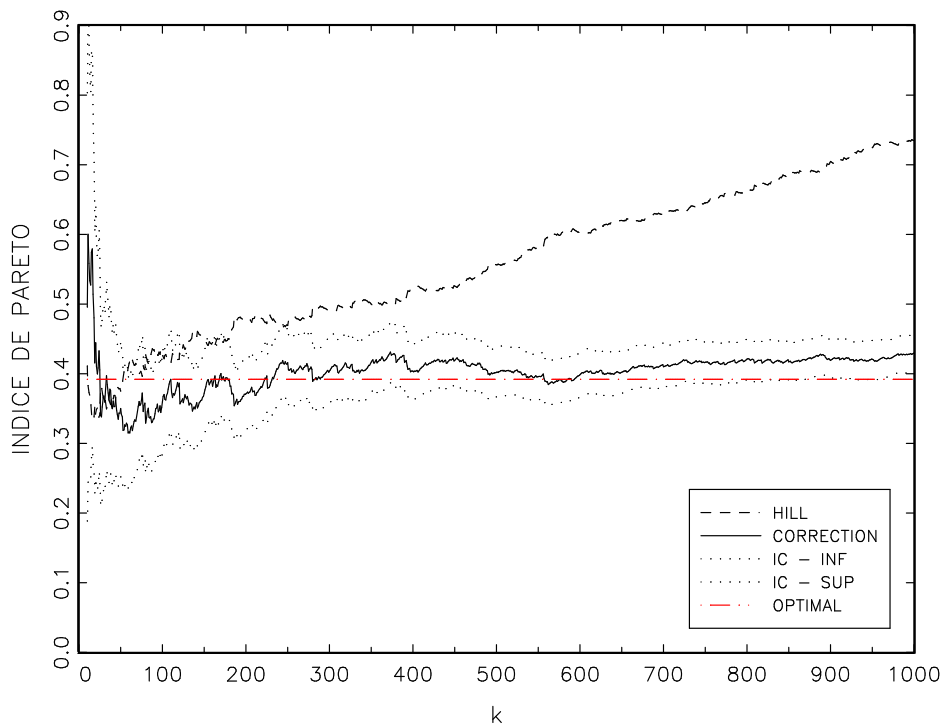


FIGURE 2. Exemple de correction de biais de l'estimateur de Hill : données méthyle mercure,  $n = 2513$ ,  $k_{\text{opt}} = 110$ ,  $\gamma_{\text{opt}} = 0.392$  (en abscisse : nombre  $k$  de valeur extrêmes retenues pour le calcul de l'estimateur ; en ordonnée : l'estimateur de Hill et la correction de biais proposée).

queue de courbe similaire à celle obtenue via un estimateur à noyau, ici de Pareto.

### 3. L'exposition vue comme un processus stochastique

Les approches statiques décrites précédemment ne sont pertinentes que dans les cas où la vitesse d'élimination du contaminant est "grande" par rapport à la durée moyenne entre prises alimentaires, c'est le cas de l'Ochratoxine A par exemple. Dans les cas de composés chimiques "à cinétique lente" par contre, tels que le méthyle-mercure ou les dioxines, le contaminant s'accumule au sein de l'organisme du fait des prises alimentaires successives. Il convient alors de décrire l'évolution de la quantité de contaminant alimentaire présent dans l'organisme au cours du temps. Ainsi, le phénomène d'exposition

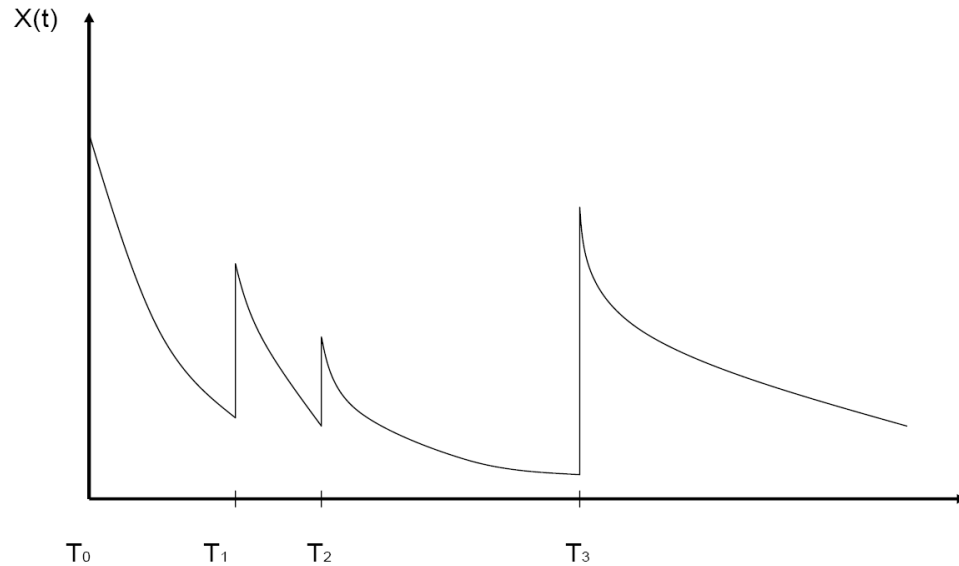


FIGURE 3. Une trajectoire du processus d'exposition : les sauts du processus, aux dates  $T_1, \dots, T_n$ , correspondent à une ingestion de contaminant, celui-ci étant ensuite lentement éliminé.

au risque alimentaire n'est alors plus décrit par une variable aléatoire, représentant la quantité de contaminant ingérée au cours d'une journée ou d'une semaine, mais par une collection de variables aléatoires, *i.e.* un processus stochastique,  $X = \{X(t)\}_{t \geq 0}$ , la variable  $X(t)$  représentant la quantité de contaminant présent dans l'organisme au temps  $t$ . La dynamique du processus  $X$  est principalement régie par deux phénomènes : le régime alimentaire induisant une accumulation du contaminant au sein de l'organisme et les propriétés pharmaco-cinétiques du contaminant, caractérisant la vitesse à laquelle il est progressivement éliminé du corps humain. Ainsi, le processus  $X$  évolue par "sauts" au moment des prises alimentaires, les sauts correspondant à la quantité de contaminant ingérée lors des prises, et selon l'équation pharmaco-cinétique décrivant l'élimination progressive du contaminant entre les prises, voir Figure 3.

**3.1. Description du modèle " KDEM ".** — Un premier modèle à temps continu, baptisé *Kinetic Dietary Exposure Model* (KDEM), fondé sur des modélisations parcimonieuses des phénomènes d'accumulation et d'élimination a récemment été proposé par [53] et étudié par [8] (voir aussi [7] pour une application plus spécifique au cas de la contamination au méthyle-mercure par voie alimentaire). Il peut être vu comme une généralisation des *modèles de stockage*

utilisés en Recherche Opérationnelle, [2]. L'accumulation par voie alimentaire y est décrite par un processus ponctuel marqué  $(T_n, U_n)_{n \in \mathbb{N}}$ ,  $T_n$  désignant la date de la  $n$ -ième prise et  $U_n$  la quantité de contaminant ingérée correspondante. Les durées "inter-prises",  $\Delta T_n = T_n - T_{n-1}$  avec  $n \geq 1$ , sont supposées indépendantes et identiquement distribuées, de loi  $G(dt) = g(t)dt$ . La suite  $\{T_n\}$  forme ainsi un processus de renouvellement standard. Par convention, l'origine des temps correspond ici à la première prise considérée  $T_0 = 0$ . Les quantités de contaminant ingérées  $\{U_n\}_{n \in \mathbb{N}}$  sont supposées former une suite de variables aléatoires i.i.d., de loi  $F_U(dx) = f_U(x)dx$ , indépendante de la série chronologique  $\{T_n\}$ . Les sauts du processus d'exposition sont ainsi les mêmes que ceux du processus à trajectoires càd-làg défini par :

$$\forall t \geq 0, W(t) = \sum_{n \leq N(t)} U_n,$$

où  $N(t) = \sum_{n \in \mathbb{N}} \mathbb{I}_{\{T_n \leq t\}}$  désigne le nombre de prises alimentaires avant  $t$ .

Un modèle pharmaco-cinétique linéaire à un compartiment est utilisé pour décrire le processus d'élimination. Précisément, entre deux prises consécutives, la quantité de contaminant présente dans l'organisme décroît avec le temps selon l'équation différentielle linéaire :

$$\frac{dx}{dt}(t) = -\theta \cdot x(t),$$

avec  $\theta > 0$ . Même si ce modèle est d'une grande simplicité puisqu'il ne dépend en effet que du paramètre d'accélération  $\theta$ , il est largement utilisé en pratique et rend compte des propriétés pharmaco-cinétiques d'un grand nombre de composés chimiques, [29], le méthyle-mercure et les dioxines en particulier. Il est généralement décrit par la *demi-vie biologique* du contaminant, *i.e.* le paramètre  $\log(2)/\theta$ , représentant le temps nécessaire pour que la quantité de contaminant au sein de l'organisme soit réduite de moitié en l'absence de prise supplémentaire.

La dynamique du processus à temps continu "KDEM" est ainsi régie par trois paramètres : les distributions  $G$  et  $F_U$  caractérisant le comportement alimentaire au regard du phénomène de contamination et la demi-vie biologique du contaminant. Ce processus est dit "déterministe par morceaux" dans la mesure où il évolue de façon déterministe entre les prises successives, selon la terminologie introduite par [20].

**3.2. Analyse probabiliste du modèle KDEM.** — Soit  $A(t)$  la durée, à l'instant  $t$ , écoulée depuis la dernière prise :  $\forall t \geq 0, A(t) = t - T_{N(t)}$ . On désignera par  $\zeta(t) = g(t)/G([t, \infty[)$  le taux de hasard relatif aux durées  $\Delta T_n$ . Par construction, le processus bivarié  $Z = \{(X(t), A(t))\}_{t \geq 0}$  est markovien, homogène dans le temps. Sa loi est donc décrite par son générateur infinitésimal ;

il est donné par :

$$\begin{aligned} \mathcal{G}\phi(x, t) &= \lim_{h \rightarrow 0} \mathbb{E}[\phi(X_{s+h}, A_{s+h}) \mid (X_s, A_s) = (x, t)] \\ &= \zeta(t) \int_{u=0}^{\infty} \{\phi(x+u, 0) - \phi(x, t)\} F_U(du) - \theta x \frac{\partial \phi}{\partial x}(x, t) + \frac{\partial \phi}{\partial t}(x, t), \end{aligned}$$

pour toute fonction  $\phi : \mathbb{R}_+^2 \rightarrow \mathbb{R}$  bornée, de dérivées premières bornées. Sous des hypothèses très générales, il a été montré dans [8] que le processus  $Z$  était *stochastiquement stable*. Plus précisément, considérons les conditions suivantes.

$H_1$  : La distribution  $G(dt)$  a une queue de distribution à droite infinie et, soit il existe  $\epsilon > 0$  tel que  $\inf_{x \in ]0, \epsilon[} g(x) > 0$ , soit  $F_U(dx)$  a une queue de distribution à droite infinie.

$H_2$  : Il existe  $\gamma \geq 1$  tel que  $\mathbb{E}[U_1^\gamma] < \infty$ .

$H_3$  : Il existe  $\delta > 0$  tel que  $\mathbb{E}[\exp(\delta T_1)] < \infty$ .

L'hypothèse  $H_1$  stipule d'une part que les durées inter-prises peuvent être arbitrairement longues avec une probabilité non nulle (quoique relativement "petite" si la condition de moment exponentiel  $H_3$  est vérifiée), et d'autre part que, soit elles peuvent aussi être arbitrairement courtes avec une probabilité non nulle, soit la valeur des prises peut être arbitrairement grandes avec une probabilité non nulle (mais relativement "petite" également si la condition de moment  $H_2$  est satisfaite). Ces hypothèses sont peu restrictives du point de vue de la modélisation et aisément vérifiables en pratique, en particulier dès lors que des familles de lois paramétriques ont été choisies pour les distributions  $G$  et  $F_U$ . Sous la condition  $H_1$ , le processus  $Z$  est *irréductible* au sens où quelque soit la valeur initiale du couple  $(X(0), A(0))$ , le processus atteint tout pavé  $[x, \infty[ \times ]t, \infty[$  au bout d'un temps fini avec une probabilité strictement positive, cf. Théorème 2.1 de [8].

Si de plus les hypothèses  $H_2 - H_3$  sont vérifiées, alors on peut établir la *condition de dérive* de type Foster-Lyapounov suivante

$$(4) \quad \mathcal{G}V(x, t) \leq -\eta V(x, t) + b \mathbb{1}_{\{(x, t) \in [0, s] \times [0, a]\}},$$

associée au *petit ensemble*  $[0, s] \times [0, a]$ , pour  $s$  et  $a$  convenablement choisis, et à la fonction test

$$(5) \quad V(x, t) = (1 + x^\gamma) \left\{ 1 + G(t) e^{-\eta t} \left( 1 + \int_{s=t}^{\infty} e^{\delta s} \frac{1 - G(s)}{1 - G(t)} ds \right) \right\},$$

avec  $0 < \eta < \delta$  et  $b < \infty$ .

On se réfèrera à [40] pour une exploitation systématique de ce type de conditions en vue d'étudier les propriétés ergodiques des processus markoviens à temps continu. En particulier, la condition (5) permet d'établir l'ergodicité géométrique du processus  $Z$ . En conséquence, sous ces conditions, il existe une loi de probabilité limite  $\mu$ , dite aussi *mesure d'équilibre*, pour le processus d'exposition  $X(t)$  : la loi de  $X(t)$  converge avec une vitesse exponentielle vers

$\mu$  lorsque  $t \rightarrow \infty$  au sens où, quel que soit l'état initial  $(x_0, a_0)$  du processus  $Z : \forall t \geq 0$ ,

$$(6) \quad \sup_{\phi} \left| \mathbb{E} [\phi(X(t)) \mid (X(0), A(0)) = (x_0, a_0)] - \int_{u=0}^{\infty} \phi(u) \mu(du) \right| \leq C_{a_0} (1 + x_0^\gamma) \beta^t,$$

pour des constantes  $C_{a_0} < \infty$  et  $\beta \in ]0, 1[$ , le supremum étant pris sur l'ensemble des fonctions  $\phi : \mathbb{R}_+ \rightarrow \mathbb{R}$  mesurables et majorées par l'application  $y \in \mathbb{R}_+ \mapsto 1 + y^\gamma$ , cf. Théorème 3.2 de [8].

La mesure  $\mu$  peut ainsi être interprétée comme une représentation de l'état d'équilibre dans lequel le processus d'exposition s'installe au fur et à mesure que le temps passe. En vertu de la loi forte des grands nombres, elle permet de décrire le comportement asymptotique des moyennes temporelles, calculées sur des intervalles de temps très longs : lorsque  $T \rightarrow \infty$ , on a ainsi

$$\frac{1}{T} \int_{t=0}^T \phi(X(t)) dt \rightarrow \int_{x=0}^{\infty} \phi(x) \mu(dx) \text{ presque-sûrement,}$$

pour toute fonction  $\phi$  intégrable par rapport à la loi  $\mu$ .

Bien qu'il soit théoriquement établi que le comportement asymptotique du processus d'exposition est décrit par une loi de probabilité  $\mu$ , il convient de vérifier, d'un point de vue pratique, que cette convergence peut effectivement être observée à l'échelle humaine, autrement dit, qu'au bout de quelques décennies tout au plus, la distribution de  $X(t)$  coïncide avec la loi à l'équilibre  $\mu$ . Des techniques de *couplage* peuvent être utilisées pour contrôler la distance en variation totale entre ces deux distributions, voir les nombreux travaux de probabilités appliquées, cf. [46] ou [47] par exemple. Les bornes de contrôle obtenues sont en général assez "lâches" et l'application de ce type de résultats ne s'est pas révélée satisfaisante dans le cas de l'exposition au méthyle-mercure. Des simulations peuvent toutefois être menées afin d'évaluer empiriquement la vitesse de convergence pour une quantité donnée, voir Figure 4 pour une approximation du temps d'atteinte de l'état stationnaire dans le cas de l'estimation de la moyenne du processus.

**3.3. Etat d'équilibre vs. comportement extrême.** — Décrire le phénomène d'exposition au risque alimentaire par un processus et non plus par une variable aléatoire conduit à s'interroger sur les quantités susceptibles de représenter le risque ou de résumer le comportement du processus. La richesse du modèle proposé permet en effet de considérer des quantités dépendant du temps, comme par exemple :

- l'exposition maximale sur une fenêtre de temps donnée

$$(7) \quad \max_{t \in [0, T]} X(t),$$

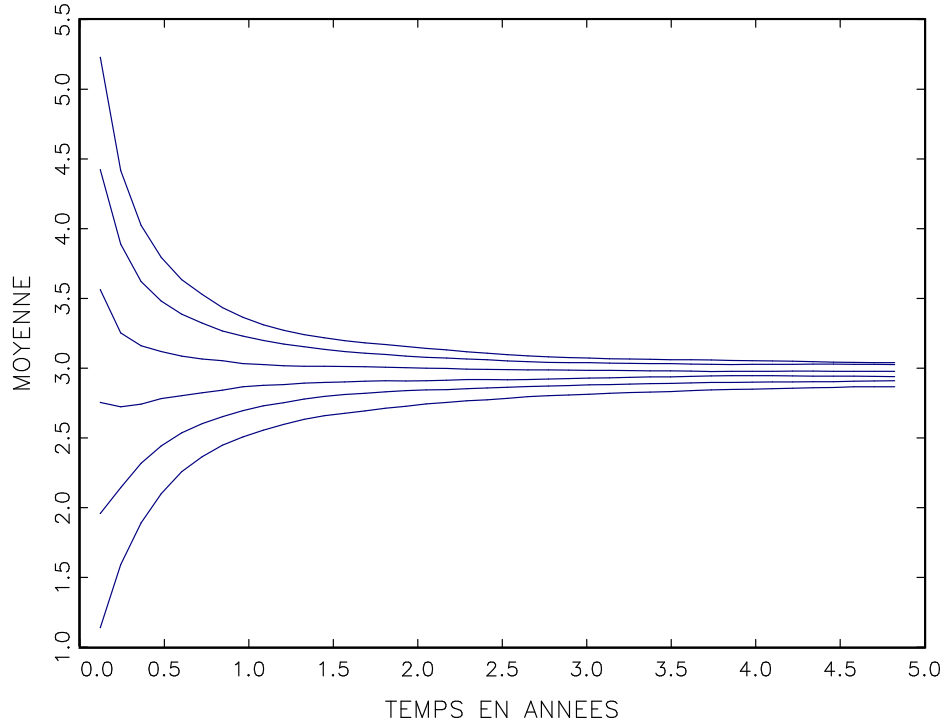


FIGURE 4. Temps d'atteinte de l'état stationnaire pour la moyenne sous la loi stationnaire  $\mu$ ,  $T$  exprimé en années - résultat moyen sur 1000 trajectoires partant respectivement de  $x_0 = 0, 1, \dots, 5$ , pour un modèle Burr-Gamma, les distributions  $F_U$  et  $G$  étant estimées sur les données méthyle mercure pour la population 'Femmes 15-45 ans', cf. application numérique de la section 3.4.

- le temps passé au dessus d'un seuil de référence  $u$  sur une période de temps donnée

$$(8) \quad \int_{t=0}^T \mathbb{I}_{\{X(t) \geq u\}} dt,$$

- le temps nécessaire pour dépasser un niveau d'exposition critique  $u$

$$(9) \quad \tau_u = \inf\{t \geq 0 \mid X(t) \geq u\}.$$

Dans ce contexte, il pourra apparaître pertinent de représenter les situations dangereuses comme des événements relatifs à ces variables aléatoires et le risque comme la probabilité d'occurrence de ces événements. Ceci nous renvoie à l'étape de *caractérisation du danger* décrite dans la première section

et requiert probablement de mettre au point de nouveaux protocoles expérimentaux, cf. section 4. Une façon directe d'étendre les indicateurs de risque "statiques" décrits dans la section précédente consiste par ailleurs à considérer l'exposition à l'état d'équilibre, la mesure  $\mu$  représentant le comportement du processus d'exposition sur le long-terme, et calculer les quantités suivantes, certaines pouvant être interprétées comme limites "presque-sûres" de moyennes temporelles calculées sur des intervalles de temps asymptotiquement grands :

- l'exposition moyenne à l'équilibre

$$\mathbb{E}_\mu[X] = \int_{x=0}^{\infty} x\mu(dx) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{t=0}^T X(t)dt,$$

- la probabilité de dépasser un seuil critique  $u$  à l'équilibre

$$\mathbb{P}_\mu\{X > u\} = \mu([u, \infty]) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{t=0}^T \mathbb{I}_{\{X(t) \geq u\}} dt,$$

- le dépassement espéré à l'équilibre d'un seuil  $u$  donné ("expected overshoot")

$$\mathbb{E}_\mu[X - u \mid X > u] = \mu([u, \infty])^{-1} \int_{x=u}^{\infty} (x - u)\mu(dx).$$

Dans cette perspective d'exposition sur le long terme, la notion de seuil de référence doit être redéfinie. Dans [53], une dose de long-terme associée à la DHTP a été proposée, nous l'appellerons ici *charge corporelle tolérable*. Elle est construite comme la limite du processus d'exposition déterministe correspondant à une ingestion hebdomadaire d'une quantité de contaminant égale à la DHTP :  $X_{ref} = \lim_{n \rightarrow \infty} x_n$ , avec

$$(10) \quad x_{n+1} = x_n \exp(-7 \times \log(2)/DV) + DHTP,$$

$DV$  désignant la demi-vie biologique du contaminant exprimée en jours. Comme l'illustre la Figure 5, la dose de référence ainsi obtenue est une valeur extrême pour le processus d'exposition. En effet, après atteinte de l'état stationnaire (partie droite du graphique), toutes les trajectoires sont bien inférieures à la charge corporelle tolérable préalablement définie.

Un autre point de vue, prolongeant l'approche "valeurs extrêmes" développée dans le cadre statique, consiste à caractériser la loi asymptotique de l'exposition maximum :  $M_T = \sup_{t \in [0, T]} X(t)$  lorsque  $T \rightarrow \infty$ . Afin d'étudier le comportement extrême du processus d'exposition, il convient de remarquer que par construction

$$(11) \quad M_T = \max_{1 \leq n \leq N(T)} X_n,$$

où  $X_n$  désigne l'exposition, calculée immédiatement après la  $n$ -ième prise, i.e.  $X_n = X(T_n)$ . Le processus à temps discret  $\{X_n\}_{n \in \mathbb{N}}$  forme une chaîne de

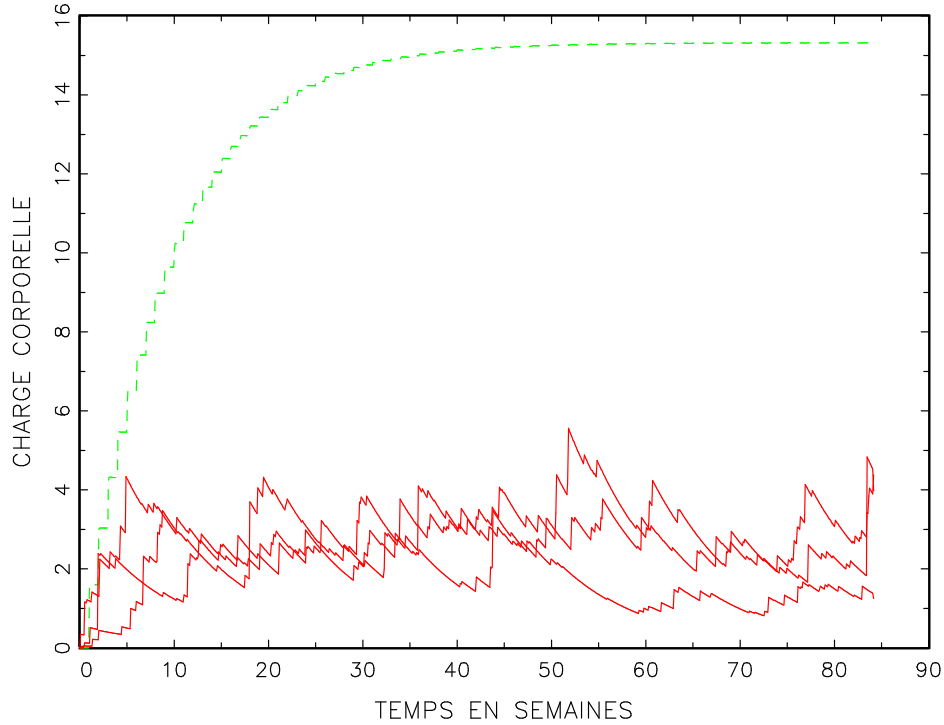


FIGURE 5. Comparaison de trajectoires type pour des femmes âgées de 15 à 45 ans (en rouge, trait plein) avec la charge corporelle tolérable définie comme la limite de (10) (en vert, ligne brisée).

Markov, plus précisément un modèle autorégressif à coefficients aléatoires :

$$(12) \quad X_{n+1} = X_n \cdot e^{-\theta \Delta T_{n+1}} + U_{n+1}.$$

Ce type de série chronologique a été largement étudié dans la littérature, voir [44] et [11] par exemple (voir aussi [3]). Sous les conditions  $H_1 - H_3$ , la chaîne de Markov  $\tilde{X} = \{X_n\}_{n \in \mathbb{N}}$  est récurrente positive, de loi stationnaire  $\tilde{\mu}$ . Il est en particulier établi que si  $F_U$  appartient au domaine d'attraction maximal de Fréchet  $MDA(\Phi_\alpha)$  avec  $\alpha > 0$  (*i.e.* si la fonction de répartition des prises  $F_U(x)$  s'écrit  $L(x) \cdot x^{-\alpha}$  où  $L(x)$  désigne une fonction à variation lente), alors  $\tilde{\mu}$  appartient aussi au domaine d'attraction maximal de la loi de Fréchet de paramètre de forme  $1/\alpha$ , laquelle, associée à l'indice extrême  $\xi$  de la chaîne  $\tilde{X}$ , définit également la loi asymptotique de l'exposition maximum. En effet, pour les séries d'observations dépendentes, les valeurs extrêmes n'arrivent pas de façon isolées mais par "paquets groupés". L'indice extrême peut être vu comme une mesure de ce phénomène (il vaut 1 dans le cas de suites i.i.d.).

On rappelle que toute chaîne de Markov récurrente positive possède un indice extrémal. Ainsi, lorsque  $n \rightarrow \infty$ , on a

$$\mathbb{P}\{\max_{k \leq n} X_k \leq u_n\} \approx \mu([0, u_n])^{n\xi},$$

pour une suite adéquate de seuils  $u_n$ , voir [25]. Si l'indice de Pareto  $\alpha$  peut être estimé au moyen des méthodes standard décrites dans la précédente section, le problème de l'estimation de l'indice extrémal est plus délicat. On pourra se référer à [6] pour un inventaire des techniques possibles. Les lois  $\mu$  et  $\tilde{\mu}$  sont liées par la relation :  $\forall u \geq 0$ ,

$$\mu([u, \infty]) = m_G^{-1} \int_{x=u}^{\infty} \int_{t=0}^{\infty} t \wedge \frac{\log(x/u)}{\theta} \tilde{\mu}(dx) G(dt),$$

où  $m_G = \int_{t \geq 0} t G(dt)$  désigne la durée moyenne entre les prises. Il en résulte que si  $\tilde{\mu}$  appartient au domaine  $MDA(\Phi_\alpha)$ , la loi stationnaire  $\mu$  du processus en temps continu appartient elle aussi à ce domaine d'attraction.

**3.4. Inférence statistique par simulation.** — Le modèle KDEM possède de nombreux avantages pour le praticien : il reproduit de manière réaliste le phénomène biologique sous-jacent et sa dynamique ne dépend que d'un petit nombre de paramètres clef. Toutefois, il convient de remarquer que les trajectoires des processus d'exposition ne sont en général pas observables en pratique et l'estimation de quantités telles que celles décrites dans le paragraphe précédent ne peut être conduite que par simulation. En effet, une idée naturelle pour inférer la valeur de fonctionnelles aussi complexes des paramètres  $G$ ,  $F_U$  et  $\theta$  consiste à calculer des contre-parties pseudo-empiriques à partir de trajectoires simulées selon la méthode Monte-Carlo. Les données disponibles en pratique permettent en effet de calculer des estimations  $\hat{G}$ ,  $\hat{F}_U$  et  $\hat{\theta}$  des paramètres. Afin de pouvoir simuler les trajectoires d'un processus possédant des propriétés similaires aux propriétés supposées du "vrai" processus d'exposition, il convient d'utiliser des méthodes d'inférence produisant des estimateurs  $\hat{G}$  et  $\hat{F}_U$  qui soient des distributions instrumentales absolument continues (permettant la simulation par inversion de la fonction de répartition) vérifiant les propriétés  $H_1 - H_3$ , ce qui exclut l'utilisation des distributions empiriques brutes (voir l'application numérique ci-dessous).

**Validité asymptotique.** Une étude de *stabilité*, la version stochastique des études de sensibilité, du modèle KDEM a été menée dans [8] et fournit un cadre de validité théorique satisfaisant pour les techniques d'estimation par simulation. En effet, si les estimateurs  $\hat{G}$ ,  $\hat{F}_U$  et  $\hat{\theta}$  à partir desquels les trajectoires  $\{\hat{X}(t)\}_{t \in [0, T]}$  sont simulées sont consistants, alors les estimateurs des distributions des quantités telles que (7) ou (8) basés sur ces "pseudo-trajectoires" le sont également, voir Théorème 4.1 dans [8]. Lorsqu'il s'agit d'estimer des

quantités relatives à l'état d'équilibre, la durée  $T$  des simulations doit être asymptotiquement longue, mais toutefois avec une vitesse inversement proportionnelle au carré du risque des estimateurs, afin de garantir la consistance des estimations, voir Théorème 2 dans [7].

**Intervalle de confiance par "bootstrap simulé".** Les distributions asymptotiques des statistiques décrites précédemment ne pouvant être explicitées, une procédure de type "bootstrap-percentile" pourra être utilisée pour produire des intervalles de confiance, [24]. Le principe consiste à ré-échantillonner les jeux de données  $\mathcal{D}_G$ ,  $\mathcal{D}_{F_U}$  et  $\mathcal{D}_\theta$  ayant permis de calculer les estimateurs  $\hat{G}$ ,  $\hat{F}_U$  et  $\hat{\theta}$  par tirages indépendants avec remise, de façon à produire successivement les *échantillons bootstrap*  $\mathcal{D}_G^*$ ,  $\mathcal{D}_{F_U}^*$  et  $\mathcal{D}_\theta^*$ , les versions bootstrap des estimateurs  $\hat{G}^*$ ,  $\hat{F}_U^*$  et  $\hat{\theta}^*$  et enfin les trajectoires bootstrap  $\{\hat{X}^*(t)\}_{t \in [0, T]}$  à partir desquelles seront calculées les versions bootstrappées des quantités estimées. En pratique, des approximations Monte-Carlo sont obtenues en itérant un grand nombre de fois la procédure décrite ci-dessus. Cette méthode, appelée de façon redondante "bootstrap simulé" de manière à insister sur le fait qu'une étape de simulation intermédiaire suit la phase de ré-échantillonnage est décrite précisément dans [7], en sous-section 3.2.

**Evénements rares.** Ainsi que l'illustre la Figure 5, l'événement  $\{\tau_{X_{ref}} \leq T\}$ , correspondant au dépassement d'un niveau d'exposition de référence  $X_{ref}$  supposé critique en un temps  $T$ , raisonnable à l'échelle humaine, peut être qualifié de "rare". Les méthodes de Monte-Carlo naïves décrites précédemment ne produisent alors pas de résultats satisfaisants puisque la plupart des trajectoires simulées n'atteignent pas le niveau en question. Une approche possible pour quantifier ce type de risque est de mettre en oeuvre différentes techniques de simulation d'événements rares, cf. la thèse d'Agnès Lagnoux-Renaudie, [36], pour une introduction en français à ce type de méthodologies et [30] pour une revue de méthodologies liées plus précisément à l'estimation de la probabilité d'événements du type  $\{\tau_{X_{ref}} \leq T\}$ . La première approche décrite en détail dans [7] est fondée sur l'échantillonnage d'importance (IS pour *Importance Sampling*) et consiste à simuler les trajectoires d'exposition selon une loi auxiliaire  $\tilde{\mathbb{P}}$  équivalente à la loi d'origine mais sous laquelle l'événement d'intérêt est beaucoup plus fréquent. Le changement de probabilité revient en pratique à augmenter la fréquence et le niveau des prises alimentaires de telle sorte que le niveau critique  $X_{ref}$  soit souvent atteint en un temps  $T$ . La probabilité d'occurrence de l'événement  $\{\tau_{X_{ref}} \leq T\}$  est alors estimée par  $\tilde{\mathbb{E}}[L_T \cdot \mathbb{I}_{\{\tau_{X_{ref}} \leq T\}}]$ , où  $L_T$  désigne le rapport des vraisemblances,  $d\mathbb{P}/d\tilde{\mathbb{P}}$ . Le choix de la loi auxiliaire  $\tilde{\mathbb{P}}$  est difficile en pratique et soulève d'importants problèmes de robustesse, [30]. La méthode multi-niveaux ou "Multilevel Splitting" semble être une bonne alternative ne nécessitant aucune transformation de la loi d'origine du processus.

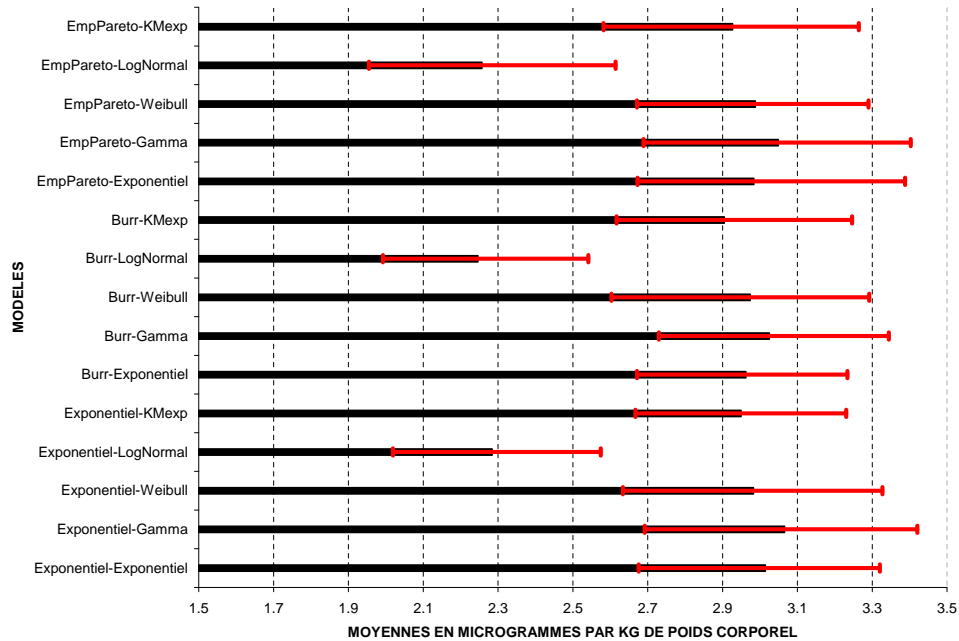


FIGURE 6. Moyenne du processus d'exposition à l'état stationnaire - 15 modèles combinant les modèles Exponentiel, Gamma, Weibull, Log Normal et semi-paramétrique (*KMexp* : estimateur de Kaplan-Meier lissé et queue exponentielle) pour  $G$ , et les modèles Exponentiel, Burr et semi-paramétrique (EmpPareto : fonction de répartition empirique lissée et queue Pareto) pour  $F_U$ .

Cette seconde approche, dite "non-intrusive", repose sur la représentation de la probabilité étudiée sous la forme d'une mesure de Feynman-Kac, cf. [41]. En pratique, elle consiste à multiplier les trajectoires ayant atteint certains niveaux intermédiaires  $u_1 < \dots < u_m$ , tels que  $u_m < u_{m+1} = X_{ref}$ , et supprimer celles qui ne les atteignent pas, cf. [12]. Une version adaptative de cet algorithme permet également de déterminer ces niveaux intermédiaires de façon optimale, [13]. On se reportera à [7] pour une description de cette procédure appliquée au problème de l'estimation de  $\mathbb{P}\{\tau_{X_{ref}} \leq T\}$  dans le cas du processus KDEM.

**Application numérique : le cas du méthyle-mercure alimentaire.** Le méthyle-mercure (MeHg) est un contaminant environnemental qui peut causer des lésions du système nerveux central du fœtus, s'il est ingéré à des doses importantes par la mère durant la grossesse, [54, 19, 31, 42]. Nous travaillons donc sur la population des femmes en âge de procréer (15-45 ans pour cette étude). En 2003, une dose hebdomadaire tolérable provisoire de

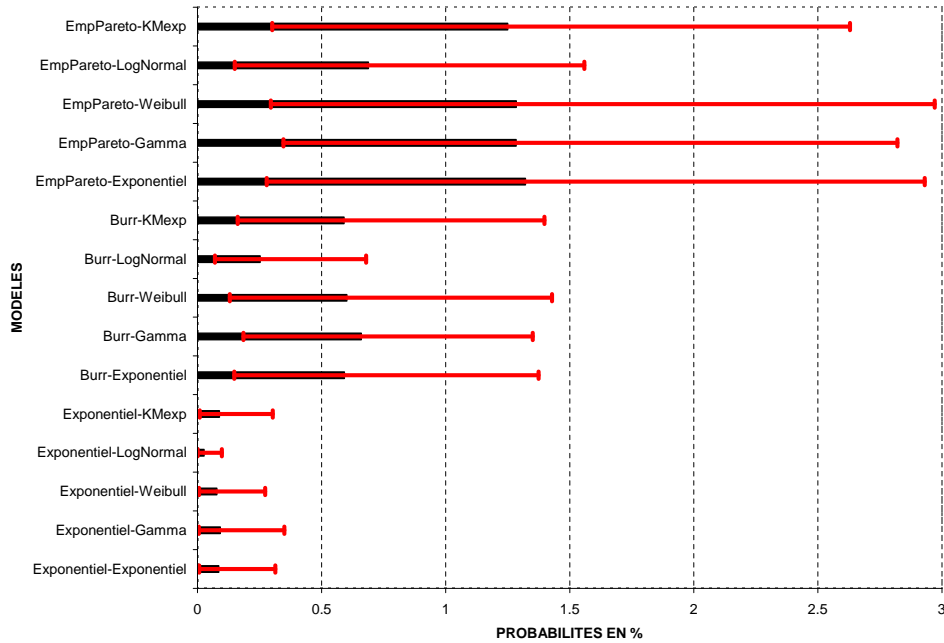


FIGURE 7. Probabilité de dépasser  $X_{ref} = 6.4 \mu\text{g}/\text{sem}/\text{kgpc}$  à l'état stationnaire (même modèles que dans la figure 6).

$1.6 \mu\text{g}/\text{semaine}/\text{kg}$  de poids corporel, a été définie par le JECFA, [27] : on en déduit une charge corporelle tolérable ou valeur de référence  $X_{ref} = 14.7 \mu\text{g}/\text{kgpc}$  pour le processus d'exposition en retenant une demi-vie biologique ( $DV$ ) de 44 jours, [48]. Le méthyle-mercure est principalement présent dans les produits de la mer et le poisson. La contamination en MeHg des différents produits entrant dans la consommation des français a été mesurée dans des plans de surveillance, [38, 34], et la consommation de ces produits (fréquence et quantité) est évaluée par l'enquête INCA, [15], l'appariement entre les deux nomenclatures "aliments" étant effectué de la même manière que dans [52] et [17]. La courte durée de l'enquête alimentaire (7 jours) crée un phénomène de censure à droite des durées inter-prises (loi  $G$ ) pris en compte dans l'estimation par maximum de vraisemblance des paramètres des modèles paramétriques et semi-paramétriques (5 pour  $G$ , 3 pour  $F_U$ ) proposés dans [7], section 4.2. On observe que la moyenne stationnaire du processus est peu affectée par le choix de modèle, cf. Figure 6, alors que l'utilisation d'une loi du domaine d'attraction du maximum de Fréchet pour  $F_U$  (Burr ou "EmpPareto") modifie de manière considérable l'estimation de la probabilité de dépasser un seuil  $X_{ref} = 6.4 \mu\text{g}/\text{kgpc}$  à l'état stationnaire, seuil issu d'une DHTP de 0.7

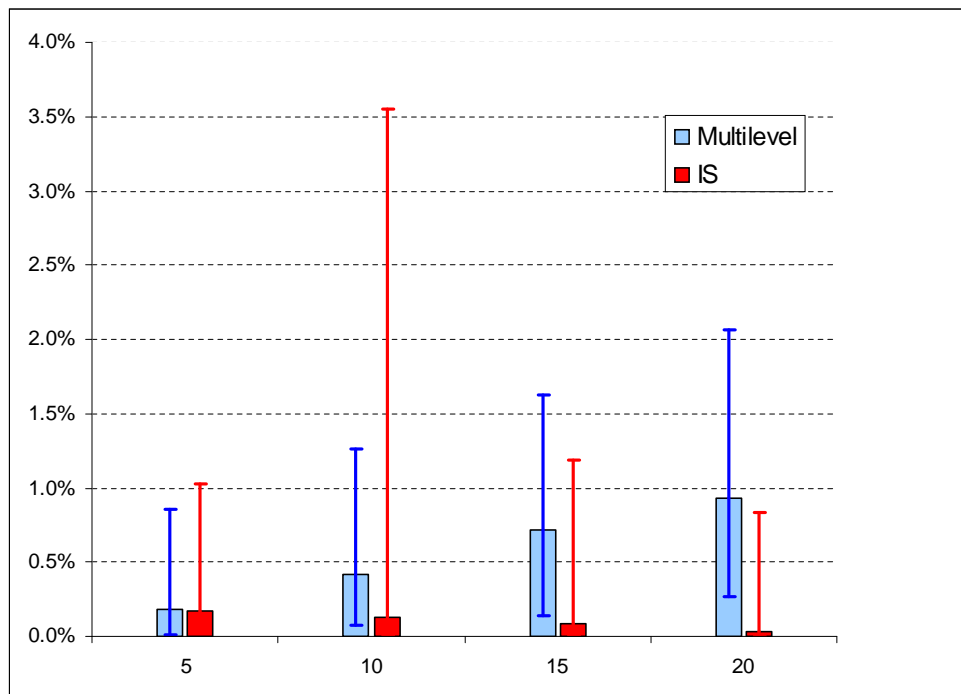


FIGURE 8. Estimation de la probabilité d'atteindre  $X_{ref} = 14.7$   $\mu\text{g}/\text{kgpc}$  en moins de  $T = 5, 10, 15$  ou  $20$  ans (IS : Importance Sampling ou échantillonnage d'importance; Multilevel : méthode multi-niveaux, intervalle de simulation de niveau 95% sur 100 répétitions).

$\mu\text{g}/\text{sem}/\text{kgpc}$  souvent utilisée aux Etats-Unis et dont le dépassement est bien moins rare que celui de  $X_{ref} = 14.7$   $\mu\text{g}/\text{kgpc}$ , associé à une DHTP de  $1.6$   $\mu\text{g}/\text{sem}/\text{kgpc}$ , cf. Figure 7. Suite à différents tests d'adéquation et pour des raisons pratiques et "conservatrices", le modèle Burr-Gamma a été retenu, [7]. On préfère en effet manipuler des lois paramétriques parcimonieuses en pratique de façon à pouvoir évaluer l'impact de la variation de quelques paramètres sur les résultats obtenus, choisir un taux de hasard croissant plutôt que constant pour les durées "inter-prise" afin d'être plus réaliste (choix de la loi Gamma pour  $G$ ) et une queue de courbe épaisse pour la distribution des prises alimentaires, de manière à ne pas sous-estimer la probabilité d'occurrence de "grandes prises", en vertu du principe de précaution (loi de Burr pour  $F_U$ ). La Figure 8 présente alors les résultats obtenus pour l'estimation de la probabilité d'atteindre  $X_{ref} = 14.7$   $\mu\text{g}/\text{kgpc}$  en moins de  $T$ , pour  $T = 5, 10, 15$  ou  $20$  ans via les techniques de simulation d'événements rares présentées ci-dessus. Les

résultats de la méthode "Multi-niveaux" sont plus convaincants que ceux de l'échantillonnage d'importance, détails disponibles dans [7].

#### 4. Discussion et perspectives de recherche

Bien qu'il marque un progrès important dans l'analyse du risque alimentaire, domaine n'ayant commencé que très récemment à bénéficier des acquis de la modélisation probabiliste, le modèle KDEM doit être modifié/généralisé de façon à pouvoir rendre compte d'un certain nombre de phénomènes importants en pratique. Dans cette première modélisation dynamique de l'exposition à un contaminant chimique via l'alimentation, nous faisons en effet plusieurs hypothèses simplificatrices qui méritent d'être discutées. Nous supposons en particulier que les durées entre les prises alimentaires successives sont indépendantes et identiquement distribuées ainsi que les prises alimentaires, les prises et les durées inter-prises étant aussi supposées indépendantes. Ce cadre est raisonnable dans le cas du méthyle mercure, présent exclusivement dans le poisson du fait que les quantités consommées sont souvent régies par une taille usuelle de portion plutôt que par la proximité temporelle de la dernière prise. Si ces hypothèses sont acceptables en première approximation, elles ouvrent de nouvelles perspectives de recherche dans au moins deux directions : la prise en compte de l'hétérogénéité des populations et la dépendance entre les prises de différents aliments susceptibles de contenir le contaminant étudié.

Il importe en effet de tenir compte de l'hétérogénéité des populations considérées au regard du comportement alimentaire de façon adéquate. Ainsi, étant donnée la non-linéarité des fonctionnelles d'intérêt mesurant le risque en les distributions d'entrée  $(G, F_U)$ , l'exposition d'une population hétérogène doit être mesurée en considérant des mélanges de processus KDEM correspondant à différentes strates de populations, décrites pas des couples de distributions  $(G, F_U)$  différents, et non par un seul processus dont les distributions d'entrée  $(G, F_U)$  seraient des mélanges de distributions représentant les strates de la population. Ceci requiert alors de mettre en oeuvre des procédures d'apprentissage statistique non-supervisé permettant d'identifier des classes de comportement alimentaires homogènes.

De plus, dans les cas où le contaminant étudié est présent dans une grande variété de produits, comme c'est le cas pour les dioxines (à la différence du MeHg présent essentiellement dans les produits de la mer ou la Pathuline que l'on ne retrouve que dans la pomme), le comportement alimentaire ne peut plus être modélisé par un processus ponctuel marqué aussi simple que celui stipulé dans l'approche KDEM originale entièrement caractérisé par le couple  $(G, F_U)$ . En effet, certains aliments s'excluent les uns des autres (c'est souvent le cas du poisson, de la viande rouge ou de la volaille, rarement présents simultanément au sein d'un même repas) ou au contraire se combinent, pour

des raisons d'équilibre nutritionnel ou de goût tout simplement. Ceci nous invite à rechercher les structures autorégressives adéquates reliant les dates de consommations des différents produits contenant la substance chimique considérée et les niveaux de consommation correspondants.

Une autre direction de recherche, impliquant une collaboration étroite avec des toxicologues, concerne la définition de la charge corporelle tolérable. Nous proposons ici une définition à partir de la DHTP, dose hebdomadaire tolérable elle-même dérivée du modèle pharmaco-cinétique liant la quantité de contaminant présente dans l'organisme et le processus d'élimination. Il serait sans doute plus robuste de déterminer directement la charge corporelle tolérable (in vivo, sur l'animal) plutôt que d'appliquer (deux fois) le modèle pharmaco-cinétique qui n'est, comme tout modèle, qu'une approximation de la réalité.

Enfin, dans la limite des données épidémiologiques disponibles, il convient également de mettre en relation la description de l'exposition avec les éventuels effets observés. Le fait que l'exposition soit décrite par un processus et non par une variable aléatoire nous incite à développer des modèles épidémiologiques, de type Cox par exemple, de nature fonctionnelle : la variable explicative pouvant être la densité de la distribution stationnaire  $\mu$  de l'exposition ou encore la courbe d'exposition moyenne  $t \mapsto \mathbb{E}[X(t)]$  par exemple.

### Remerciements

Les auteurs remercient les deux rapporteurs dont les commentaires ont permis d'améliorer considérablement la présentation de l'article. La recherche du second auteur est en partie financée par la Hong Kong RGC grant #601906.

### Références

- [1] O. ALLAIS & J. TRESSOU – « Using decomposed household food acquisitions as inputs of a kinetic dietary exposure model », *Statistical Modelling : an International Journal* (2009), à paraître, <http://hal.archives-ouvertes.fr/hal-00139914>.
- [2] S. ASMUSSEN – *Applied probability and queues*, Springer-Verlag, New York, 2003.
- [3] P. BARBE & W. P. MCCORMICK – *Asymptotic expansions for infinite weighted convolutions of heavy tail distributions and applications*, Memoirs of the American Mathematical Society, American Mathematical Society, 2009.
- [4] S. M. BARLOW, J. B. GREIG, J. W. BRIDGES, A. CARERE, A. J. M. CARPY, C. L. GALLI, J. KLEINER, I. KNUDSEN, H. B. W. M. KOËTER, L. S. LEVY, C. MADSEN, S. MAYER, J. F. NARBONNE, F. PFANNKUCH, M. G. PRODANCHUK, M. R. SMITH & P. STEINBERG – « Hazard identification by methods of animal-based toxicology », *Food Chem. Tox.* **40** (2002), p. 145–191.

- [5] J. BEIRLANT, G. DIERCKX, Y. GOEGEBEUR & G. MATTHYS – « Tail index estimation and an exponential regression model », *Extremes* **2** (1999), no. 2, p. 177–200.
- [6] P. BERTAIL, S. CLÉMENÇON & J. TRESSOU – « Extreme values statistics for Markov chains via the (pseudo-) regenerative method. », *Extremes* (2008), à paraître, <http://hal.archives-ouvertes.fr/hal-00165652>.
- [7] ———, « Statistical analysis of a dynamic model for food contaminant exposure with applications to dietary methylmercury contamination », 2008, en révision, <http://hal.archives-ouvertes.fr/hal-00308881>.
- [8] ———, « A storage model for modelling exposure to food contaminants. », *Mathematical Biosciences and Engineering* **5** (2008), no. 1, p. 35–60.
- [9] P. BERTAIL, M. FEINBERG, J. TRESSOU & P. V. (COORDINATEURS) – *Analyse des risques alimentaires*, TEC&DOC, Lavoisier, Paris, 2006.
- [10] P. BERTAIL & J. TRESSOU – « Incomplete generalized U-Statistics for food risk assessment », *Biometrics* **62** (2006), no. 1, p. 66–74.
- [11] P. BOUGEROL & N. PICARD – « Strict stationarity of generalized autoregressive processes », *Ann. Probab.* **20** (1992), no. 4, p. 1714–1730.
- [12] F. CÉROU, P. DELMORAL, F. LEGLAND & P. LEZAUD – « Genetic genealogical models in rare event analysis », *Alea* **1** (2006), p. 183–196.
- [13] F. CÉROU & A. GUYADER – « Adaptative splitting for rare event analysis », *Stoch. Anal. Proc.* **25** (2007), no. 2, p. 417–443.
- [14] D. CLAYTON & M. HILLS – *Statistical models in epidemiology*, Oxford Univ. Press, 1993.
- [15] CREDOC-AFSSA-DGAL – *Enquête inca (individuelle et nationale sur les consommations alimentaires)*, TEC&DOC éd., Lavoisier, Paris, 1999, (coordinateur : J.L. Volatier).
- [16] A. CRÉPET, H. HARARI-KERMADEC & J. TRESSOU – « Using empirical likelihood to combine data : application to food risk assessment », *Biometrics* **65** (2009), no. 1, à paraître, <http://dx.doi.org/10.1111/j.1541-0420.2008.01051.x>.
- [17] A. CRÉPET, J. TRESSOU, P. VERGER & J. C. LEBLANC – « Management options to reduce exposure to methyl mercury through the consumption of fish and fishery products by the French population », *Regul. Tox. Pharm.* **42** (2005), no. 2, p. 179–189.
- [18] J. J. DAUDIN & C. DUBY – *Techniques mathématiques pour l'industrie agroalimentaire*, TEC&DOC éd., Paris, 2002.
- [19] P. DAVIDSON, G. MYERS, C. COX, C. F. SHAMLAYE, T. CLARKSON, D. MARSH, M. TANNER, M. BERLIN, J. SLOANE-REVES, E. CERNICHIARI, O. CHOISY, A. CHOI & T. W. CLARKSON – « Longitudinal neurodevelopmental study of seychellois children following in utero exposure to mehg from maternal fish ingestion : Outcomes at 19-29 months », *Neurotoxicology* **16** (1995), p. 677–688.
- [20] M. DAVIS – *Applied stochastic analysis*, Stochastics Monographs, Taylor & Francis, 1991.
- [21] A. S. DEATON & J. MUELLBAUER – « An almost ideal demand system », *American Economic Review* **70** (1980), no. 3, p. 323–326.

- [22] E. DYBING, J. DOE, J. GROTEN, J. KLEINER, J. O'BRIEN, A. G. RENWICK, J. SCHLATTER, P. STEINBERG, A. TRITSCHER, R. WALKER & M. YOUNES – « Hazard characterisation of chemicals in food and diet : dose response, mechanisms and extrapolation issues », *Food Chem. Tox.* **40** (2002), p. 237–282.
- [23] L. EDLER, K. POIRIER, M. DOURSON, J. KLEINER, B. MILESON, H. NORDMANN, A. RENWICK, W. SLOB, K. WALTON & G. WÜRTZEN – « Mathematical modelling and quantitative methods », *Food Chem. Tox.* **40** (2002), p. 283–326.
- [24] B. EFRON & R. J. TIBSHIRANI – *An introduction to the Bootstrap*, Monographs on Statistics and Applied Probability, Chapman & Hall, 2004.
- [25] P. EMBRECHTS, C. KLÜPPELBERG & T. MIKOSCH – *Modelling extremal events for insurance and finance*, Applications of Mathematics, Springer-Verlag, 1997.
- [26] FAO/WHO – « Application of risk analysis to food standard issues », Tech. report, Report of the joint FAO-WHO consultation, Geneva, Switzerland, 1995, 13-17 march 1995.
- [27] ———, « Evaluation of certain food additives and contaminants for methylmercury », Sixty first report of the Joint FAO/WHO Expert Committee on Food Additives, Technical Report Series, WHO, Geneva, Switzerland, 2003.
- [28] A. FEUERVERGER & P. HALL – « Estimating a tail exponent by modelling departure from a Pareto Distribution », *Ann. Statist.* **27** (1999), p. 760–781.
- [29] M. GIBALDI & D. PERRIER – *Pharmacokinetics*, Drugs and the Pharmaceutical Sciences : a Series of Textbooks and Monographs, Marcel Dekker, New York, 1982, Second Edition.
- [30] P. GLASSERMAN, P. HEIDELBERGER, P. SHAHABUDDIN & T. ZAJIC – « Multilevel splitting for estimating rare event probabilities », *Oper. Research* **47** (1999), no. 4, p. 585–600.
- [31] P. GRANDJEAN, P. WEIHE, R. WHITE, F. DEBES, S. ARAKI, K. YOKOYAMA, K. MURATA, N. SORENSEN, R. DAHL & P. JORGENSEN – « Cognitive deficit in 7-year-old children with prenatal exposure to methylmercury », *Neurotox. Teratol.* **19** (1997), p. 417–428.
- [32] L. DE HAAN & L. PENG – « Comparison of tail index estimators », *Statist. Neerlandica* **52** (1998), p. 60–70.
- [33] B. M. HILL – « A simple general approach to inference about the tail of a distribution », *Ann. Statist.* **3** (1975), p. 1163–1174.
- [34] IFREMER – « Résultat du réseau national d'observation de la qualité du milieu marin pour les mollusques (RNO) », 1994-1998.
- [35] R. KROES, D. MÜLLER, J. LAMBE, M. R. H. LOWIK, J. VAN KLAVEREN, J. KLEINER, R. MASSEY, S. MAYER, I. URIETA, P. VERGER & A. VISCONTI – « Assessment of intake from the diet », *Food Chem. Tox.* **40** (2002), p. 327–385.
- [36] A. LAGNOUX-RENAUDIE – « Analyse des modèles de branchement avec duplication des trajectoires pour l'étude des événements rares », Thèse, Univ. Paul Sabatier, Toulouse III., 2006, <http://tel.archives-ouvertes.fr/tel-00129752>.
- [37] A. J. LEE – *U-statistics : Theory and practice*, Statistics : textbooks and monographs, vol. 110, Marcel Dekker, Inc, New York, USA, 1990.

- [38] MAAPAR – « Résultats des plans de surveillance pour les produits de la mer », Ministère de l’Agriculture, de l’Alimentation, de la Pêche et des Affaires Rurales, 1998-2002.
- [39] T. McMEEKIN, J. OLLEY, T. ROSS & D. RATKOWSKY – *Predictive microbiology : theory and application*, Research Studies Press, LTD, Taunton., 1993.
- [40] S. MEYN & R. TWEEDIE – « Stability of Markovian processes III : Foster-Lyapunov criteria for continuous time processes », *Adv. Appl. Probab.* **25** (1993), no. 3, p. 518–548.
- [41] P. D. MORAL – *Feynman-kac formulae : Genealogical and interacting particle systems with applications*, Probability and its applications, Springer, New York, 2004.
- [42] NRC (NATIONAL RESEARCH COUNCIL) – « Committee on the toxicological effects of methylmercury », Tech. report, Washington DC, 2000.
- [43] A. OWEN – *Empirical likelihood*, Chapman & Hall/CRC, 2001.
- [44] S. RACHEV & G. SAMORODNITSKY – « Limit laws for a stochastic process and random recursion arising in probabilistic modelling », *Adv. Appl. Probab.* **27** (1995), no. 1, p. 185–202.
- [45] A. G. RENWICK, S. M. BARLOW, I. HERTZ-PICCIOTTO, A. R. BOOBIS, E. DYBING, L. EDLER, G. EISENBRAND, J. B. GREIG, J. KLEINER, J. LAMBE, D. J. MÜLLER, M. R. SMITH, A. TRITSCHER, S. TUIJTELAARS, P. A. VAN DEN BRANDT, R. WALKER & R. KROES – « Risk characterisation of chemicals in food and diet », *Food Chem. Tox.* **41** (2003), p. 1211–1271.
- [46] G. O. ROBERTS & R. L. TWEEDIE – « Rates of convergence of stochastically monotone and continuous time Markov models », *J. Appl. Probab.* **37** (2000), no. 2, p. 359–373.
- [47] J. S. ROSENTHAL – « Minorization conditions and convergence rates for Markov chain Monte Carlo », *J. Amer. Stat. Assoc.* **90** (1995), p. 558–566.
- [48] J. SMITH & F. FARRIS – « Methyl mercury pharmacokinetics in man : A reevaluation », *Toxicol. Appl. Pharmacol.* **137** (1996), p. 245–252.
- [49] J. TRESSOU – « Méthodes statistiques pour l’évaluation du risque alimentaire », Thèse, Univ. Paris X, Nanterre, 2005, <http://tel.archives-ouvertes.fr/tel-00139909>.
- [50] ———, « Non parametric modelling of the left censorship of analytical data in food risk exposure assessment », *J. Amer. Stat. Assoc.* **101** (2006), no. 476, p. 1377–1386.
- [51] ———, « Bayesian nonparametrics for heavy tailed distribution. application to food risk assessment », *Bayesian Analysis* **3** (2008), no. 2, p. 367–392, A paraître, <http://hal.archives-ouvertes.fr/hal-00184755>.
- [52] J. TRESSOU, A. CRÉPET, P. BERTAIL, M. H. FEINBERG & J. C. LEBLANC – « Probabilistic exposure assessment to food chemicals based on extreme value theory. application to heavy metals from fish and sea products », *Food Chem. Tox.* **42** (2004), no. 8, p. 1349–1358.
- [53] P. VERGER, J. TRESSOU & S. CLÉMENÇON – « Integration of time as a description parameter in risk characterisation : application to methylmercury », *Regul. Tox. Pharm.* **49** (2007), no. 1, p. 25–30.

- [54] WHO – « Methylmercury, environmental health criteria 101 », Tech. report, Geneva, Switzerland, 1990.

---

*janvier 2009*

STÉPHAN CLÉMENÇON, Telecom ParisTech (TSI) - LTCI UMR Institut Telecom/CNRS  
No. 5141 • *E-mail* : [stephan.clemencon@telecom-paristech.fr](mailto:stephan.clemencon@telecom-paristech.fr)  
*Url* : <http://www.tsi.enst.fr/~clemenco/>

JESSICA TRESSOU, Unité Mét@risk - INRA UR1204, AgroParisTech, 16 rue Claude Bernard,  
75234 Paris Cedex 5, France • The Hong Kong University of Science and Technology -  
Department ISOM (prev. ISMT) - Clear Water Bay - Hong Kong. Tél : +852 2358  
7750 - Fax : +852 2358 2421 • *E-mail* : [Jessica.Tressou@agroparistech.fr](mailto:Jessica.Tressou@agroparistech.fr)  
*Url* : [http://www.paris.inra.fr/metarisk/members/tressou\\_jessica](http://www.paris.inra.fr/metarisk/members/tressou_jessica)