Bayesian nonparametric model for clustering individual co-exposure to pesticides found in the French diet.

Amélie Crépet, Jessica Tressou

Bayesian nonparametric model for clustering individual co-exposure to pesticides found in the French diet.

Amélie Crépet [a] & Jessica Tressou[b]

February 3, 2011

[a]ANSES, French Agency for Food, Environmental and Occupational Health Safety, 27-31 Av. Général Leclerc, 94701 Maisons-Alfort, France

[b]INRA-Met@risk, Food Risk Analysis Methodologies, National Institute for Agronomic Research, 16 rue Claude Bernard, 75231 Paris, France

Keywords

Dirichlet process; Bayesian nonparametric modeling; multivariate Normal mixtures; clustering; multivariate exposure; food risk analysis.

Abstract

This work introduces a specific application of Bayesian nonparametric statistics to the food risk analysis framework. The goal was to determine the cocktails of pesticide residues to which the French population is simultaneously exposed through its current diet in order to study their possible combined effects on health through toxicological experiments. To do this, the joint distribution of exposures to a large number of pesticides, which we called the co-exposure distribution, was assessed from the available consumption data and food contamination analyses. We proposed modelling the co-exposure using a Dirichlet process mixture based on a multivariate Gaussian kernel so as to determine groups of individuals with similar co-exposure patterns. Posterior distributions and optimal partition were computed through a Gibbs sampler based on stick-breaking priors. The study of the correlation matrix of the sub-population co-exposures will be used to define the cocktails of pesticides to which they are jointly exposed at high doses. To reduce the computational burden due to the high data dimensionality, a random-block sampling approach was used. In addition, we propose to account for the uncertainty of food contamination through the introduction of an additional level of hierarchy in the model. The results of both specifications are described and compared.

1 Introduction

Each food product may contain several pesticide residues, consequently meals ingested daily may include a wide range of pesticides. All consumers are thus exposed to complex cocktails of pesticides

whose combined effects on health are still unknown. This work proposes an original methodology to respond to the following question "what are the cocktails of pesticides to which the French population is simultaneously and most heavily exposed?" Cocktails of pesticides were selected based on their joint probability of occurring at high doses in the French diet, according to the following process. The French population exposure to P different pesticides found in the diet, called the co-exposure, was estimated considering both the food residue level patterns generated from national food monitoring administrations in charge of food control and the dietary habits of n surveyed individuals from the INCA 2 French consumption survey (AFSSA, 2009). A nonparametric mixture model was developed to cluster the population co-exposures in order to define groups of individuals with similar patterns of exposures to the P pesticides. For the biggest clusters made of the highly exposed individuals, the correlations between exposures to the P pesticides were studied to characterize the relevant cocktails.

To define homogeneous groups of individuals, the population co-exposure to the P pesticides was modeled with a Bayesian nonparametric model relying on the use of the Dirichlet process, (Ferguson, 1973; Antoniak, 1974). This approach consists in building a mixture model, in which the number of mixture components is potentially unlimited, and is itself a random variable that is part of the overall model. In our pesticide study, a multivariate Normal distribution was chosen as the kernel density of the mixture. Thus, the correlations of the P pesticides were modeled and individuals were clustered according to both their co-exposure levels and their co-exposure correlations. The mixing distribution was modeled with a Dirichlet process (DP) which is the most popular choice of prior distribution for infinite mixture models in a Bayesian nonparametric context, Lo (1984). Indeed, the DP can be viewed as a probabilistic measure on the space of probability measures which has the required properties defined by Ferguson (1973), to be used as a prior distribution. Models combining both mixture model and Dirichlet process are called Dirichlet Process Mixture (DPM) models. This approach provides a way of putting a distribution on all possible partitions of the data and then, through the corresponding posterior distribution or classification likelihood, selecting the most likely cluster formations. To include the uncertainty of individual exposure to each pesticide, we modified the base model using a hierarchical DP approach, similar to the one proposed by Teh et al. (2006). In the case of a real application to a large sample size, like the one under study, scanning all possibilities is computationally unfeasible and thus there is a need to resort to simulation methods such as Markov Chain Monte Carlo techniques. Popular choices for such simulation technique include Gibbs samplers based either on the Pólya urn scheme which is closely related to the Chinese restaurant process (CRP, Blackwell and MacQueen, 1973; Pitman and Yor, 1996), or on the stick-breaking (SB) representation of the DP (Sethuraman, 1994). The latter was retained to account for the complexity of our model's hierarchy within an effective algorithm, Ishwaran and James (2001). The stick-breaking priors can be simply constructed using a sequence of independent Beta random variables and the SB algorithm does not require individuals to be

reassigned one by one like the CRP, which becomes computationally complex with a large sample size. Still, to reduce this computational burden further, the random-block sampling proposed in the technical report of the Department of Informations Systems, Business Statistics and Operations Management, Hong Kong University of Science and Technology (Cabrera et al., 2009), for the CRP was applied to the SB algorithm to reduce computational time. This procedure consists in subsampling d variates among the available P dimensions at each iteration.

The following section describes the data on residue levels, quantities consumed and the co-exposure estimate. In Section 3, the Dirichlet Process Mixture model applied to the modeling of pesticide co-exposure is outlined. Then, the chosen prior distributions and the model in its hierarchical form are presented. For reproducibility purposes, the SB and the random-block SB algorithms are detailed. Finally, in section 4 the models are applied to a set of simulated data and to the French population's co-exposure to pesticides.

2 Co-exposure to pesticides

Pesticide food exposure was estimated with the individual food consumption data from the French dietary survey and with residue levels obtained from French pesticide residue monitoring programs. It is necessary to identify and take into account all foodstuffs in which significant residues might occur, as well as all pesticides that may be present in the food. Therefore, a first step consists in identifying the food / pesticide combinations to consider for the exposure assessment.

2.1 Data on pesticide residues levels

The data source on pesticide residues in food and drinking water corresponds to the annual monitoring programmes implemented in 2006 by the French administrations (Ministry of Economy, Ministry of Agriculture, Ministry of Health). These surveys provide sample distributions of residues for up to 300 pesticides measured in about 150 raw agricultural commodities (RACs). The number of samples varies from about 10 for minor commodities up to 480 for staples (apple, lettuce, etc.). Most of analytical results are left-censored, i. e. the residue level stands between 0 and the limit of reporting (LOR) from the laboratory but cannot be quantified or detected. For one pesticide, when more than 90% of results were censored in each food commodity then, it was considered of no interest to take it into account for the co-exposure calculation. Some pesticides have however been included in the study, when the determined residue levels were of the same order as the corresponding LOR. In such cases, it was considered that the pesticide may really be present but could not be determined due to analytical limitations. Thus, 79 pesticides were selected for the analysis. Residues of the selected pesticides were analyzed in 120 RACs and in drinking water consumed by the INCA2 population. A total of $306,899$ analytical results corresponding to $8,364$ combinations of pesticide/commodity were used in this work.

2.2 Food consumption data

Consumption data were provided by the second "Individual and National Study on Food Consumption", INCA2 survey, carried out by the French Food Safety Agency, AFSSA (2009). The study was conducted in three fieldwork phases between late 2005 and April 2007 in order to cover seasonal variations. Two independent population groups were included in the study: $2,624$ adults aged 18-79 years and $1,455$ children aged 3-17 years. Participants were selected using a three-stage random probability design stratified by region of residence, size of urban area and population category (adults or children). Each subject was asked to complete a seven-day food diary as well as other questionnaires on anthropometric and socio-economic factors. Food were subsequently categorized into $1,305$ "as consumed" food items (INCA2 classification). In order to match the consumption data to the residues ones, the food items defined in the INCA2 survey were broken down into 181 RACs. To do this, 763 standardized recipes were used, which were defined by the French Food Safety Agency taking account of industrial processes, home cooking habits and edible portions for the INCA2 survey (AFSSA, 2009).

In the INCA2 survey, sampling weights are provided for each surveyed individual representing its frequency in the entire French population. Based on the sampling weights provided for each individual, two samples of adults and children were built out of the original samples by carrying out random trials with replacement. Only normal-reporters, i.e. individuals whose energy needs are covered by the declared consumptions, were considered for this study. Therefore, two normal-reporter samples of $1,898$ adults and $1,439$ children were used for the analysis.

2.3 Dietary co-exposure assessment

To estimate acute exposure, i.e the exposure during 24-hour day, one day of consumption was randomly selected for each individual from the 7 days recorded in the INCA2 survey. The individual daily consumption of a commodity denoted c_{ia} corresponds to the sum of all the quantities of commodity a consumed by the individual i during the selected day. For each commodity a treated with the pesticide p, the daily consumption c_{ia} was multiplied by one residue level q_{pa} and adjusted by the body weight w_i of the consumer i. The intakes calculated for each commodity were summed to obtain a total daily exposure in milligrams of the pesticide p per kilogram of body weight of the consumer per day (mg/kg bw/d). This process was performed for $m = 1, \ldots, M$ values randomly selected from the contamination distribution of each pesticide/commodity combination to account for the residue level uncertainty. The final data set consisted of a series of M possible daily exposures $x_{pim} = \sum_{a=1}^{A_p} (c_{ia} \times q_{pam})/w_i$ to each pesticide $p = 1, \ldots, P$, for each individual $i = 1, \ldots, n$. In order to deal with quantitative values, each censored datum was uniformly selected between 0 and its censoring value (LOR). Similarly, for each pesticide/commodity combination, random contamination values were uniformly selected among the different residues levels, with respect to

the probability of lying in the interval between two consecutive observed residues levels. Scaling problems between pesticide exposure levels were ruled out by using a log scale, centering around the mean and rescaling by the standard deviation for each pesticide and across individuals.
Then, two datasets were created, one considering the 95^{th} percentile of the distribution of the M exposures to the pesticide p of each individual i (one high exposure per individual), the other one considering the entire distribution empirically described by the M exposure values. High percentiles such as the 95^{th}, 99^{th} or the maximum of the exposure distribution are usually employed to study the worst case of exposure to a chemical. In the context of co-exposure assessment, this scenario could be quite unrealistic in the sens that along a day the probability of being jointly exposed to high values for all pesticides is very low. Random co-exposures are more likely to occur in current life. In this way, using the distribution of exposure for each pesticide permits to attribute to each individual random values of exposure integrating the uncertainty of residue levels of each pesticide. Therefore, computations were carried out, in the first case with a co-exposure matrix of size $n \times P$ and in the second case with an array of size $n \times P \times M$.
Sensitivity of the exposure estimate to the value of M was tested in comparing the distributions of the exposure obtained with $M = 100$ and $M = 1000$ using a Wilcoxon test. For the alternative hypothesis "two sided", the p-value ranges between 0.80 and 0.96 depending on the pesticide and for the alternative hypothesis "greater", the p-value ranges between 0.40 and 0.55. So no statistical difference between both distributions were observed, and therefore computations were performed with $M = 100$.

3 Methodology

3.1 Bayesian nonparametric model-based clustering

Dirichlet Process Mixture model

A common approach to assign data to clusters is to construct a model in which data are generated from a mixture of probability distributions. In this way, the co-exposures of the n individuals to the P pesticides arise from a distribution composed of different sub-distributions, namely the mixture components. Therefore, the groups of individuals with similar patterns of pesticides co-exposure are identified as the ones sharing the same sub-distributions. Let the observed co-exposures $x = (x_1, \ldots, x_i, \ldots, x_n)$ with x_i a P dimensional vector $x_i = (x_{i1}, \ldots, x_{ip}, \ldots, x_{iP})$, be distributed with a probability density

$$f(x_i) = \int_{\Theta} k(x_i|\theta) G(d\theta) \tag{1}$$

where $k(.|\theta)$ is the known density of the mixture components called the kernel density, with parameter $\theta \in \Theta$ and G the unknown mixing distribution. Under a nonparametric perspective, the unknown distribution G is one of an infinite-dimensional function space. Equation (1) can

be broken down by introducing the latent variables θ_i that will be shared among individuals with similar patterns of pesticides co-exposure, as follows

$$\begin{aligned} x_i|\theta_i &\sim k(dx|\theta_i) \\ \theta_i|G &\sim G(d\theta) \\ G &\sim P(G). \end{aligned} \tag{2}$$

In a Bayesian approach, the challenge is to place an appropriate prior $P(G)$ on the distribution G. A collection of distribution functions called random probability measures (RPMs) can be assigned to the density G (Walker et al., 1999; Muller and Quintana, 2004). Ferguson (1973) stated the properties of this class of measures and introduced the Dirichlet Process (DP) as one of the RPMs. The DP is defined by two parameters, a scaling parameter γ and a base probability measure H. The probability distribution G is drawn from a DP, denoted $G \sim DP(\gamma, H)$, if and only if for any partition $(A_1, \ldots, A_k)$ of Ω, the vector of random probabilities $(G(A_1), \ldots, G(A_k))$ is drawn from a Dirichlet distribution $(G(A_1), \ldots, G(A_k)) \sim Dir(\gamma H(A_1), \ldots, \gamma H(A_k))$.

Partition Models

A clustering of n objects can be represented by a partition, denoted as $\mathbf{p}$ (Quintana and Iglesias, 2003). The partition $\mathbf{p}$ separates the n vectors x_i into $n(\mathbf{p})$ groups of individuals. The partition of size $n(\mathbf{p}) \in \{1, \ldots, n\}$, can be represented as $\mathbf{p} = \{C_1, \ldots, C_{n(\mathbf{p})}\}$ where C_j denotes the j^{th} cluster for $j = 1, \ldots, n(\mathbf{p})$. Equation (1) can be expressed conditionally on the partition $\mathbf{p}$, as a classification likelihood (Lau and Lo, 2007)

$$f(x|\mathbf{p}) = \prod_{j=1}^{n(\mathbf{p})} k(x_i, i \in C_j),$$

where $k(x_i, i \in C_j)$ is given by $k(x_i, i \in C_j) = \int_\Theta \prod_{i \in C_j} k(x_i|\theta) G(d\theta)$.

In the context of this classification likelihood, the partition $\mathbf{p}$ is the parameter for which a prior/posterior analysis is required. The prior distribution of $\mathbf{p}$ is induced by the distribution $P(G)$ and is proportional to $\prod_{j=1}^{n(\mathbf{p})} g(C_j)$, where g (known as the *cohesion*) is a function of the cluster only, e.g. its size. By Bayes theorem, the posterior distribution of $\mathbf{p}$ is the following product over partition components

$$\pi(\mathbf{p}|x) \propto \prod_{j=1}^{n(\mathbf{p})} g(C_j) k(x_i, i \in C_j). \tag{3}$$

Choosing a Dirichlet process $DP(\gamma, H)$ as the prior distribution $P(G)$ is equivalent to considering $g(C_j) = \gamma \Gamma(e_j)$, where e_j is the size of cluster C_j, see Lau and Lo (2007) for other settings.

An estimate of the optimal partition is the one that maximizes the posterior distribution (3),

which is approximated in this paper with a Gibbs sampler as described in Section 3.3. The number of clusters in the optimal partition represents the number of sub-populations of individuals with similar pesticide co-exposure patterns.

Stick-breaking representation of the Dirichlet process

If $G \sim DP(\gamma, H)$ then it can be represented as an infinite mixture of point masses

$$G = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$$

where ϕ_k are random variables sampled from H, δ_{ϕ_k} refers to a point mass concentrated at atom ϕ_k and β_k are the "stick-breaking weights" depending on γ. Drawing θ_i from G in equation 2 means that θ_i is equal to one of ϕ_k with the associated probability β_k.

One way to deal with such infinite mixtures is to use truncation (see Walker (2007) for a slice sampler based alternative). Ishwaran and James (2001) have shown that when truncating the sum to a reasonable N ($N < \infty$), the quality of approximation of G is good. The random weights β_k can therefore be built from auxiliary weights $\beta_k^* \sim Beta(1, \gamma)$ through the stick-breaking procedure given by

$$\beta_1 = \beta_1^*, \qquad \beta_k = \beta_k^* \prod_{l=1}^{k-1} (1 - \beta_l^*) \text{ for } k = 2, \ldots, N-1, \text{ and } \beta_N = 1 - \sum_{k=1}^{N-1} \beta_k.$$

In our study, clusters of individuals with similar patterns of co-exposure are identified as the ones sharing the same atoms ϕ_k.

3.2 Specific models

Multivariate Normal mixture model

A P dimensional multivariate Normal distribution $N_P(\mu, \tau^{-1})$ with mean vector $\mu \in R^P$ and random covariance matrix $\tau^{-1} \in R^{P \times P}$ is assigned to the kernel density k.

The distribution H is chosen to be a Wishart-Normal $(\mu, \tau) \sim WN(\alpha, \Psi, m, t)$ distribution due to its conjugate properties with the multivariate Normal distribution. Writing $(\mu, \tau) \sim WN(\alpha, \Psi, m, t)$ means that a Wishart distribution is used for the symmetric and positive definite precision matrix τ as $\tau \sim W(\alpha, \Psi)$, where α is a scalar degree of freedom and Ψ a $P \times P$ scale matrix; and conditionally to τ, the random vector μ is assigned a P-dimensional Normal distribution $\mu|\tau \sim N_P\left(m, (t\tau)^{-1}\right)$. Hence, we have

$$\begin{aligned} H(d\mu, d\tau) &= \left\{ 2^{-\alpha P/2} |\Psi|^{\alpha/2} (\Gamma_P(\alpha/2))^{-1} \times |\tau|^{(\alpha - P - 1)/2} \exp\left[-\frac{1}{2} Tr(\Psi\tau) \right] \right\} \\ &\quad \times \left\{ (2\pi)^{-P/2} |t\tau|^{1/2} \exp\left[-\frac{t}{2} (\mu - m)' \tau (\mu - m) \right] \right\} d\mu d\tau \end{aligned}$$

where Γ_P is the multivariate Gamma function $\Gamma_P(\alpha/2) = \pi^{P(P-1)/4} \prod_{r=1}^{d} \Gamma(\frac{\alpha+1-r}{2})$ and $Tr(A)$ is the trace of the matrix A.

The marginal density $k(x_i, i \in C_j)$ is obtained as

$$k(x_i, i \in C_j) = \int \int \prod_{i \in C_j} k(x_i|\mu, \tau) H(d\mu, d\tau).$$

With respect to our distribution choice, the marginal density is written as

$$k(x_i, i \in C_j) = \frac{\Gamma_P(\alpha_j^*/2)}{\Gamma_P(\alpha/2)} \frac{t^{P/2}}{\pi^{Pe_j/2} \left(t_j^*\right)^{P/2}} \frac{|\Psi|^{\alpha/2}}{\left|\Psi_j^*\right|^{\alpha_j^*/2}},$$

where $(^*_j)$ indicate the updated values of the parameters of the Wishart-Normal denoted $WN(\alpha_j^*, \Psi_j^*, m_j^*, t_j^*)$, and equal to

$$\alpha_j^* = \alpha + e_j,\ m_j^* = \frac{tm + e_j \overline{x_j}}{t_j^*},\ t^* = t + e_j,\ \Psi_j^* = \Psi + S_j + \frac{e_j t}{t_j^*}(m - \overline{x_j})(m - \overline{x_j})',$$

where e_j is the number of observations classified in the cluster C_j, $\overline{x_j} = \frac{1}{e_j} \sum_{i \in C_j} x_i$ is the mean of the cluster C_j and $S_j = \sum_{i \in C_j} (x_i - \overline{x_j})(x_i - \overline{x_j})$ is the corresponding covariance matrix.

The optimum number of mixture components $n(\mathbf{p})$ and the posterior distribution of the component mixture parameters are obtained by maximizing the following criterion (similar to Eq. 14 in Lau and Lo (2007)) corresponding up to a constant to the posterior empirical log likelihood (i.e. the log of equation (3))

$$Q(\mathbf{p}) = n(\mathbf{p}) \times \ln(\gamma) + \sum_{j=1}^{n(\mathbf{p})} \ln \Gamma(e_j) + \sum_{j=1}^{n(\mathbf{p})} \ln k(x_i|i \in C_j). \tag{4}$$

Hierarchical model to account for the uncertainty of the exposure

An additional Dirichlet process is used to account for the uncertainty of the exposure when considering the set of data $x_{im} = \{x_{pim}, p = 1, \ldots, P\}$ for each individual $i = 1, \ldots, n$ and the contamination value $m = 1, \ldots, M$

$$\begin{aligned} x_{im}|\theta_{im} &\sim k(.|\theta_{im}) \\ \theta_{im}|G_i &\sim G_i \\ G_i &\sim DP(\alpha_i, G_0) \\ G_0 &\sim DP(\gamma, H). \end{aligned} \tag{5}$$

With such a model, the co-exposure to the P pesticides of each individual i is composed of several sub-distributions identified as the ones sharing the same θ_{im}. Therefore, the form of the individual exposures to each pesticide given by the M values is considered into the clustering process.

3.3 Algorithm

3.3.1 Stick-breaking algorithm

3.3.1.a Base model

The algorithm of the base model described in Equation (2) is presented below in three steps[1]. Considering $G \sim DP(\gamma, H)$, we use the stick-breaking representation of the Dirichlet process $G = \sum_{k=1}^{N} \beta_k \delta_{\phi_k}(.)$, where ϕ_k denotes the hidden parameters of the multivariate Gaussian distribution (μ_k, τ_k), $\beta = (\beta_1, \ldots, \beta_N)$ are the stick-breaking weights, and N is the maximum number of atoms of G. A vector $K = (K_i)$ is introduced to store the assignment of each data point $x_i = (x_{ip}, p = 1, \ldots, P)$ for $i = 1, \ldots, n$ to an atom $(\phi_k)_{k=1,\ldots,N}$ so that K_i is an integer from 1 to N. Only N^* of the N available atoms are distinct values whose set is denoted by K^*.

1. Sampling $(\phi|K, \beta, X)$: for those k in K^*, sample ϕ_k with respect to the "updated" base measure H_k^* (a Wishart-Normal with parameters $\alpha_k^*, \Psi_k^*, m_k^*, t_k^*$ obtained from the posterior distribution given by $\{x_i, K_i = k\}$) and for the remaining $(N - N^*)$ atoms, get ϕ_k from the base measure H (the prior Wishart-Normal(α, Ψ, m, t)).

2. Sampling $(K|\beta, \phi, X)$: for $k = 1, \ldots, N$, and $i = 1, \ldots, n$

$$\Pr(K_i = k) \propto \beta_k \times k(x_i|\phi_k). \tag{6}$$

3. Sampling $(\beta|\phi, K, X)$: based on the $\beta_k^* \sim Beta(1 + e_k, \gamma + \sum_{l=k+1}^{N} e_l)$ with e_k corresponding to $\#\{x_i, K_i = k\} \leq n$ for $k = 1, \ldots, N$, then

$$\beta_1 = \beta_1^*, \qquad \beta_k = \beta_k^* \prod_{l=1}^{k-1}(1 - \beta_l^*), \quad \text{for} \quad k = 2, \ldots, N-1, \qquad \beta_N = 1 - \sum_{l=1}^{N-1} \beta_l$$

3.3.1.b Hierarchical model

The algorithm for the stick-breaking representation of the hierarchical Dirichlet process presented in equation (5) requires the sampling of additional intermediate weights $\pi = (\pi_{ik})$ and the definition of a matrix $K = (K_{im})$ describing the assignment of each data point $x_{im} = (x_{pim}, p = 1, \ldots, P)$ to

[1]The code is available on the webpage `http://www.paris.inra.fr/metarisk/members/tressou_jessica/publications`

one of the N atoms, for $i = 1, \ldots, n$ and $m = 1, \ldots, M$. Steps 2 and 3 are replaced with steps 2' and 3' below.

2'. Sampling $(K|\pi, \beta, \phi, X)$ for $k = 1, \ldots, N$, $i = 1, \ldots, n$ and $m = 1, \ldots, M$

$$\Pr(K_{im} = k) \propto \pi_{ik} \times k(x_{im}|\phi_k)$$

3'. Sampling $(\pi|\beta, \phi, K, X)$, independently on the fixed i's and based on

$$\pi^*_{ik} \sim Beta\left(\alpha_0\beta_i + e_{ik}, \quad \alpha_i\left(1 - \sum_{l=1}^{k}\beta_l\right) + \sum_{l=k+1}^{N} e_{il}\right),$$

with e_{ik} corresponding to $\#\{x_{im}, K_{im} = k\} \leq M$ for $i = 1, \ldots, n$ and $k = 1, \ldots, N$, then

$$\pi_{i1} = \pi^*_{i1}, \qquad \pi_{ik} = \pi^*_{ik}\prod_{l=1}^{k-1}(1 - \pi^*_{il}), \quad \text{for} \quad k = 2, \ldots, N-1, \qquad \pi_{iN} = 1 - \sum_{l=1}^{N-1}\pi_{il}.$$

Note that $\pi^*_{iN} \sim Beta(\alpha_i\beta_N + e_{iN}, 0)$ is a properly defined Beta distribution.

Finally sampling $(\beta|\pi, \phi, K, X)$ is performed exactly as described in the original step 3.

3.3.2 Learning about the parameter γ

Considering γ as a random parameter with prior distribution $\Gamma(a_\gamma, b_\gamma)$ leads to an additional last step as proposed in Escobar and West (1995).

4. Sampling $(\gamma|\phi, K, \pi, \beta, X)$ based on an auxiliary variable $\gamma^* \sim Beta(\gamma + 1, n)$, according to the following mixture distribution

$$\gamma \sim w_{\gamma^*} \times \Gamma(a_\gamma + k, b_\gamma - \ln\gamma^*) + (1 - w_{\gamma^*}) \times \Gamma(a_\gamma + k - 1, b_\gamma - \ln\gamma^*), \tag{7}$$

with weights w_{γ^*} defined by $\frac{w_{\gamma^*}}{1-w_{\gamma^*}} = \frac{a_\gamma + k - 1}{b_\gamma - \ln\gamma^*} = c_{\gamma^*}$ that is $w_{\gamma^*} = \frac{c_{\gamma^*}}{1+c_{\gamma^*}}$.

3.3.3 Starting values of hyperparameters

Starting values of the hyperparameters were taken to be equal to $\alpha = P$, $\Psi = 0_{P\times P}$, $m = 0_P$, $t = 1$, as proposed in Cabrera et al. (2009) in order to use vague prior distributions. Different settings were tested on the simulated datasets for the starting value related to the parameter γ: no prior distribution but γ is set to 1, a prior Gamma distribution $\Gamma(a_\gamma, b_\gamma)$ with hyperparameters equal to: $(a_\gamma, b_\gamma) = (2, 4)$ for informative prior and equal to $(a_\gamma, b_\gamma) = (1, 1)$ and $(a_\gamma, b_\gamma) = (0.01, 0.01)$ for more vague prior distributions, yielding the posterior given in (7). In the case of the hierarchical model, the weights α_i were set to 1 for each $i = 1, \ldots, n$.

3.3.4 Algorithm convergence checking

The convergence of the Gibbs sampler to the optimal partition was visually checked by plotting the Q-criterion and the number of clusters $n(\mathbf{p})$ against the number of realized iterations. The number of iterations necessary to reach the optimal partition depends on the size of the dataset. Therefore, the algorithm is stopped after checking that the Q-criterion stabilizes over a minimum of $20,000$ iterations. The optimal partition is determined over all the iterations.

3.3.5 Random-block Gibbs Stick-breaking

Cabrera et al. (2009) introduced a novel procedure called the random-block Gibbs weighted Chinese restaurant process algorithm to reduce the heavy computational burden in estimating the optimal partition induced by the Gibbs sampler and the high dimensionality of the data. We propose to apply this method to the SB algorithm. The principle is to randomly reduce the dimension of the data by selecting a number d ($d < P$) of pesticides from the original number P at each Gibbs cycle. Therefore, given the sequence of random integers $v_d = \{l_1, \ldots, l_d\}$, a subset of observations $x_i = (x_{il_1}, \ldots, x_{il_d})$ is used instead of the $x_i = (x_{i1}, \ldots, x_{iP})$ for the $i = 1, \ldots, n$ individuals. This procedure will be referred to as RB-SB below.

4 Application

4.1 Simulated datasets

To investigate the quality of the clustering estimates under various settings, a simulation study was conducted for the stick-breaking algorithms applied to two datasets based on the settings proposed by Cabrera et al. (2009). The datasets were built from a three component mixture of $P = 5$ dimensional multivariate Normal distributions denoted $N_P(\mu_k, \Sigma_k)_{k=1,\ldots,3}$. The parameters (μ_k, Σ_k) are detailed in Table 1.

4.1.1 Dataset for the base model

A sample of $1,000$ values (x_i) was built from $f(x) = \sum_{k=1}^{3} \beta_k N_P(\mu_k, \Sigma_k)$ with $(\beta_1, \beta_2, \beta_3) = (0.3, 0.3, 0.4)$. After $30,000$ iterations, and with the parameter γ set to 1, the optimal partition was obtained at the 34^{th} iteration with the Q-criterion of equation (4) being equal to $-3,687$, see Fig. 1(a). The optimal partition was composed of 3 clusters corresponding to the 3 components of the mixture dataset, see Fig. 1(b). Performing the RB-SB algorithm with the dimension reduced to $d = 2$, the maximum Q-criterion is reached at the 650^{th} iteration, also with an optimal partition matching the generating one (see Figure 1(b)). When using different prior distributions for the parameter

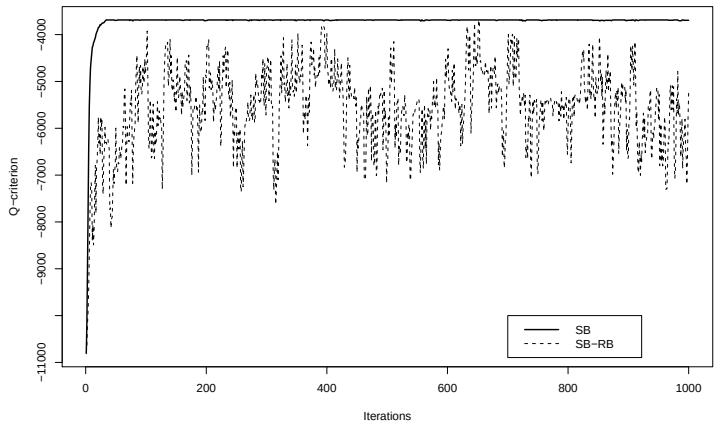
Table 1: Parameters of the multivariate Normal distributions $N_P(\mu_k, |Sigma_k)_{k=1,...,3}$ used to generate the simulated datasets

k	μ_k	Σ_k				
	2	1.0	0.5	0.2	0.1	0.1
	2	0.5	2.0	0.5	0.2	0.1
1	4	0.2	0.5	1.0	0.5	0.2
	5	0.1	0.2	0.5	1.0	0.5
	6	0.1	0.1	0.2	0.5	3.0
	-2	1.0	0.5	0.2	0.1	0.1
	-2	0.5	2.0	0.5	0.2	0.1
2	-4	0.2	0.5	3.0	0.5	0.2
	-5	0.1	0.2	0.5	2.0	0.5
	-6	0.1	0.1	0.2	0.5	1.0
	-5	3.0	0.5	0.2	0.1	0.1
	5	0.5	1.0	0.5	0.2	0.1
3	-7	0.2	0.5	2.0	0.5	0.2
	7	0.1	0.2	0.5	2.0	0.5
	-9	0.1	0.1	0.2	0.5	3.0

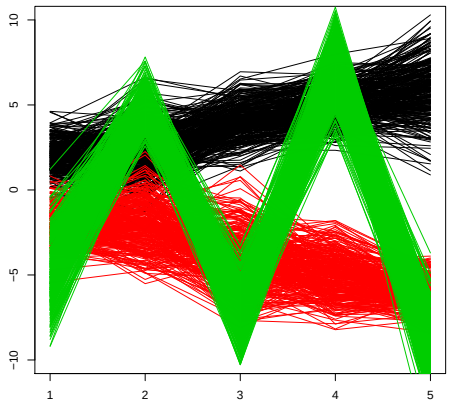
γ, the same maximum value of $Q = -3,687$ is reached, as with γ being set to 1. Figure 2 shows the 3 different prior distributions attributed to the parameter γ and their corresponding posterior distributions. 50% of posterior values of γ are below $0.27 < 1$, showing that it is possible, though perhaps not crucial here, to learn about γ.

4.1.2 Dataset for the hierarchical model

To reproduce the co-exposure data structure including uncertainty of exposure, 240 individuals were generated from the $f(x) = \sum_{k=1}^{3} \beta_k N_P(\mu_k, \Sigma_k)$ with $(\beta_1, \beta_2, \beta_3) = (0.33, 0.17, 0.5)$. For each individual i, $M = 100$ values were sampled resulting in a total sample of $24,000$ observations. The convergence of the SB and the RB-SB algorithms to the optimal partition was very slow. After $200,000$ iterations, the number of clusters which maximized the Q-criterion was 11. The size of the clusters ranged from 269 to $16,089$ observations. The biggest clusters had means and covariances similar to those used to generate the dataset. This suggested that if the number of iterations was increased, three main clusters would become apparent and the optimal partition would eventually converge to that generating the data.



(a) Q-criterion of the first $1,000$ iterations performed with the SB and the SB-RB algorithms



(b) Parallel coordinates of the optimal partition obtained with the SB or the RB-SB algorithm

Figure 1: The Stick-Breaking (SB) and the Random-Block Stick-Breaking (RB-SB) algorithms for the base model applied to a simulated dataset ($N = 30$ atoms and $30,000$ iterations).
Note: To read parallel coordinates consider that the x-axis is the dimension j, the y-axis is the value of x_j, a line connects x_1 to x_2, x_2 to x_3 and so on. The color indicates the 3 clusters resulting from the SB or RB-SB algorithms, which both match the generating ones described in Table 1

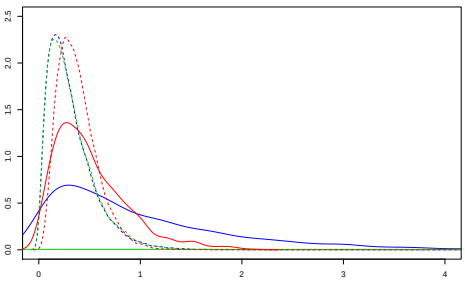


Figure 2: Densities of the $Gamma(a_\gamma, b_\gamma)$ prior (solid line) and the posterior (dashed line) distributions of the parameter γ. Red line: $(a_\gamma, b_\gamma) = (2, 4)$, blue line: $(a_\gamma, b_\gamma) = (1, 1)$, green line: $(a_\gamma, b_\gamma) = (0.01, 0.01)$, prior densities are plotted in logarithmic scale.

4.2 Real datasets on co-exposure to pesticides

4.2.1 Base model for the 95^{th} percentile of co-exposure

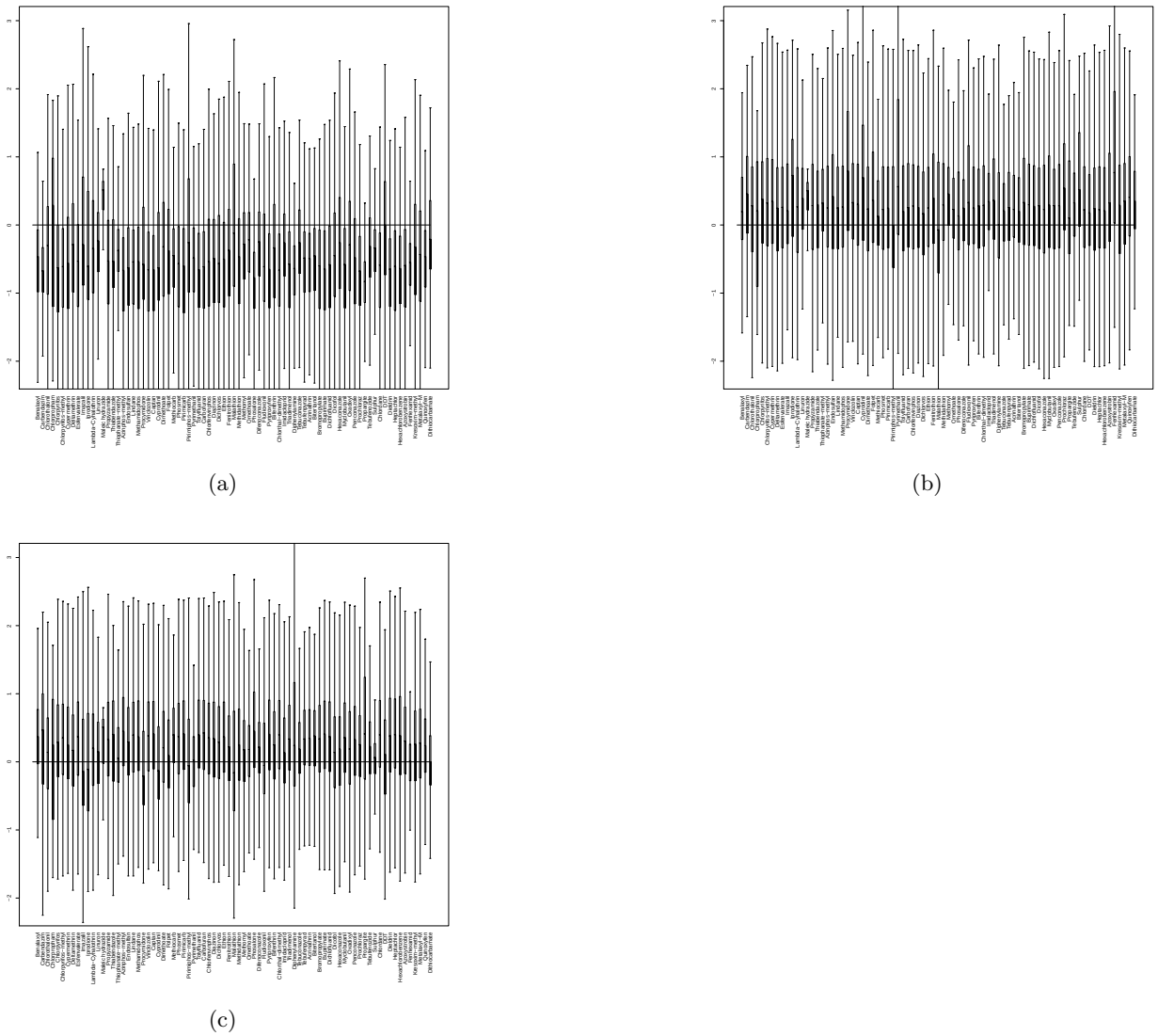
At first, a non hierarchical version of the model is considered as shown in equation (2): x_{ip} is the 95^{th} percentile of exposure to pesticide p for individual i. Different values of d ($d = \{15, 25, 41, 79\}$) were tested running $100,000$ iterations of the algorithm and using a Gamma(0.01, 0.01) distribution for γ as a vague prior. The results shown are for the value of d which maximizes the Q-criterion: $d = 41$. To test the convergence of the algorithm to the optimal partition an extra $200,000$ iterations were performed with this value of d.

For the adults sample, the optimal partition was obtained after $196,445$ iterations and is composed of 17 clusters. The adult population was clustered into 3 main sub-populations composed of 582, 412 and 870 individuals. The other 14 clusters were discarded as they jointly only accounted for 34 individuals. For each main cluster, the box plots of the 79 pesticide exposures are shown in Fig. 3. The sub-populations of the clusters 2 and 3 are the most highly exposed to a large number of pesticides. For these 2 populations, the correlation matrices of the pesticide exposure are drawn from the posterior distribution of the parameter τ and shown in cf. Fig. 4(a) and Fig. 4(c). To define the cocktails, we focused on pesticides with at least one correlation greater than 0.95 (see Fig. 4(b) and Fig. 4(d)). With this criterion, from the 79 pesticides and the two sub-populations, 34 pesticides have been selected and combined into 20 cocktails.

For the children sample, the optimal partition was obtained after $98,362$ iterations and is composed of 16 clusters. As with the adults, 14 clusters accounting for only 45 individuals were discarded to focus on the 2 main ones. The first cluster is composed of 743 children who are highly exposed to a large number of pesticides, and the second one consists of 651 children. As with the adults sample, the correlation matrix of the most heavily exposed sub-population was analyzed to determine cocktails of pesticides. There are 39 pesticides with correlations over 0.95, divided

into 13 different cocktails. Of these 39 pesticides, 28 are similar to those obtained with the adult population.

The analyses of the pesticide exposures of the main clusters show that the individuals are either highly exposed to all pesticides or have a low exposure to all pesticides. A preliminary conclusion is that this clustering is mostly related to the consumption behavior of individuals rather than the contamination levels.



(a) (b)

(c)

Figure 3: Box plots of the 79 pesticide exposures for the cluster 1 (a), cluster 2 (b) and cluster 3 (c).

Comparison with PCA

To compare the results obtained with the base model to those from a classical method, a principal component analysis (PCA) for the adult population was realized and described in this section. The

axis 1, which represents 68% of the total variance is mainly determined by the 34 pesticides selected with our base model. Indeed, among the 24 pesticides which mainly contribute to axis 1, 23 were also selected with the base model. The second axis only represents 6% of the total variance. The coordinates of the 79 pesticides are positive on axis 1, while the individuals are found on both sides of the axis. The results obtained with the PCA lead to the same conclusion that there are two groups of individuals, one highly exposed and one only marginally exposed to all the pesticides.

4.2.2 Hierarchical model for distributions of individual co-exposures

The hierarchical model was developped to deal with distributions of individual co-exposure to the P pesticides in order to account for uncertainty of residue levels. Because of the algorithm's long computational time, the hierarchical model was only applied to the 34 pesticides selected with the base model for the adult population. From $30,000$ iterations, the optimal partition was found after $25,853$ iterations and was composed of 6 clusters. From these clusters, three were composed of large sets of observations according to the clustering obtained with the base model (see previous subsection). Moreover, two of these clusters were composed of highly exposed individuals to a large proportion of the 34 pesticides. The distribution of co-exposure of each individual was mostly found to consist of 2 or 3 component mixtures. The analysis of the correlation matrix of the two main clusters showed that the correlations between pesticides were very low, ranging between 0.2 and 0.45, compared to those obtained with the base model. These low correlations could be due to the high uncertainty associated with the exposure to each pesticide. Indeed, for a random set of contamination selected for the P pesticides, an individual can be exposed to a low level for one pesticide and a high level for another, leading to low correlations. The inclusion of the exposure uncertainty is more realistic in terms of exposure assessment but implies some difficulties in the definition of the cocktails of pesticides as we can place ourselves at the different hierarchy levels to define the cocktails, which results in too many different cocktails.

5 Conclusion

This paper presents a Bayesian nonparametric model based on Dirichlet process mixtures, applied to cluster the co-exposures of the French population to various pesticides in order to define cocktails of pesticides which are relevant to study in terms of human health effects. The hierarchical model applied in this work is original in the field of food risk analysis although it has been applied in other fields, mainly to deal with functional data (Rodríguez et al., 2009; Teh et al., 2006). Compared to the PCA or other clustering methods, this type of Bayesian nonparametric model has several advantages, for example it is not based on the linearity assumption. Also, it can comprise an infinite number of mixture components which is particularly suited for the case of high data dimensionality. Moreover, the number of clusters is automatically determined through the estimation process, no

parametric assumption with regard to the form of the co-exposure distribution is required and the structure of the data set may be introduced through a specific hierarchy to account for exposure uncertainty. As a result, the model applied to the 95^{th} percentile of the French population co-exposure to pesticides clustered the adult population into two sub-populations: individuals highly exposed to a large proportion of pesticides and individuals slightly exposed to a large proportion of pesticides. Thus, cocktails defined from individuals highly exposed are numerous and composed of various pesticides, for example for the adults sample 34 pesticides shared into 20 cocktails have been selected. Nevertheless, this approach relies on the major assumption that the non-detect values are uniformly distributed between 0 and the LOR. This assumption may not be realistic for pesticides residues, for which a real 0 can occur when the pesticide is not used on a crop, (EFSA, 2010). Then, refinement on the way to deal with non-detects is necessary in order to define more realistic cocktails.

Acknowledgment

This work is a part of a project funded by the French National Research Agency (ANR) and the French Agency for Environmental and Occupational Health Safety (AFSSET). The authors thank Fanny Héraud for her relevant comments.

References

AFSSA (2009). "INCA 2 (2006-2007). Report of the Individual and the National Study on Food Consumption."

Antoniak, C. E. (1974). "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems." *Annals of Statistics*, 2: 1152–1174.

Blackwell, D. and MacQueen, J. B. (1973). "Ferguson distributions via Polya urn schemes." *Annals of Statistics*, 1: 353–355.

Cabrera, J., Lau, J. W., and Lo, A. Y. (2009). "Random Block Sampling for high dimensional clustering (from the Bayesian point of view)." *Working Paper*. Available from the second author upon request.

EFSA (2010). "Management of left-censored data in dietary exposure assessment of chemical substances." Technical Report 3. `doi:10.2903/j.efsa.2010.1557`. Available online: `www.efsa.europa.eu`.

Escobar, M. D. and West, M. (1995). "Bayesian density estimation and inference using mixtures." *Journal of the American Statistical Association*, 90: 577–588.

Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems." *Annals of Statistics*, 1: 209–230.

Ishwaran, H. and James, L. F. (2001). "Gibbs Sampling Methods for Stick-Breaking Priors." *Journal of the American Statistical Association*, 96: 161–173.

Lau, J. W. and Lo, A. Y. (2007). "Model based clustering and weighted Chinese restaurant processes." In *Advances in Statistical Modeling and Inference: Essays in Honor of Kjell A. Doksum*, 405–424.

Lo, A. Y. (1984). "On a class of Bayesian nonparametric estimates: I. Density Estimates." *Annals of Statistics*, 12(1): 351–357.

Muller, P. and Quintana, F. (2004). "Nonparametric Bayesian data analysis." *Statistical Science*, 19(1): 95–110.

Pitman, J. and Yor, M. (1996). "Some developments of the Blackwell-MacQueen Urn scheme." *IMS, Hayward, California*.

Quintana, F. and Iglesias, P. (2003). "Bayesian Clustering and Product Partition Models." *Journal of the Royal Statistical Society Series B*, 65(2): 557–574.

Rodríguez, A., Gelfand, A. E., and Dunson, D. B. (2009). "Bayesian nonparametric functional data analysis through density estimation." *Biometrika*, 96(1): 149–162.

Sethuraman, J. (1994). "A constructive definition of Dirichlet prior." *Statistica Sinica*, 4: 639–650.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). "Hierarchical Dirichlet Processes." *Journal of the American Statistical Association*, 101(476): 1566–1581.

Walker, S. (2007). "Sampling the Dirichlet mixture model with slices." *Communications in Statistics*, 36: 45–54.

Walker, S., Damien, P., Laud, P., and Smith, A. (1999). "Bayesian nonparametric inference for random distributions and related functions." *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 61(3): 485–527.

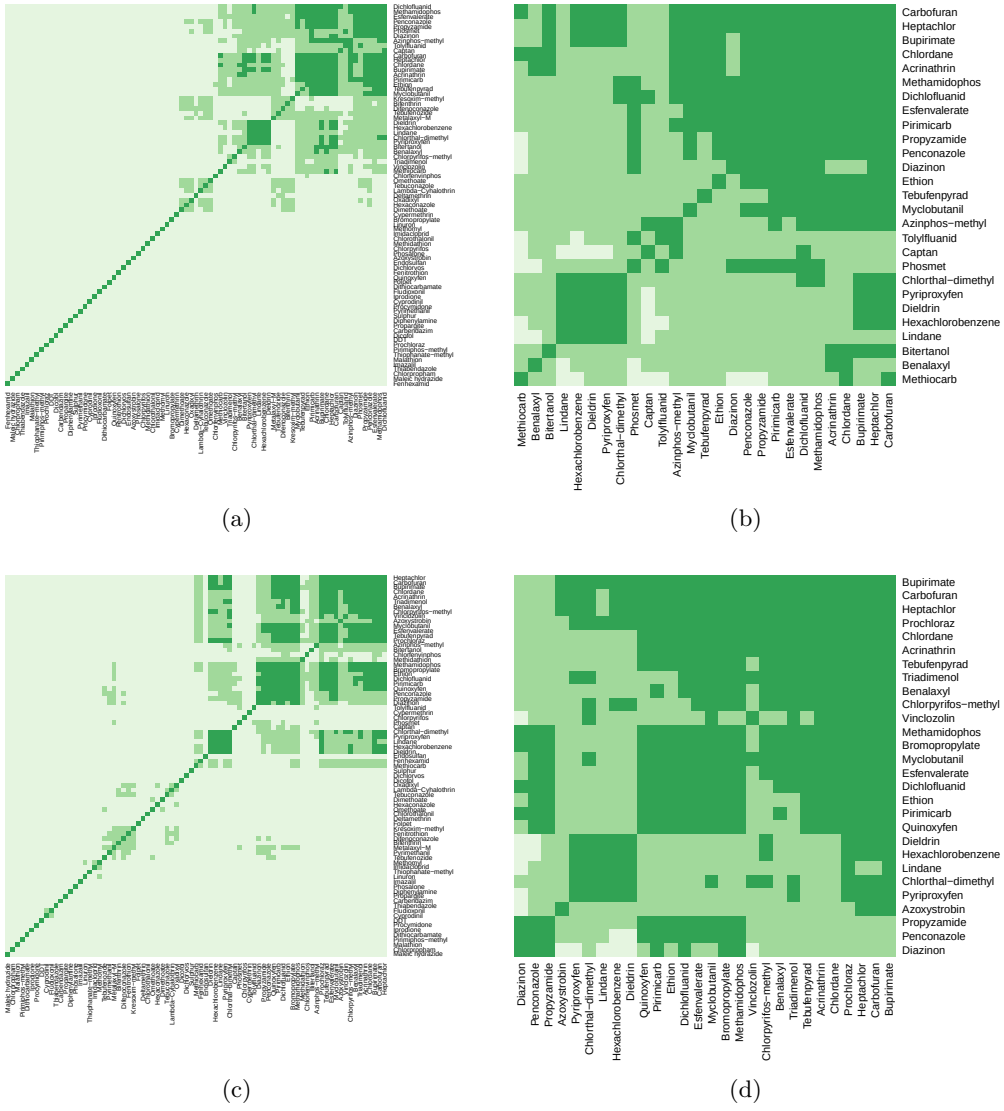


Figure 4: Heat map of the correlation matrix of the cluster 2 (a,b) and the cluster 3 (c,d). Right maps focus on pesticides with correlations over 0.95