# Modélisation des Co-Expositions aux Pesticides : une Approche Bayésienne Nonparamétrique

Amélie Crépet, Jessica Tressou

# Modelisation des Co-Expositions aux Pesticides : une Approche Bayésienne Nonparamétrique

Amélie Crépet[a] & Jessica Tressou[b]

[a] AFSSA-DERNS-AQR-PC, Agence Française de Sécurité Sanitaire des Aliments, 27-31 Av. Général Leclerc, 94701 Maisons-Alfort, France

[b] INRA-Mét@risk, Méthodologies d'Analyse de Risque Alimentaire, Institut National de Recherche Agronomique, 16 rue Claude Bernard, 75231 Paris, France

**Mots-clés :** Processus de Dirichlet ; Modèle Bayésien Nonparamétrique ; Mélanges de lois normales multivariées ; Clustering ; Exposition multivariée ; Analyse de Risque Alimentaire

**Abstract**

This work introduces a specific application of the Bayesian nonparametric methodology in the food risk analysis framework. The goal is to determine mixture of pesticides residues which are simultaneously present in the diet, to give directions for future toxicological experiments for studying possible combined effects of those mixtures. Namely, the joint distribution of the exposures to a large number of pesticides is assessed from the available consumption data and contamination analyses. We propose to model the co-exposures by a Dirichlet process mixture based on a multivariate Gaussian kernel so as to determine clusters of pesticides jointly present in the diet at high doses. The posterior distributions and the optimal partition are computed through a Gibbs sampler based on stick-breaking priors. To reduce computational time due to the high dimensional data, a random block sampling is used. Finally, the clustering among individuals also obtained as an auxiliary output of these analyses is discussed in a risk management perspective.

**Résumé**

Ce travail présente l'utilisation d'un modèle bayésien nonparamétrique dans le cadre de l'évaluation du risque alimentaire. L'objectif est de déterminer les cocktails de pesticides auxquels la population française est exposée afin de mieux appréhender les risques liés à la présence de plusieurs substances chimiques dans l'alimentation. Ainsi, l'exposition de la population aux pesticides est estimée en combinant les données de consommation des aliments avec les données de contamination disponibles des denrées alimentaires. Nous proposons ensuite de modéliser ces co-expositions par un mélange de lois normales multivariées et d'en estimer les différentes composantes en choisissant pour la loi du mélange une distribution *a priori* selon un processus de Dirichlet, l'objectif étant de déterminer les clusters de pesticides simultanément présents à des niveaux élevés dans le régime alimentaire. Les distributions *a posteriori* ainsi que la partition optimale sont obtenues par un algorithme de Gibbs appliqué à la représentation "stick-breaking" du processus de Dirichlet. La grande dimension des données induit une convergence lente de l'algorithme que nous tentons de réduire en sous échantillonnant aléatoirement les variables d'expositions (random block Gibbs). Enfin, la classification des individus également obtenue est discutée d'un point de vue de gestion du risque.

# 1 Introduction

Ces travaux présentent un modèle bayésien nonparamétrique développé pour déterminer les mélanges de pesticides auxquels la population française est exposée. Ces mélanges seront ensuite étudiés d'un point de vue toxicologique afin de mieux appréhender les risques liés à la présence simultanée de plusieurs substances chimiques dans l'alimentation. Chaque aliment est en effet susceptible de contenir différents résidus de pesticides. Les consommateurs sont alors exposés à des mélanges de pesticides dont les effets combinés sur la santé sont inconnus. Selon les procédures actuelles, les substances actives sont évaluées individuellement et les normes sanitaires sont établies pour chaque substance. Cette approche se justifie par l'idée que la multiplicité des combinaisons et des modes d'actions possibles de ces substances rend la survenue d'effets combinés inattendue. Cependant, il existe très peu de données expérimentales permettant d'étayer cette hypothèse. Se posent alors deux questions : comment estimer l'exposition combinée des consommateurs à différents pesticides et déterminer les mélanges pertinents ? Comment évaluer leurs possibles effets combinés ?

Pour répondre à la première question, nous proposons de modéliser la co-exposition à l'ensemble des pesticides par un mélange de lois normales multivariées, dans un cadre bayésien. L'utilisation d'un processus de Dirichlet comme loi *a priori* de la distribution des composantes du mélange permet de ne pas faire d'hypothèse sur le nombre de ses composantes et ainsi de couvrir un large ensemble de distributions.

Dans une première partie, nous présentons les données disponibles et décrivons la construction de la distribution de la co-exposition aux pesticides. Puis le modèle bayésien nonparamétrique développé ainsi que sa mise en oeuvre pratique par un échantillonnage de Gibbs sont explicités. Nous proposons en particulier des outils pour réduire les temps de calcul particulièrement élevés dans le cas de données de grande dimension. Enfin, nous concluons par la présentation de résultats préliminaires.

# 2 Estimation de la co-exposition

L'exposition de la population aux substances chimiques par l'alimentation ne peut en général pas être mesurée directement. Elle est donc estimée en combinant les données de consommation alimentaire et les données de contamination des denrées. Nous nous intéressons uniquement aux pesticides présentant un risque aigu, c'est-à-dire les pesticides pour lesquels l'effet intervient peu de temps après l'ingestion.

**Données de contamination** Les données de contamination en résidus de pesticides des denrées alimentaires en France proviennent principalement des plans de surveillance et de contrôle mis en place par les ministères de l'agriculture et de l'économie, ainsi que de la base nationale gérée par le ministère de la santé pour l'eau destinée à la consommation humaine. L'année 2006 correspondant à celle étudiée pour la consommation est retenue.

Parmi les 300 résidus de pesticides analysés dans plus de 150 denrées d'origine animale et végétale, 54 pesticides ont été détectés (*i.e.*, au moins une analyse se révèle supérieure à la limite de détection, LOD) dans 141 denrées et eau de boisson, ce qui correspond à 6644 couples pesticide/denrée différents. Pour chaque couple pesticide/denrée, plusieurs analyses ont été effectuées, ce qui permet de construire une distribution empirique de contamination. Cette distribution est construite par tirage aléatoire uniforme, soit entre les différentes mesures quantifiées, soit entre 0 et la LOD pour les données censurées, en respectant la probabilité d'appartenance à chaque intervalle.

**Données de consommation** Les données de consommation des Français sont issues de l'*Enquête Individuelle Nationale de Consommation Alimentaire " INCA2"*, pilotée par l'Agence Française de Sécurité Sanitaire des Aliments, (Volatier, 2007). Cette enquête détaille la consommation alimentaire de 4079 individus sur une semaine en 2006. Le nombre d'aliments ou plats répertoriés est de 1280. Nous travaillerons sur deux sous-populations distinctes, celle composée de 1918 adultes et celle composée de 1444 enfants, après retrait des 717 sous-déclarants.

**Calcul de la co-exposition** L'exposition aiguë à un pesticide $p$ est calculée de la manière suivante : pour chaque denrée $a$ contaminée par le pesticide $p$, la quantité consommée $c_{ia}$ par l'individu $i$ de poids corporel $w_i$ sur un des 7 jours d'enquête pris au hasard est multipliée par une valeur $q_{pam}$ tirée aléatoirement dans la distribution de contamination de la denrée $a$ par le pesticide $p$. En sommant sur les différentes denrées $a = 1, ..., A_p$, contenant le pesticide $p$, on obtient une valeur d'exposition $x_{pim} = \sum_{a=1}^{A_p}(c_{ia} \times q_{pam})/w_i$ pour l'individu $i$ exprimée en milligrammes de pesticide $p$ par kilogramme de poids corporel (mg/kg pc) ($p = 1, \ldots, P$, $i = 1, \ldots, n$, $m = 1, \ldots, M$). Afin d'exprimer l'ensemble des expositions dans une même échelle, les données sont normalisées après avoir subies une log-transformation.

# 3 Modèle bayésien nonparamétrique

Les modèles nonparamétriques peuvent généralement être définis comme des modèles paramétriques avec une infinité de paramètres, (Bernardo and Smith, 1994 ; Muller and Quintana, 2004).

**Mélange de lois et processus de Dirichlet** Une approche classique pour grouper des données est de considérer que les données sons issues d'un mélange de lois de probabilité. Considérons que les co-expositions décrites par les $i = 1, ..., n$ vecteurs $x_i = \{x_{i1}, \ldots, x_{iP}\}$ sont indépendamment et identiquement distribuées selon la densité de probabilité suivante :

$$f(x) = \int_{\Theta} k(x|\theta)G(d\theta)$$

où $k(.|\theta)$ est la densité mélangée connue de paramètre $\theta$ (par exemple, une gaussienne avec $\theta = (\mu, \Sigma)$) et $G$ est la distribution de mélange supposée inconnue. Dans un cadre

bayésien, on choisit une distribution *a priori* $P(G)$ pour le paramètre $G$, distribution aléatoire appartenant à un espace de fonctions de dimension infinie. L'équation précédente peut être reformulée sous la forme hiérarchique suivante :

$$\begin{aligned} G &\sim P(G) \\ \theta_i|G &\sim G(d\theta) \\ x_i|\theta_i &\sim k(dx|\theta_i) \end{aligned} \tag{1}$$

Plusieurs distributions *a priori* pour $G$ sont envisageables, (Muller and Quintana, 2004 ; Walker et al., 1999). Dans le cadre de l'estimation de densité, les processus de Dirichlet (Dirichlet Process, DP) ou leurs extensions restent les plus largement employés, (Muller and Quintana, 2004 ; Lo, 1984). Ces processus ont été introduits par (Ferguson, 1973) et sont définis à partir de deux paramètres, un paramètre d'échelle $\gamma$ et une mesure aléatoire de base $H$. On dit que la fonction $G$ à valeurs dans $E$ est distribuée selon un processus de Dirichlet de paramètres $\gamma$ et $H$ si pour toute partition $(A_1, \ldots, A_k)$ de $E$, le vecteur de probabilités aléatoires $(G(A_1), \ldots, G(A_k))$ suit une distribution de Dirichlet vectorielle standard de paramètre $(\gamma H(A_1), \ldots, \gamma H(A_k))$. On note ceci simplement par $G \sim DP(dG|\gamma, H)$. Parmi les propriétés de ces processus, notons que pour tout borélien $B$ de $E$

$$\mathbb{E}[G(B)] = H(B) \text{ et } \mathbb{V}[G(B)] = \frac{H(B)(1 - G_H(B))}{1 + \gamma}.$$

Dans notre application, nous choisissons d'utiliser pour le noyau $k$ une loi normale multivariée afin de prendre en compte la corrélation entre les expositions aux différents pesticides. En ne considérant dans un premier temps qu'une seule des $M$ valeurs de co-exposition (par exemple celle correspondant au $95^{\text{ème}}$ percentile d'exposition de chaque pesticide), on suppose que $x_i = (x_{i1}, \ldots, x_{iP}) \sim \int_\Theta k(x_i|\theta)G(d\theta)$, où $\theta = (\mu, \tau)$ désigne le vecteur des $P$ moyennes et la matrice de précision (inverse de la matrice de variance-covariance) de la loi normale multivariée. L'utilisation du processus de Dirichlet comme loi *a priori* pour les paramètres du mélange assure que le nombre de valeurs distinctes de $\theta_i$ est inférieur à $n$, permettant ainsi de constituer des clusters d'individus partageant le même type de co-exposition. Les cocktails de pesticides seront recherchés au sein de ces clusters homogènes. Pour la mesure de base $G_0$ du processus de Dirichlet, nous retenons la loi Wishart-Normale notée $WN(\alpha, \Psi, m, t)$, loi conjuguée pour la normale multivariée, définie par $\tau \sim W(\alpha, \Psi)$ et $\mu|\tau \sim N(m, (t\tau)^{-1})$.

La variabilité de la contamination peut être prise en compte dans la modélisation en intégrant un niveau hiérarchique supplémentaire sur la distribution des variables latentes, (Teh et al., 2006). Ainsi en notant $x_{im} = (x_{pim}, p = 1, \ldots, P)$ pour $i = 1, \ldots, n$ et

$m = 1, \ldots, M$, le modèle devient

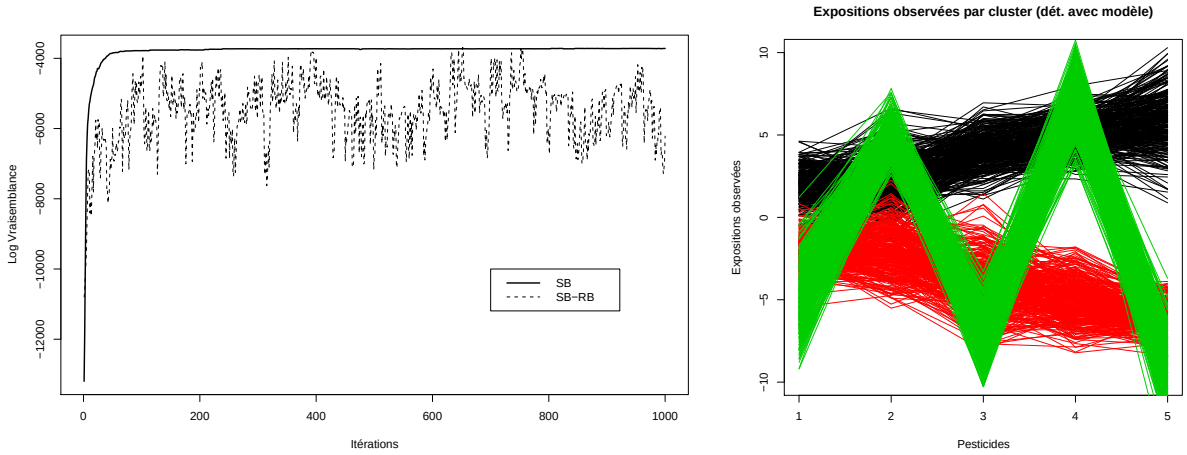
$$\begin{aligned} G_0 &\sim DP(dG|\gamma, H) \\ G_i &\sim DP(dG|\alpha_i, G_0) \\ \theta_{im}|G_i &\sim G_i \\ x_{im}|\theta_{im} &\sim k(.|\theta_{im}) \end{aligned} \tag{2}$$

**Inférence par échantillonnage de Gibbs** L'inférence de tels modèles fondés sur des processus de Dirichlet est en général conduite à partir de méthodes de Monte carlo par Chaînes de Markov (MCMC), et plus particulièrement selon des algorithmes de Gibbs (Ishwaran and James, 2001). Les plus utilisés sont fondés sur la distribution en urne de Pólya ou processus du restaurant chinois (Blackwell and MacQueen, 1973 ; Escobar and West, 1995 ; Lau and Lo, 2007), et la représentation "stick-breaking" (SB) du processus de Dirichlet (Ishwaran and James, 2001). La différence entre les deux approches relève principalement du fait que, pour le premier, il faut disposer d'une règle de prédiction pour affecter les individus à un cluster sans effectuer de tirage des variables latentes alors que l'approche stick-breaking inclut le tirage de ces variables latentes. Nous avons retenu l'approche stick-breaking en particulier pour la simplicité de sa mise en oeuvre dans le cas du modèle hiérarchique (2). Si $G \sim DP(dG|\alpha_0, H)$, la distribution $G$ peut s'écrire $G = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$ où les $\phi_k$ sont des variables aléatoires issues de la distribution $H$, $\delta_{\phi_k}$ est une masse de Dirac en l'atome $\phi_k$ et $\beta_k$ sont les poids SB dépendant du paramètre $\gamma$. Pour une valeur raisonable de $N < \infty$, $G$ peut être approximée par les $N$ premières masses de Dirac, voir le théorème 2 d'Ishwaran and James (2001). Les $N$ poids SB sont construits à partir de poids auxiliaires $\beta_k^* \sim Beta(1, \gamma)$ de la manière suivante : $\beta_1 = \beta_1^*, \qquad \beta_k = \beta_k^* \prod_{l=1}^{k-1}(1 - \beta_l^*)$ pour $k = 2, ..., N-1$, et $\beta_k = 1 - \sum_{k=1}^{N-1} \beta_k$. Pour le modèle (1), un cycle de Gibbs consiste en le tirage successif des paramètres $\theta_j^*$ dans chaque cluster $j$, de la réaffectation aléatoire des individus à une nouvelle partition (n'ayant pas nécessairement le même nombre de clusters) à partir des poids SB, du noyau $k(.|\theta)$ et des $\theta_j^*$ tirés dans l'étape précédente, et enfin à la mise à jour des poids SB, les $\beta_k$, sachant cette nouvelle partition. Dans le cas du modèle (2), $n$ nouvelles séries de poids SB (on peut les noter $\pi_{ik}$, $k = 1, \ldots, N$, $i = 1, \ldots, n$) sont introduites et une étape de tirage de ces poids est ajoutée dans le cycle de Gibbs. Dans chaque cas, la ou les partitions optimales sont déterminées en maximisant la log-vraisemblance *a posteriori* obtenue sur chaque cycle de Gibbs.

Les algorithmes de Gibbs fondés sur le calcul itératif sont chronophages, en particulier lorsque la dimension des données est grande (ici la dimension de la loi normale multivariée est $P$). Nous proposons de réduire les temps de calculs selon une méthode analogue à celle proposée par Cabrera et al. (2009) appelée "Random Block Gibbs Weighted Chinese Restaurant" et qui consiste à sous échantillonner aléatoirement les variables observées, *i.e.,* en n'utilisant que $d$ valeurs au hasard parmi les $P$ disponibles dans chaque cycle de Gibbs, ceci dans le cas d'un algorithme de type processus de restaurant chinois.

Nous adoptons cette approche dans le cas de la représentation stick-breaking ("Random Block Stick-Breaking", RB-SB dans la suite). Ainsi, en considérant le vecteur des entiers sélectionnés $v_d = \{l_1, \ldots, l_d\}$, le jeu de données sous-échantillonné est composé des expositions $\{x_{i[v_d]}, i = 1, \ldots, n\}$ plutôt que de l'ensemble des $\{x_i, i = 1, \ldots, n\}$, permettant de réduire de manière importante les temps de calcul (tirages aléatoires et calcul de densités gaussiennes en dimension $d << P$).

# 4 Résultats préliminaires




(a) Log-vraisemblance sur les 1000 premières itérations

(b) Partition optimale obtenue avec l'algorithme SB : mélange de 3 gaussiennes

FIG. 1 – Comparaison des algorithmes Stick-breaking (SB) et "Random-Block Stick-Breaking" (RB-SB) sur données simulées ($N = 30$ atomes, 30000 itérations au total).

**Sur données simulées** En utilisant le premier jeu de $n = 1000$ données de Cabrera et al. (2009) correspondant à un mélange de 3 gaussiennes en dimension $P = 5$, la partition optimale est obtenue pour une valeur de log vraisemblance de $-3706$ après 1511 itérations seulement (sur un total de 30000 itérations), *cf.* Fig. 1(a). Cette partition comporte 3 clusters et les valeurs des paramètres sont similaires à celles qui ont permis de générer les données, *cf.* Fig. 1(b). En utilisant l'agorithme RD-SB avec $d = 2$, le maximum de vraisemblance est atteint plus rapidement (à l'itération 652), *cf.* Fig. 1(a).

**Sur données réelles** Le modèle n'a été testé que sur une partie des données d'exposition. L'analyse des matrices de corrélation de chaque cluster permet de mettre en évidence des groupes de pesticides présents simultanément à des niveaux élevés dans l'alimentation de certaines sous-populations. Les résultats seront présentés lors de la conférence.

## Bibliographie

[1] J. Bernardo and A. Smith. *Bayesian theory*. 1994.
[2] D. Blackwell and J. B. MacQueen. Ferguson distributions via polya urn schemes. *Annals of Statistics*, 1 :353–355, 1973.
[3] J. Cabrera, J. W. Lau, and A. Y. Lo. Random block sampling for high dimensional clustering (from the bayesian point of view). *Working Paper*, 2009.
[4] M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90 :577–588, 1995.
[5] T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1 :209–230, 1973.
[6] H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96 :161–173, 2001.
[7] J. W. Lau and A. Y. Lo. Model based clustering and weighted Chinese restaurant processes. In *Advances in Statistical Modeling and Inference : Essays in Honor of Kjell A. Doksum*, pages 405–424, 2007.
[8] A. Y. Lo. On a class of bayesian nonparametric estimates : I. density estimates. *Annals of Statistics*, 12(1) :351–357, 1984.
[9] P. Muller and F. Quintana. Nonparametric bayesian data analysis. *Statistical Science*, 19(1) :95–110, 2004.
[10] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476) :1566–1581, 2006.
[11] J. L. Volatier. Etude INCA2 : objectifs et méthodes. colloque PNNS la situation nutritionnelle en france en 2007 , 2007.
[12] S. Walker, P. Damien, P. Laud, and A. Smith. Bayesian nonparametric inference for random distributions and related functions. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 61(3) :485–527, 1999.