Identifying pesticide mixtures at country-wide scale

Milena CAIRO[1], Anne-Christine MONNET[1], Stéphane ROBIN[1,2], Emmanuelle PORCHER[1], Colin FONTAINE[1]

[1] Centre d'Écologie et des Sciences de la Conservation (CESCO), Muséum national d'Histoire naturelle, Centre National de la Recherche Scientifique, Sorbonne Université, CP 135, 57 rue Cuvier 75005 Paris, France

[2] Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, F-75005 Paris, France

Corresponding author: Milena Cairo, milena.cairo1@mnhn.fr, Centre d'Écologie et des Sciences de la Conservation (CESCO), Muséum national d'Histoire naturelle, CP 135, 57 rue Cuvier 75005 Paris, France

ABSTRACT

Wild organisms are likely exposed to complex mixtures of pesticides owing to the large diversity of substances on the market and the broad range agricultural practices. The consequences of such exposure are still poorly understood, first because of potentially strong synergistic effects, making cocktails effects not predictable from the effects of single compounds, but also because little is known about the actual exposure of organisms to pesticide mixtures *in natura*.

We aimed to identify the number and composition of pesticide mixtures potentially occurring in French farmland, using a database of pesticide purchases in postcodes. We developed a statistical method based on a model-based clustering (mixture model) to cluster postcodes according to the identity, purchase probability and quantity of 279 active substances.

We found that the 5,642 French postcodes can be clustered into a small number of postcode groups (ca. 20), characterized by a specific pattern of pesticide purchases, i.e. pesticide mixtures. Substances defining mixtures can be sorted into “core” substances highly probable in most postcode groups and “discriminating” substances, which are specific to and highly probable in some postcode groups only, thus playing a key role in the identity of pesticide mixtures. We found 12 core substances: two insecticides (deltamethrin and lambda-cyhalothrin), six herbicides (glyphosate, diflufenican, fluroxypyr, MCPA, 2,4-d, triclopyr) and four fungicides (fludioxonil, tebuconazole, difenoconazole, thiram). The number of discriminating substances per postcode group ranged from 2 to 74. These differences in substance purchases seemed related to differences in crop composition but also potentially to regional effects.

Overall, our analyses return (1) sets of molecules that are likely to be part of the same pesticide mixtures, for which synergetic effects should be investigated further and (2) areas within which biodiversity might be exposed to similar mixture composition. This information will hopefully be of interest for future ecotoxicological studies to characterise the actual impacts of pesticide cocktails on biodiversity in the field.

Keywords: Active substances, Cluster, mixture model, expectation-maximization algorithm, risk assessment

INTRODUCTION

Since the mid-20th century, pesticides have become of common use in agriculture and their effects on both the environment and human health are a growing concern. For example, systemic pesticides are known to affect a broad range of organisms, from invertebrates, both terrestrial and aquatic, to amphibians or birds (Humann-Guilleminot et al., 2019; Mahmood et al., 2016; Yang et al., 2008), thereby questioning the sustainability of agroecosystem functioning and related services (Deguines et al., 2014; Dudley et al., 2017; Furlan et al., 2018; Geiger et al., 2010). Pesticides are also identified as a concern for human health, with numerous pesticide poisonings reported across developing countries (Boedeker et al., 2020) and recent evidence of relationships between diseases such as Parkinson’s or cancers and exposure to organophosphate insecticides (Sheahan et al., 2017; Tassin de Montaigu and Goulson, 2020).

The effect of pesticides on biodiversity are usually demonstrated with a focus on a single substance or a limited set of substances in general (e.g. thiamethoxam, clothianidin, imidacloprid, thiacloprid or glyphosate (Botías et al., 2015; Busse et al., 2001; Rundlöf et al., 2015; Van Bruggen et al., 2018). Yet, wild organisms are exposed to complex mixtures (Dudley et al., 2017), owing to the diversity of substances available and used in farmlands. Hence, studying substance mixtures is considered a central task for environmental risk assessment (Lydy et al., 2004a), notably because the effects of pesticide cocktails can strongly exceed the additive effects of single compounds (Bopp et al., 2016; Junghans et al., 2006). Laboratory experiments demonstrate synergetic interactions among substances within mixtures, affecting the effect of the cocktails in non-additive ways (Cedergreen, 2014; Hernández et al., 2017; Heys et al., 2016). While the importance of studying the effects of cocktails beyond those of single substances was highlighted as soon as the late sixties (Keplinger and Deichmann, 1967), and their evaluation is mandatory in the European Union since 2009 (EC No 1107/2009), few attempts to do so exist outside laboratories (Gibbons et al., 2015).

Studies examining the effects of substance cocktails use two approaches: bottom-up or top-down (Altenburger et al., 2013; Hernández et al., 2017; Relyea, 2009). The bottom-up approach aims at testing all possible mixture compositions, starting from pairs of substances to more complex combinations. This method makes it challenging to consider more than a handful of substances. For example, ten substances represent 45 possible pairs and over a thousand possible combinations of three or more substances (Lydy et al., 2004a). Moreover, such approach might be more suited to experiments in controlled rather than natural environments, as the latter are recognized as strongly contaminated (Tang et al., 2021), making the control of mixture composition difficult. The top-down approach proposes to compare the effect of cocktails, starting from potentially frequent mixtures including a high number of substances, but at the cost of not testing all combinations. In addition, the few existing field studies generally focused on the effects of pesticide cocktails composed of a restricted number of substances, on specific crops or on restricted spatial extent, thereby limiting a broad understanding of cocktail effects (e.g. Brittain et al., 2010; Hallmann et al., 2014; Millot et al., 2017, but see Schreiner et al., 2016 & (Fritsch et al., 2022). The top-down approach makes it critical to identify relevant mixture compositions, i.e. those actually occurring in the fields. The number of actual mixtures encountered in agroecosystems should be much lower than the number of possible combinations of substances because each substance is often intended for a limited set of crops only and because agricultural production is regionally specialised on particular crops. Such regional specialisation implies that existing mixtures are likely to be spatially structured. However, we still miss an overall picture of the pesticide mixture composition and its spatial structure over large spatial extents.

Here, we introduce a new statistical method to identify relevant pesticide mixtures, i.e. actual combinations of substances potentially co-occurring in agroecosystems, across Metropolitan France. We overcame the general problem of limited availability of data on temporal and spatial use of pesticides (Navarro et al., 2021) by taking advantage of the recent publication of an up-to-date database on pesticide purchases in France, the French national bank of pesticide sales database (https://www.data.gouv.fr/fr/datasets/ventes-de-pesticides-par-departement/). This database has registered mandatory reporting of quantities of active substances purchased in France since 2013 (law n°2006-1772) at a relatively fine spatial grain (postcode of the buyer). France is also the seventh largest user of pesticides in the world (FAO 2020) and has a wide range of agricultural types (Urruty et al., 2016), which makes it a well-suited case country to identify pesticide mixtures encountered in the field by wild organisms, as well as their spatial variation.

Applying an Expectation/Maximization algorithm to a model-based clustering, we aimed to cluster French postcodes on the basis of their composition of active substances purchased. We addressed three main questions: 1) How many groups of postcodes best describe the patterns of pesticide purchase in France? 2) How are these groups spatially distributed? 3) What are the mixtures of active substances characterizing these groups? Because pesticide use is at least partially related to crop identity, and because of crop regional specialization in France, we expect a limited number of postcode groups, that are strongly structured in space. Such groups with homogeneous pesticide mixtures could subsequently be used to identify potentially important pesticide substances and mixtures deserving further investigation.

METHODS

1.1 Pesticide data

Data on active substances were obtained from the French national bank of pesticide sales (BNV-d; https://bnvd.ineris.fr). The BNV-d database registers active substances under mandatory reporting. The seller indicates the amount of each active substance purchased and the postcode of the buyer in the database. This database thus indicates the quantity of active substances purchased at the spatial resolution of the postcode of the buyer. Postcode are the third level of administrative division in France, lower than the European Union NUTS3 level (administrative departments) and range from 0.17 km² to 614.39 km² in metropolitan France (median = 62.79 km², Q1 = 19.59 km², Q3 =140.36 km²). Substances are identified with their generic name and a unique identifier, the Chemical Abstracts Service number. We modified generic names when synonyms were found. We only retained substances with a license fee (i.e. under compulsory reporting) because we can expect thorough reporting for these.

The years registered in the database ranged from 2013 to 2020. We discarded the year 2013 because of incomplete data during the first reporting year, and the two latest years of the time series (2019 and 2020) because additions and changes in the database are allowed for two years after reporting. Also, note that the legislation has kept changing until 2016, with consequences for the mandatory nature of reporting for some substances or treatments. In particular, until 2016 the geographical information associated with seed coating substances was that of the seed coating company, not of the buyer. Hence, 2017 can be considered the most accurate and thorough year within the period 2013-2020.

The data provides the total mass (in g) bought per substance with mandatory reporting, of which in 2017 there were 279. We analysed these quantitative data at the postcode level, assuming that substances purchased in a given postcode would be used within the same postcode or in close vicinity. Given the spatial extent of farms, pesticides may not always be spread exactly in the postcode where farmers are domiciled, but are unlikely to be used beyond the neighbouring postcodes, with one exception that we discarded. Using specific postcodes (CEDEX) that enable the identification of private companies, we discarded the data related to the national railroad company (SNCF): SNCF is a major buyer with central purchasing bodies that do not use the substances within the postcode of purchase. We converted all remaining CEDEX codes to their corresponding regular postcodes. We were thus left with 5,642 postcodes with information about the quantities (in g) of 279 active substances purchased in 2017. We classified these substances into fungicides, herbicides, insecticides following the Pesticide Properties Data Base (PPDB) (Lewis et al., 2016) and the European commission pesticide database (ec.europa.eu/food/plant/pesticides/eu-pesticides-database/active-substances). There were also 32 substances with other target groups (e.g. rodents or molluscs; Table S1 for a complete list) that we classified as "other targets".

To relate the use of active substances to the area of arable land in postcodes, we extracted the total area of cropland from the 2017 French Land Parcel Identification System (LPIS, "Registre Parcellaire Graphique", Agence de Services et de Paiements, 2015). This database is a geographic information system developed under the European Council Regulation No 153/2000, for which the farmers provide annual information about their fields and crop rotation. We grouped the 16 categories of cropland types used in LPIS into 11 sub-groups (Figure S9) (Cantelaube and Carles, 2010; Levavasseur et al., 2016). We summed the area of all types of cropland but meadows to obtain the total crop area per postcode.

1.2 Model-based Clustering

1.2.1 Input data

As described above, the dataset consisted of n (= 5,642) postcodes and p (=279) substances. For each postcode i ($1 \leq i \leq n$) and substance j ($1 \leq j \leq p$), we denoted by X_{ij} the presence/absence variable, which is 1 if substance j is bought in postcode i and 0 otherwise, and by Y_{ij} the log of the quantity of substance j bought in postcode i (when used) normalized with the cropland area of postcode i:

$$Y_{ij} = \log\left(\frac{\text{quantity of substance } j \text{ bought in postcode } i}{\text{cropland area of postcode } i}\right)$$

(Y_{ij} is NA when substance j is not bought in postcode i).

1.2.2 Model

We aimed to provide a clustering of the postcodes according to the quantity of the various substances bought. Mixture models (McLahan and Peel, 2000) provide a classical framework to achieve such a clustering. To avoid any confusion with "pesticide mixtures" we will use "Model-based Clustering" when referring to the statistical "mixture models". The model we consider assumes that the n postcodes are spread into K groups and that the respective use of the different substances depends on the group they belong to. Mixture models or model-based clustering precisely aim at recovering this unobserved group structure from the observed data.

1.2.2.1.1 Groups definition

We denoted by Z_i the group to which postcode i belongs. We assumed the Z_i are all independent and that each postcode i belongs to group k ($1 \leq k \leq K$) with respective proportions π_k:

$$\pi_k = \Pr\{Z_i = k\}. \qquad (1)$$

Note that the π_k consists of only $K - 1$ independent parameters, as they have to sum to 1 ($\sum_{k=1}^{K} \pi_k = 1$).

1.2.2.1.2 Emission distribution

The model then describes the distribution of the observed data conditional on the group to which each postcode belongs. The distribution of the presence/quantity pair (X_{ij}, Y_{ij}) is built in two stages: first, if postcode *i* belongs to group k, substance j is used in the postcode with probability γ_{kj}:

$$\gamma_{kj} = \Pr\{X_{ij} = 1 | Z_i = k\}, \quad (2)$$

then, if substance j is used in postcode i, its log-quantity is assumed to have a Gaussian distribution:

$$(Y_{ij} | X_{ij} = 1, Z_i = k) \sim \mathcal{N}(\mu_{kj}, \sigma_{kj}^2). \qquad (3)$$

with μ_{kj} and σ_{kj}^2 the mean and variance of the log-quantity of substance j used in a postcode from group k, provided that the substance is bought in the postcode. In addition to the $(K-1)$ proportions π_k and the $K \times p$ probabilities γ_{jk}, this model involves $K \times p$ mean parameters μ_{kj} and as many variance parameters σ_{kj}^2. This makes a total of $K - 1 + 3Kp$ parameters to be estimated.

Combining Equations (2) and (3), we defined the conditional distribution f_{jk} for substance j in a postcode from group k:

$$f_{jk}(x_{ij}, y_{ij}) = x_{ij}\gamma_{kj}\phi(y_{ij}; \mu_{kj}, \sigma_{kj}^2) + (1 - x_{ij})(1 - \gamma_{kj})$$

denoting by $\phi(\cdot; \mu, \sigma^2)$ the probability density function of the Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$.

To avoid over-parametrization, we also considered models with constrained variance, assuming either that the variance depends on the substance but not on the group: $\sigma_{kj}^2 \equiv \sigma_j^2$, or that the variance is the same for all substances in all groups: $\sigma_{kj}^2 \equiv \sigma^2$.

1.2.3 Inference

Model-based clustering belongs to incomplete-data models, which can deal with situations where part of the relevant information is missing. For the sake of brevity, we denoted by Y the set of observed variables (i.e. all the (X_{ij}, Y_{ij})) and by Z the set of unobserved variables (i.e. the Z_i). We further denoted by θ the whole set of parameters to be estimated: $\theta = (\{\pi_k\}, \{\gamma_{kj}\}, \{\mu_{kj}\}, \{\sigma_{kj}^2\})$.

A classical way to estimate the set of parameters θ is to maximize the log-likelihood of the data $\log p(Y; \theta)$ with respect to the parameters. An important feature of incomplete-data models is that this log-likelihood is not easy to compute, and even harder to maximize, as its calculation requires integrating over the unobserved variable Z. However, the so-called 'complete' log-likelihood, which involves both the observed Y and the unobserved Z, $\log p(Y, Z; \theta)$ is often tractable.

1.2.3.1.1 Expectation-Maximization algorithm

The Expectation-maximization (EM) algorithm (Dempster et al., 1977) resorts to the complete log-likelihood to achieve maximum-likelihood inference for the parameters. More specifically, because $\log p(Y,Z;\theta)$ cannot be evaluated (as Z is not observed), EM uses the conditional expectation of the complete likelihood given the observed data, namely $\mathbb{E}[\log p(Y,Z;\theta)|Y;\theta]$, as an objective function, to be maximized with respect to θ.

The EM algorithm alternates the steps 'E' (for expectation) and 'M' (for maximization) until convergence. It can be shown that the likelihood of the data $\log p(Y;\theta)$ increases after each EM step. The reader may refer to Dempster et al. (1977) or McLahan and Peel (2000) for a formal justification of the procedure.

1.2.3.1.2 E step

This step aimed at recovering the relevant information to evaluate the objective function. In the case of model-based clustering, the E steps only amounts to evaluating the conditional probability τ_{ik} for the postcode i to belong to group k given the data observed for the postcode and the estimate of the parameter θ_{ik} after iteration $h-1$:

$$\tau_{ik}^{(h-1)} = \Pr\{Z_i = k|\{(X_{ij},Y_{ij})\}_{1\le j\le p};\theta^{(h-1)}\}$$

The calculation of τ_{ik} simply resorts to Bayes formula. In the following, we drop the iteration superscript (h) for the sake of clarity, and we use the notation $\hat{\theta}$ to indicate the current estimate. Because the substance are assumed to be independent, we get

$$\hat{\tau}_{ik} = \hat{\pi}_k \prod_{j=1}^{p} \hat{f}_{jk}(x_{ij},y_{ij}) / \left(\sum_{\ell=1}^{K} \widehat{\pi_\ell} \prod_{j=1}^{p} \hat{f}_{j\ell}(x_{ij},y_{ij})\right).$$

1.2.3.1.3 M step

The M step updates the parameter estimate by maximizing $\mathbb{E}\left[\log p(Y,Z;\theta)|Y;\theta^{(h-1)}\right]$ with respect to θ. The objective function can be calculated using the conditional probabilities τ_{ik}s

$$\mathbb{E}\left[\log p(Y,Z;\theta)|Y;\theta^{(h)}\right] = \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{\tau}_{ik}\left(\log \pi_k + \sum_{j=1}^{p} \log f_{kj}(x_{ij},y_{ij})\right).$$

The maximization of this function yields in close-form update formulas for all parameters. All estimates can be viewed as weighted versions of intuitive proportions, means or variance. Let us first define

$$\widehat{N}_k = \sum_{i=1}^{n} \hat{\tau}_{ik}, \widehat{M}_{kj} = \sum_{i=1}^{n} \hat{\tau}_{ik} x_{ij}.$$

$\widehat{N}_k$ is the current estimate of the number of entities belonging to group k; $\widehat{M}_{kj}$ is the current estimate of the number of entities from group k where substance j is bought. For the proportions and probability of use, we get the following updates:

$$\hat{\pi}_k = \widehat{N}_k/n \,, \hat{\gamma}_{kj} = \widehat{M}_{kj}/\widehat{N}_k.$$

For the quantitative part of the model, we get additionally:

$$\hat{\mu}_{kj} = \frac{1}{\widehat{M}_{jk}}\sum_{i=1}^{n} \hat{\tau}_{ik}x_{ij}y_{ij} \quad \hat{\sigma}^2_{kj} = \left(\frac{1}{\widehat{M}_{jk}}\sum_{i=1}^{n} \hat{\tau}_{ik}x_{ij}y_{ij}^2\right) - (\hat{\mu}_k)^2 .$$

Similar estimates of σ_j^2 and σ^2 can be derived for the models with constrained variances.

1.2.4 Model selection

To select the number of groups K and to choose between the models with unconstrained and constrained variances, we used the Bayesian Information Criterion (BIC, Schwarz, 1978). We adopted the same form as in Fraley and Raftery [1999], that is:

$$BIC = \log p(Y; \hat{\theta}) - \frac{n}{2}\log(\#\text{independent parameters}).$$

As indicated above, the number of independent parameters is:

- $K - 1 + 3Kp$ with unconstrained variances σ_{jk}^2,
- $K - 1 + 2Kp + p$ with constant variance for each substance $\sigma_{jk}^2 \equiv \sigma_j^2$,
- $K + 2Kp$ with constant variance $\sigma_{jk}^2 \equiv \sigma^2$.

1.2.5 Estimated parameters

The output of the model-based clustering yielded K groups with their corresponding estimated parameters, that is $\hat{\tau}_{ik}, \hat{\gamma}_{kj}, \hat{\mu}_{kj}, \hat{\sigma}^2_{kj}$, with k one of the K groups obtained, j an active substance and i a postcode. These estimated parameters gave information on groups of postcodes and substances bought per group.

$\hat{\tau}_{ik}$ was the conditional probability that a postcode i belongs to each group k given the quantities of substances bought in the postcode. We used this probability to associate each postcode to its most probable group.

$\hat{\gamma}_{kj}$ was the probability of a substance j to be used in a postcode of group k. We used this probability to study the composition of active substances in each group k.

$\hat{\mu}_{kj}$ and $\hat{\sigma}^2_{kj}$ were the estimated mean and variance of the log-quantity of substance j per square meter of cropland purchased in a postcode from group k. These quantities were used to refine our understanding of the subtance composition of postcode groups.

1.3 Analyses on estimated parameters

1.3.1 Spatial structure of postcode groups

To characterise the spatial structure of postcode groups, we quantified the spatial spread of postcodes belonging to a same group via the area of the convex hull of the group. The convex hull of a group is the smallest convex set that contains all postcodes of the group. Regardless of their spatial aggregation, most groups contain a few scattered postcodes, such that the convex area of all groups generally contains most of France, making comparisons of the area irrelevant. To circumvent this difficulty, we merged all contiguous postcodes within a group into single polygons and retained only the largest polygons, representing 80% of the total area of a group. This eliminated the scattered postcodes outside the main core of postcodes within a group.

We also characterized the similarity among the K groups in terms of substance use via hierarchical clustering on distances between groups. To obtain a matrix of between-group distances, we used results from the model-based clustering and calculated a maximum-likelihood inference when two randomly chosen groups were merged (see method in 1.2). We repeated this step for each possible group pair. We thus obtained a matrix of between-group distances, characterized as differences in likelihood between clusterings. Using this matrix, we computed an agglomerative nesting clustering, using Ward criterion, implemented in the R package *cluster* (Maechler et al.,2019, R Core Team 2021).

1.3.2 Searching for the drivers of the substance composition of groups

We tried to identify some of the possible drivers of the substance composition of groups using two complementary approaches. First, we tested whether the groups obtained with the model-based clustering, which by construction differ in terms of active substances purchased, also differed in terms of crop composition. To compare the proportion of area covered with different crops among groups, we performed a log-ratio analysis (LRA). This approach was implemented in the R package *easyCODA* (Greenacre, 2019, R Core Team 2021). Second, we used Mantel tests (Mantel & Valand 1970) to estimate the correlations between three distance matrices among postcode groups: distances in the composition of substances purchased in the group (see above), distances in crop composition, and geographic distances. We used a spearman method and used 9999 permutations, computed with the *vegan* package (Oksanen and Simpson, 2022)

1.3.3 Test of the temporal robustness of the model-based clustering

To test robustness of the results of the model-based clustering run on the pesticide purchase data from the year 2017 vs. a longer time period, we also run the clustering on BNV-d data over the period 2015 to 2018. To do so, we aggregated all purchase data from 2015 to 2018 and analysed these data in the same way as those from 2017. In the following, the groups obtained with the model-based clustering applied on the 2017 data (respectively 2015-2018 data) are referred to as the "2017 groups" (respectively the "2015-2018 groups").

We used postcode probabilities to be in group k (i.e. $\hat{\tau}_{ik}$) to compare results from the two model-based clusterings, with the 2017 groups as a reference. We compared each 2017 group with all 2015-2018 groups by calculating the proportion of postcodes in each 2017 group that belong to each 2015-2018 group. We thus obtained a matrix with the percentage of postcodes from 2017 groups that were found in the various *2015-2018* groups (Gelbard et al., 2007).

RESULTS

1.4 The model-based clustering yields a small number of groups of postcodes

The model-based clustering with unconstrained variances had the highest BIC and classified the 5,642 postcodes into 19 groups on the basis of 2017 purchase data for 279 active substances (Figure S2). Most postcodes were unambiguously attributed to a single of these groups, as shown by the bimodal distribution of the probability for a postcode *i* to belong to group k, with most values close to 0 or 1 (Figure S3). Only 13 out of 5,642 postcodes had a maximum probability to be in a group lower than 0.7.

Most groups of postcodes identified by the model-based clustering were spatially aggregated, albeit of contrasting sizes (Figure 1). The number of postcodes per group ranged from 135 to 493 (median = 270, Q1 = 215.5, Q3 = 378.5), which translated into a cropland area per group ranging from 38.7 km^2 to 24,184 km^2 (median = 5,573.7 km^2, Q1 = 1,547.55 km^2, Q3 = 13,959 km^2). The cropland area of groups was negatively related to the area of the convex envelop encompassing it, such that groups with the largest cropland area tended to be the most spatially clustered (Figure 2). Such a spatial clustering of postcodes purchasing similar pesticide substances was expected as agricultural practices are spatially structured (see below) but keep in mind that the model-based clustering did not incorporate spatial information.

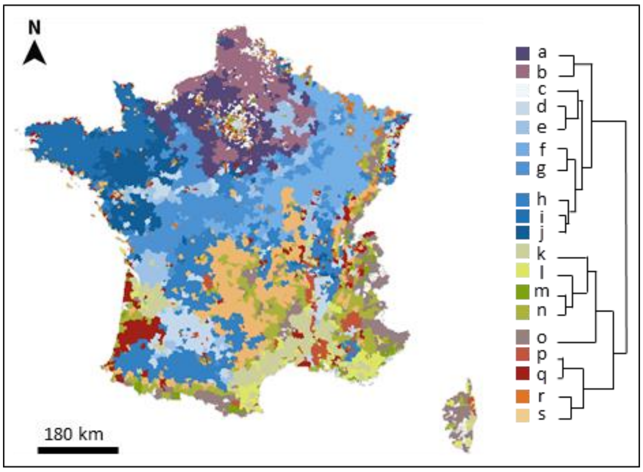


Figure 1: Map of France split into postcode groups obtained from the model-based clustering on the basis of active substances purchased within postcodes in 2017. Postcodes within a group share the same colour. The dendrogram was obtained using an agglomerative hierarchical clustering.

Postcode groups corresponded to specific geographical and/or agricultural regions. For example, group *i* corresponded mostly to Brittany (the western peninsula) and group *b* was predominantly located in Northern France. Groups *e* and *d* were more scattered across the country but overlapped almost perfectly with wine regions (*Figure 2*). Note that a couple of groups were composed of a limited number of postcodes spatially scattered across France (e.g. groups *m and o* Figure 2). In particular, group *m* represented less than 39 km^2 of cropland and is generally discarded in the following.

The groups identified by the model-based clustering were relatively robust to a change in the temporal range of the data, as shown by the results of the clustering on the 2015-2018 data (Figure S7). This second clustering yielded 24 groups and the percentage of shared postcodes between the 2017 groups and their most similar 2015-2018 groups varied between 41% and 80% (median = 62%, Q1 = 53%, Q3 = 66%). For example, groups in Normandie (group *a* vs. group 15) or part of the Languedoc region (group *k* vs. 10) were stable over time (Figure S7). The higher number of groups obtained with the 2015-2018 model-based clustering (24 vs. 19) was often due to the split of some 2017 groups into two 2015-2018 groups. For example, for 2017 group *i*, there was 53% similarity with 2015-2018 group 16 and 40% similarity with group 20 (Figure S7). Because of this temporal consistency in the clustering, we only present in the following the analyses on the 2017 dataset, which is thought be more accurate (see 1.1).

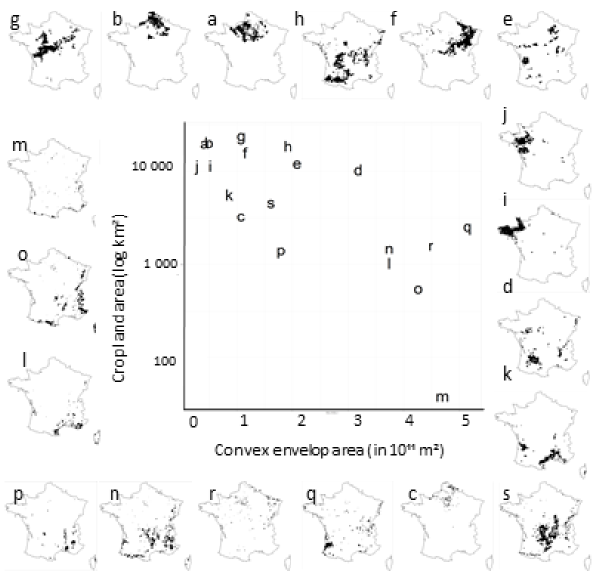
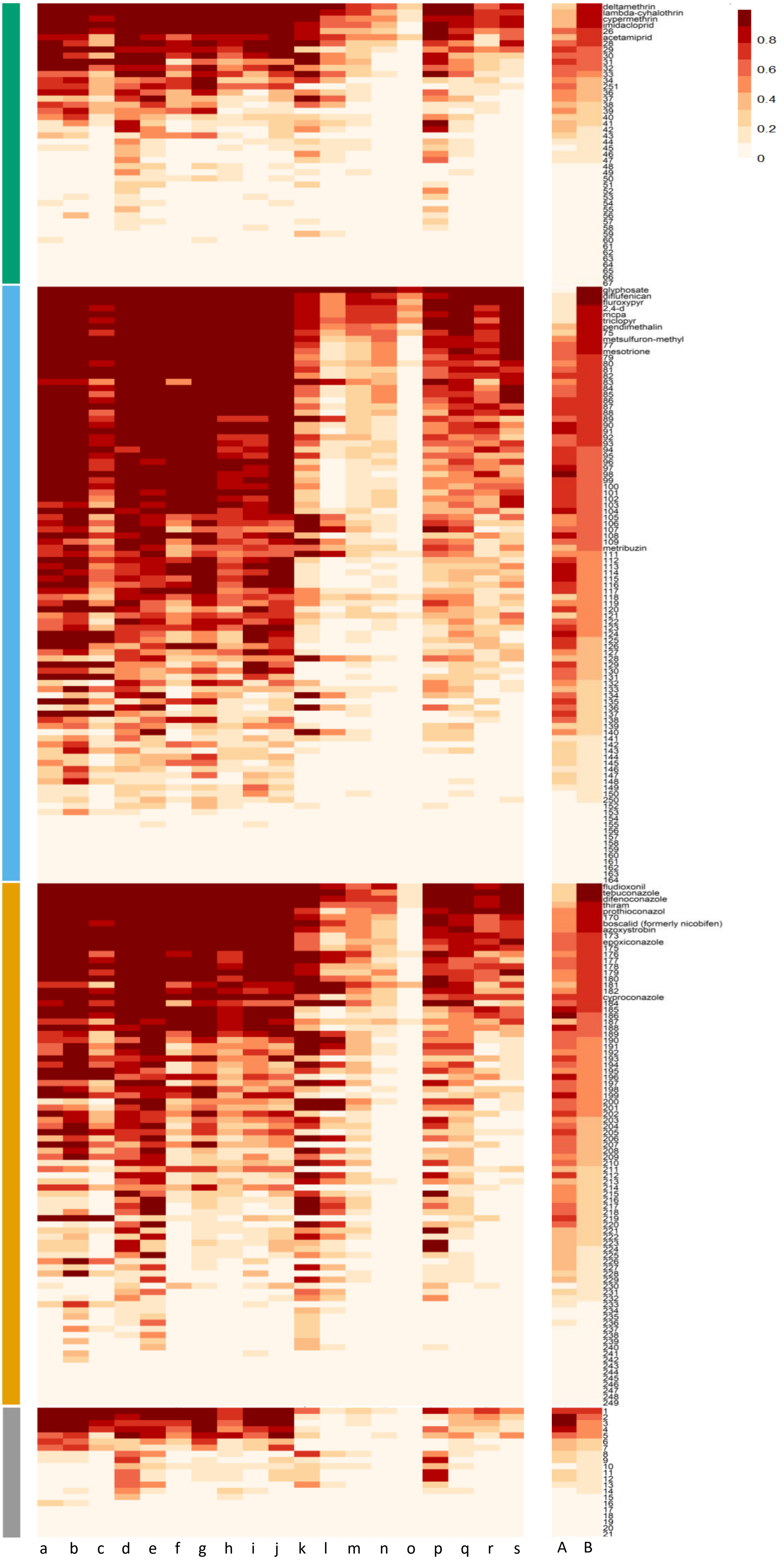


Figure 2: Relationship between cropland area (log scale) and convex area, a proxy for spatial extent, of groups. The spatial distribution of each group is plotted around the relationship, with one map of France per group, in which postcodes forming each group are highlighted in black. Groups are ordered clockwise from top left in decreasing cropland area. Note that the focus on cropland area (not total area) in a postcode makes some groups with little cropland (e.g. mountain areas, q or m) appear with a relatively large black area on the maps, although they are ranked low in terms of cropland area.

1.5 Substance composition of postcode groups: core and discriminating substances

Postcode groups differed in terms of the composition of substances purchased (*Figure 3*), as expected from the clustering algorithm, but may also share common substances. Group composition was inferred, and can be characterised by, (1) the probability of a substance to be purchased in a postcode from a given group ($\hat{\gamma}_{kj}$), and, if the substance is purchased, (2) the estimated mean quantity purchased ($\hat{\mu}_{kj}$) as well as (3) the estimated variance in the latter quantity (σ_{jk}^2). In the following, for the sake of simplicity, we chose to focus on the probability of substances to be purchased, knowing that this probability was positively related with the estimated mean quantity (Figure S4 & Figure S6, *r* = 0.2) and negatively related with the estimated variance (Figure S4, *r* = -0.07). For a given substance, this probability can also vary substantially across groups, and we used this variability to distinguish two main types of substances with interest for the definition of postcode groups and for the identification of relevant pesticide mixtures : core substances and discriminating substances (Figure 4).



deltamethrin
lambda-cyhalothrin
cypermethrin
imidacloprid
acetamiprid
glyphosate
diflufenican
fluroxypyr
2,4-d
mcpa
triclopyr
pendimethalin
metsulfuron-methyl
mesotrione
metribuzin
fludioxonil
tebuconazole
difenoconazole
thiram
prothioconazol
boscalid (formerly nicobifen)
azoxystrobin
epoxiconazole
cyproconazole
0.8
0.6
0.4
0.2
0
a b c d e f g h i j k l m n o p q r s
A B

Figure 3: Heatmap of the probability γ_{kj} in each group, in each of four categories of substances: insecticides (green), herbicides (blue), fungicides (orange), other targets (grey). Within each category, substances are ordered in increasing average probabilities of use across groups. For readability, substance names are not displayed and can be found in Figure S8. On the right of the figure, column A corresponds to the mean probability of use and column B corresponds to the scaled (0,1) variance in probability of use across groups.

Core substances, defined as substances with a high average and low variance of probability to be purchased across groups, were by definition found in most groups; they were widespread molecules that were likely to form the backbone of mixtures encountered by living organisms in farmland. Using an arbitrary threshold value of mean purchase probability of 0.85, we found 12 such core substances with high probabilities (Figure *3* & Figure S5): two pyrethroid insecticides (deltamethrin, lambda-cyhalothrin), six herbicides of different chemical families (glyphosate, diflufenicanil, fluroxypyr, MCPA, 2,4-d, triclopyr) and four fungicides (fludioxonil, tebuconazole, difenoconazole and thiram). Because they were found with high probability in most groups, these substances were unlikely to weight strongly in the definition of postcode groups, although they can contribute via differences in the mean quantities used across groups. For example, the average estimated amount of glyphosate purchased ranged from 19 to 928 kg/ m^2 of cropland (median = 44, Q1 = 38, Q3 = 35) among groups.

Discriminating substances are defined as substances with medium to high mean probability of purchase, mechanically associated with a large variance across groups in this probability (Figure S5). Because of their contrasting probability of purchase across groups, discriminating substances were likely to contribute greatly to the formation of groups. We used the arbitrary range of average probabilities from 0.5 to 0.85 to define discriminating substances. Using these thresholds, we found a set of 84 discriminating substances, including 45 herbicides, 25 fungicides, 10 insecticides and 4 with other targets (Supplementary information 2). In the following, we focus on discriminating substances that are highly probable ($\hat{\gamma}_{kj} > 0.85$) in at least one postcode group, i.e. substances that are likely major components of pesticide mixtures occurring in a given group. We found seven widespread discriminating substances purchased with a probability higher than 0.85 in at least 12 out of 19 groups: azoxystrobin, boscalid, cypermethrin, mesotrione, metsulfuron-methyl, pendimethalin and prothioconazole. These substances are very close to core substances. Conversely, four substances were highly specific, being purchased with high probability (> 0.85) in less than four groups (e.g. metribuzin in groups *d and b*). Within a group, the number of discriminating substances with high probability of purchase (> 0.85) varied strongly among groups, from 2 for group *r* to 80 for group *g* (mean = 43 ± 27). This cross-group variation in the number of highly probable discriminating substances has implication for the composition and complexity of pesticide mixtures in French agroecosystems: from relatively "simple" (12 core substances and 11 discriminating substances in group *g*) to highly complex (12 core substances and 74 discriminating substances in group *g*).

The 156 remaining substances, with a low average probability to be purchased (< 0.5), also had a role in group identification, but were seldom purchased and will not be described further (Figure 3).

1.6 Postcode groups differ in terms of crop composition, but active substance purchase may not be solely driven by crop identity

Groups of postcodes, which by construction are composed of different mixtures of substances, also differed in terms of proportions of cropland grown with various crops, such that groups with close pesticide composition sometimes, but not always, also exhibited similar crop usage (Figure 4). The possible relations between pesticide composition and crop composition can be visualized either on Figure 4, where crop composition of groups similar in terms of pesticides purchases are plotted next to each other, or on the biplot of the log ratio analysis (Figure 5), in which groups with similar crop composition are plotted next to each other. For example, groups *k* and *l*, characterized by a large proportion of vineyards, were close to each other both in the log-ratio analysis, which is indicative of similar crop compositions (Figure 5) and in the hierarchical clustering, which is indicative of similar pesticide purchases (Figure 4). The same was true for groups *b, c* and *i*, and, to a lesser extent, *a*, characterized by an appreciable proposition of crops from the legume/flower category. However, some groups such as *h* and *g* were different in terms of substances (not in the same sub-group, Figure 4) while exhibiting comparable proportions of crop types (Figure 4). Alternatively, some groups that were closely related in terms of substance purchases, such as groups *i* and *h*, could be characterized by dissimilar crop compositions. The latter patterns may suggest regionalisation of substance use, such that neighbouring regions tend to use similar products or substances even with variations in crops grown (e.g. *i* and *h*).

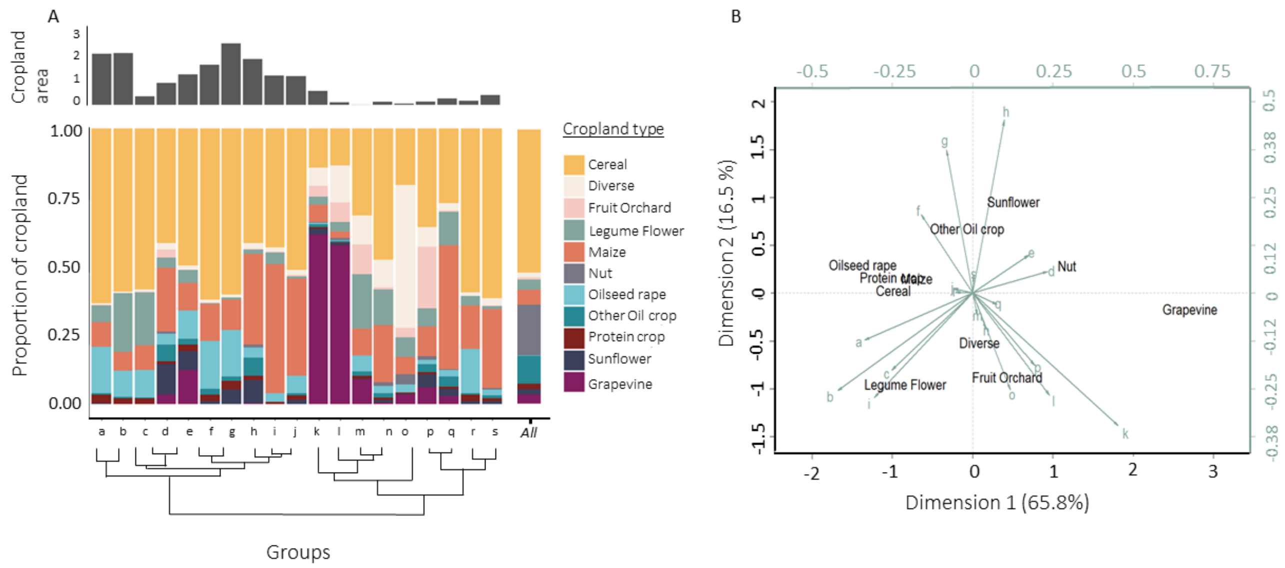


*Figure 4: **A.** Distribution of crop type area across groups. The top grey histogram shows the distribution of total cropland area across groups (in 10^4 km²). The dendrogram was obtained using an agglomerative hierarchical clustering on the basis of Ward's method among groups (see 2.2.1). **B.** Biplot of the log ratio analysis relating the proportion of crop types in each group. Only groups identified as spatially coherent are displayed (see 3.2). For readability, the groups and crop types are displayed on two different scales: black for crop types, green for groups. The size of arrows corresponds to the contribution of each group. Groups that appear close to each other on the biplot have similar crop composition, which can be inferred from the contribution of each crop type to the axes.*

Despite the abovementioned associations between crop composition and active substance compositions of groups, we found no significant correlation between distance matrices: the distance in substance composition among groups was not correlated with the distance in crop composition, although the relationship was marginally significant (Mantel test, ρ = 0.13, P = 0.057). Neither did we found a correlation between the geographic distance and active substance composition of groups (Mantel test, ρ = -0.01, P = 0.53) indicating that adjacent postcode groups do not necessarily exhibit similar composition of active substances adjacent.

DISCUSSION

A major challenge in pesticide risks assessment is to characterise mixtures of pesticides used in the field (Lydy et al., 2004), partly because of the large number of substances used but also because of the limited information on the combinations of substances contaminating the environment. Here, we developed a methodology to analyse a newly available database on pesticide purchases across France. It aimed to identify groups of postcodes with similar compositions of pesticide purchases and characterise their spatial structure, two critical pieces of information to unravel the composition of pesticide mixtures. Our method resulted in the clustering of the 5,642 French postcodes into a relatively low number of groups. These groups represent as many potential pesticide mixtures, which is much lower than the possible combinations among the 279 substances included in the data. In the following, we discuss how our findings can help understand the impacts of pesticides in the environment (e.g. by identifying relevant pesticide mixtures), how this approach can be improved in the future, and the possible mechanisms underlying the groups.

1.7 Significance of the identification of highly probable active substances, and of mixtures of active substances characteristic of postcode groups, for the study of the impacts of pesticides in the environment

The identification of active substances that are purchased with high probability in all (core substances) or a subset (discriminating substances) of postcode groups might contribute to reducing the potential street light effect, whereby most research efforts focus on molecules that are either easy to study (Hendrix, 2017) or that were popularized by previous studies (Tsvetkov and Zayed, 2021). Unsurprisingly, most core substances identified here are already well-known, widely-used substances. Glyphosate is the most widely used broad-spectrum herbicide (Jatinder Pal Kaur Gill et al. 2017; Myers et al. 2016), with associated concerns regarding pervasive direct and indirect effects (Van Bruggen et al., 2018). Tebuconazole and difenoconazole, two triazole fungicides, are widely used and studied (Zubrod et al., 2019). Deltamethrin and lambda-cyhalothrin, two pyrethroids impacting nervous systems (Ray and Fry, 2006; Soderlund and Bloomquist, 1989), are known to have adverse effects on a large range of non-target species such as fish, birds and amphibians (Ali et al. 2011). Yet, a preliminary literature search on these 12 core substances suggests that the research effort on their adverse effects on biodiversity is still highly variable. For core herbicides, a simple search of the molecule name together with "biodiversity" or "ecotoxicology" in the abstract of articles on ISI Web of Science yields more than two hundred research articles for glyphosate and around seventy for 2,4-d, but only 2 to 17 articles for diflufenican, fluroxypyr, MCPA, triclopyr and pendimethalin. For core insecticides, the same search returns ca. 40 articles for lambda-cyhalothrin and deltamethrin. The four core fungicides were no exception, with a number of research articles below ten for thiram, fludioxonil and difenoconazole and around thirty for tebuconazole. Ultimately, our method eases the bottom-up approach in the laboratory by providing a selection of understudied substances deserving further attention.

Studying all possible (combinations of) substances is prohibitive (Wolska et al., 2007); beyond the identification of single substances, our approach chiefly contributes to identifying combinations of active substances that are likely to be encountered in farmland environments, i.e. pesticide mixtures. The model-based clustering identified a relatively small number of postcode groups (19 to 24 depending on the temporal coverage of pesticide data). Each group is characterized by a specific combination of purchases of active substances and can be interpreted as a potential mixture of pesticides occurring in the location of the postcodes, under the assumption that all purchased substances are used within the buying area during the year of purchase (see "Limitations and perspectives" below). Among the 279 active substances considered in these analyses, we highlighted the core substances included in most mixtures and the discriminating substances specific to particular mixtures. Within each postcode group, both types of substances might be a good starting shortlist of substances within which one can investigate potential interactive effects on biodiversity. Indeed, these substances are purchased with high probability in at least some large groups of postcodes, hence are potentially part of widespread mixtures. Although this list is much shorter than the total list of authorized active substances, it still contains 12 core substances, plus 2 to 80 discriminating substances depending on the postcode group. Since our approach to identifying core and discriminating substances was based on probability of purchase only, this shortlist of substances could be narrowed down further by selecting active substances bought in large quantities (see also "Limitations and perspectives") or with high toxicity. The appreciable number of core and discriminating substances composing mixtures is anyway consistent with surveys showing that active substances are rarely found alone in the environment (Silva et al., 2019). It also further substantiates the need for a broader assessment of the synergistic effects of pesticides on biodiversity, often completed on a limited set of substances only (Schreiner et al., 2016; Silva et al., 2019). For core substances, for example, some cocktail effects have already been studied but mostly on pairs of substances (Brodeur et al., 2014; Peluso et al., 2022) and more rarely for cocktails of three or more substances (Cedergreen, 2014; Glinski et al., 2018; Van Meter et al., 2018). Focusing on the reasonable number of relatively complex mixtures identified by the present approach would contribute to improve our understanding of the synergistic effects of realistic cocktails on organisms.

1.8 Limitations & perspectives

1.8.1 Limited spatio-temporal resolution of the BNV-d data

The first limitation of our study is associated with the BNV-d database, which provides information on quantity and year of pesticide purchase, as well as on the administrative location of the buyer, but not on the actual date and location of pesticide treatments, nor on the actual pesticide contamination of the various postcodes. For simplicity, we assumed that the pesticides were used in the year of purchase and in the postcode of purchase and that all substance are equally likely to contaminate the environment. These assumptions may not be verified under all circumstances because farmers are sometimes known to store some pesticide products despite their high prices, e.g. to anticipate increased taxes, and because farms are sometimes spread across several postcodes. Further, not all substances are equally likely to contaminate the environment, e.g. because they vary in terms of degradability or because weather conditions such as wind and rain can affect the way they contaminate the environment. The relationships between pesticide purchase and the ensuing environmental contamination will therefore need further investigation. Yet, there are a couple of indications that the assumption of immediate and local use of pesticides is generally correct. For example, our results are consistent with those of an extensive European study on soil contamination (Silva et al., 2019) which identified glyphosate and the fungicides boscalid, epoxiconazole, and tebuconazole as the most frequent and most abundant contaminants. These substances either belong to the core substances we identified (glyphosate and tebuconazole) or to discriminant substances (boscalid and epoxiconazole) with a high probability of being used over half of the postcode groups.

Although our estimation of pesticide mixture composition may be roughly correct at the resolution of a postcode and of a year, the actual use of pesticides in space and time varies at much finer scales than those of available data. Pesticide substances bought within a given postcode and year may be spread in contrasting fields and times and may not be found together in the environment, depending on their half-life and transport in the environment. The actual mixture composition of a site hence depends, among others, on the crop cover in the landscape and associated farming practices. In particular, the amount of organic farming within the identified postcode groups may affect local heterogeneity in the quantity and composition of substances used, although pesticides approved for organic farming were generally not part of our analysis and may add up to pesticides used for conventional farming. Downscaling the BNV-d database to the field scale is challenging (Cahuzac et al. 2018; Ramalanjaona, 2020), but it might reveal other patterns than the ones we highlighted here, probably decreasing the number of substances that are part of local mixtures. Such fine-grained data on pesticides might be more relevant to assess the impact of pesticide contamination on biodiversity.

1.8.2 Going beyond the use of purchase probabilities and arbitrary thresholds to identify the substances of interest for risk assessment

The method we developed is continuous, with quantitative estimates of purchase probabilities, as well as mean and variance of quantities purchased per postcode group. Still, we used arbitrary thresholds to identify core and discriminating substances. The mixture compositions we highlighted here are thus dependent on the chosen thresholds. Depending on the question of interest, these thresholds can and should be adapted. For example, by changing the threshold to 0.80, there are nine more core substances, and among these substances there are, for example, imidacloprid and boscalid, both known for high use and effects on biodiversity (Lopez-Antia et al., 2015; Qian et al., 2018; Simon-Delso et al., 2017; Yang et al., 2008).

In addition, most of our interpretation of pesticide mixture composition relies on the estimated purchase probabilities, but these mixtures were also identified using information on the mean and variance of purchased amounts within postcodes, hence mixtures differ for these variables as well. For example, glyphosate, a core substance with high purchase probability in all postcode groups, was bought in contrasting quantities across postcode groups: the average amount was 53.9 kg/km^2 and ranged from 7.8 kg/km^2 in group *p* to 146 kg/km^2 in group *i*. Although the purchase probability was positively correlated to the mean purchased quantity and negatively to its variance, the correlation is not strong, and further analysis is needed to fully uncover variation in substance quantities within the mixtures we identified.

1.8.3 Taking into account the yearly variation in pesticide use

Our analysis appeared relatively robust to the time period of the pesticide purchase data, as suggested by the comparison of postcode groups obtained with the 2017 and the 2015-2018 datasets. This strong correlation between the 2017 and the 2015-2018 analysis is not entirely surprising because of the presence of the 2017 data in both analyses. Yet, adding three years of data into the analysis did not affect much the composition of postcode groups, which suggests relatively stable patterns of pesticide purchase in France over a short time period. Nonetheless, we observed some differences, mainly due to the split of some groups, which were also expected due to climatic variation, changes in legislation on pesticide use (Urruty et al., 2016) or changes in crop areas (Levavasseur et al., 2016). A better integration of the temporal dynamics of pesticide purchases in the characterisation of pesticide mixtures is needed if we are to monitor pesticide mixtures across France. This can be achieved by applying the model-based clustering to each year of data separately. Investigating the spatial stability of groups and mixture compositions across years would contribute to either estimate annual mixtures or to find temporarily stable mixtures. Finding recurrent mixtures could facilitate risk assessment over years. Indeed, this could provide key information on the frequency of mixtures encountered by organisms as repeated contact might increase risks (Stuligross and Williams, 2021).

1.9 Postcode groups are related to the crop they grow, as well as to other regional factors, but the underlying mechanisms remain to be fully identified

Although no spatial information was included in the model-based clustering analysis, the postcode groups exhibited a strong spatial structure, in which most groups are strongly aggregated and only a few small groups are scattered across France. Such spatial structure was expected since pesticide use is strongly crop-dependent. For example, acetamiprid, a substance used to protect fruit trees or grapevine against aphids, is bought with high probability in groups *l*, *e* and *d*, with high proportion of fruit orchards and grapevines. Similarly, cyproconazole, a substance with a broader spectrum of use, is bought with high probability in several groups with contrasting crop compositions (*a, b, e, f, g, h, j, k, l, n, o, q, r* Figure 4). However, deviations from this pattern were found: some adjacent postcode groups can have different sets of crops but similar substance purchases or some spatially distant postcode groups can have similar sets of crops but different substance purchases. This observation suggests that local conditions, such as climate or pests, or some regional patterns in the pesticide market and/or distribution, can drive the purchase of active substances more than the set of crops grown (Silva et al., 2019; Storck et al., 2017). Hence, the differences among postcode groups were related to a combination of crop identity effects and other regional effects that will need additional analysis to be identified. A straightforward perspective for the model-based clustering approach would thus be to incorporate environmental covariates in the model, and evaluate how clusters are modified.

CONCLUSION

This study shows that a reasonably low number of substance mixtures can be identified at the scale of France. Pursuing ecotoxicological studies on the synergistic effects of mixtures will make it possible to identify risks and better understand the effects of pesticides on organisms. The mapping of these pesticide mixtures enables the identification of regions under different regimes of pesticide contamination. This might be particularly useful to plan *in situ* tests for both pesticide contamination and effects on biodiversity. Here we did not investigate the effects of cocktails on wild organisms, and further work should be done on this aspect.

Acknowledgement

This project was funded and supported by ANSES (grant agreement 2019-CRB-03_PV19) via the tax on sales of plant protection products. The proceeds of this tax are assigned to ANSES to finance the establishment of the system for monitoring the adverse effects of plant protection products, called ‘phytopharmacovigilance’ (PPV), established by the French Act on the future of agriculture of 13 October 2014. We wish to thank the steering committee of the project: Fabrizio Botta, Sandrine Charles, Marc Girondot, Olivier Le Gall, Thomas Quintaine, and Lynda Saibi-Yedjer. Milena Cairo was supported by ANR project VITIBIRD (ANR-20-CE34-0008) while working on this project. This work also benefitted from the support of the project ECONET (ANR-18-CE02-0010) and of the “Chaire Modélisation Mathématique et Biodiversité”.

Conflict of interest

The authors declare they have no conflict of interest relating to the content of this article

SUPPLEMENTARY MATERIALS

Supplementary materials to this article can be found online at https://doi.org/10.5281/zenodo.7693149

REFERENCES

Ali, S. F., Shieh, B. H., Alehaideb, Z., Khan, M. Z., Louie, A., Fageh, N., & Law, F. C. (2011). A review on the effects of some selected pyrethroids and related agrochemicals on aquatic vertebrate biodiversity. Canadian Journal of Pure & Applied Sciences, 5(2), 1455-1464.

Altenburger, R., Backhaus, T., Boedeker, W., Faust, M., Scholze, M., 2013. Simplifying complexity: Mixture toxicity assessment in the last 20 years. Environ. Toxicol. Chem. 32, 1685–1687. https://doi.org/10.1002/etc.2294

Boedeker, W., Watts, M., Clausing, P., Marquez, E., 2020. The global distribution of acute unintentional pesticide poisoning: estimations based on a systematic review. BMC Public Health 20, 1–19. https://doi.org/10.1186/s12889-020-09939-0

Bopp, S.A.K., Klenzier, A., van der Linden, S., Lamon, L., Paini, A., Parissis, N.,

Richarz, A.-N., Triebe, J., Worth, A., 2016. Review of case studies on the human and environmental risk assessment of chemical mixtures. https://doi.org/10.2788/272583
Botías, C., David, A., Horwood, J., Abdul-Sada, A., Nicholls, E., Hill, E., Goulson, D., 2015. Neonicotinoid Residues in Wildflowers, a Potential Route of Chronic Exposure for Bees. Environ. Sci. Technol. 49, 12731–12740. https://doi.org/10.1021/acs.est.5b03459
Brittain, C.A., Vighi, M., Bommarco, R., Settele, J., Potts, S.G., 2010. Impacts of a pesticide on pollinator species richness at different spatial scales. Basic Appl. Ecol. 11, 106–115. https://doi.org/10.1016/j.baae.2009.11.007
Brodeur, J.C., Poliserpi, M.B., D'Andrea, M.F., Sánchez, M., 2014. Synergy between glyphosate- and cypermethrin-based pesticides during acute exposures in tadpoles of the common South American Toad Rhinella arenarum. Chemosphere 112, 70–76. https://doi.org/10.1016/j.chemosphere.2014.02.065
Busse, M.D., Ratcliff, A.W., Shestak, C.J., Powers, R.F., 2001. Glyphosate toxicity and the effects of long-term vegetation control on soil microbial communities. Soil Biol. Biochem. 33, 1777–1789. https://doi.org/10.1016/S0038-0717(01)00103-1
Cantelaube, P., Carles, M., 2010. Le registre parcellaire graphique : des donn é es g é ographiques pour d é crire la couverture du sol agricole.
Cedergreen, N., 2014. Quantifying synergy: A systematic review of mixture toxicity studies within environmental toxicology. PLoS One 9. https://doi.org/10.1371/journal.pone.0096580
Deguines, N., Jono, C., Baude, M., Henry, M., Julliard, R., Fontaine, C., 2014. Large-scale trade-off between agricultural intensification and crop pollination services. Front. Ecol. Environ. 12, 212–217. https://doi.org/10.1890/130054
Dempster;A.P, Laird, N.., Rubin, D.., 1977. Maximum Likelihood from Incomplete data via the EM Algorithm.
Dudley, N., Attwood, S.J., Goulson, D., Jarvis, D., Bharucha, Z.P., Pretty, J., 2017. How should conservationists respond to pesticides as a driver of biodiversity loss in agroecosystems? Biol. Conserv. 209, 449–453. https://doi.org/10.1016/j.biocon.2017.03.012
Fritsch, C., Appenzeller, B., Burkart, L., Coeurdassier, M., Scheifler, R., Raoul, F., Driget, V., Powolny, T., Gagnaison, C., Rieffel, D., Afonso, E., Goydadin, A.C., Hardy, E.M., Palazzi, P., Schaeffer, C., Gaba, S., Bretagnolle, V., Bertrand, C., 2022. Pervasive exposure of wild small mammals to legacy and currently used pesticide mixtures in arable landscapes. Sci. Rep. 1–22. https://doi.org/10.1038/s41598-022-19959-y
Furlan, L., Pozzebon, A., Duso, C., Simon-Delso, N., Sánchez-Bayo, F., Marchand, P.A., Codato, F., Bijleveld van Lexmond, M., Bonmatin, J.M., 2018. An update of the Worldwide Integrated Assessment (WIA) on systemic insecticides. Part 3: alternatives to systemic insecticides. Environ. Sci. Pollut. Res. 1–23. https://doi.org/10.1007/s11356-017-1052-5
Geiger, F., Bengtsson, J., Berendse, F., Weisser, W.W., Emmerson, M., Morales, M.B., Ceryngier, P., Liira, J., Tscharntke, T., Winqvist, C., Eggers, S., Bommarco, R., Pärt, T., Bretagnolle, V., Plantegenest, M., Clement, L.W., Dennis, C., Palmer, C., Oñate, J.J., Guerrero, I., Hawro, V., Aavik, T., Thies, C., Flohre, A., Hänke, S., Fischer, C., Goedhart, P.W., Inchausti, P., 2010. Persistent negative effects of pesticides on biodiversity and biological control potential on European farmland. Basic Appl. Ecol. 11, 97–105.

https://doi.org/10.1016/j.baae.2009.12.001
Gelbard, R., Goldman, O., Spiegler, I., 2007. Investigating diversity of clustering methods: An empirical comparison. Data Knowl. Eng. 63, 155–166. https://doi.org/10.1016/j.datak.2007.01.002
Gibbons, D., Morrissey, C., Mineau, P., 2015. A review of the direct and indirect effects of neonicotinoids and fipronil on vertebrate wildlife. Environ. Sci. Pollut. Res. 22, 103–118. https://doi.org/10.1007/s11356-014-3180-5
Glinski, D.A., Purucker, S.T., Van Meter, R.J., Black, M.C., Henderson, W.M., 2018. Endogenous and exogenous biomarker analysis in terrestrial phase amphibians (Lithobates sphenocephala) following dermal exposure to pesticide mixtures. Env. chem 60, 1–24. https://doi.org/10.1071/EN18163.
Greenacre, M., 2019. Variable Selection in Compositional Data Analysis Using Pairwise Logratios. Math. Geosci. 51, 649–682. https://doi.org/10.1007/s11004-018-9754-x
Hallmann, C.A., Foppen, R.P.B., Van Turnhout, C.A.M., De Kroon, H., Jongejans, E., 2014. Declines in insectivorous birds are associated with high neonicotinoid concentrations. Nature 511, 341–343. https://doi.org/10.1038/nature13531
Hendrix, C.S., 2017. The streetlight effect in climate change research on Africa. Glob. Environ. Chang. 43, 137–147. https://doi.org/10.1016/j.gloenvcha.2017.01.009
Hernández, A.F., Gil, F., Lacasaña, M., 2017. Toxicological interactions of pesticide mixtures: an update. Arch. Toxicol. 91, 3211–3223. https://doi.org/10.1007/s00204-017-2043-5
Heys, K.A., Shore, R.F., Pereira, M.G., Jones, K.C., Martin, F.L., 2016. Risk assessment of environmental mixture effects. RSC Adv. 6, 47844–47857. https://doi.org/10.1039/c6ra05406d
Humann-Guilleminot, Ségolène, Binkowski, Ł.J., Jenni, L., Hilke, G., Glauser, G., Helfenstein, F., 2019. A nation-wide survey of neonicotinoid insecticides in agricultural land with implications for agri-environment schemes. J. Appl. Ecol. 56, 1502–1514. https://doi.org/10.1111/1365-2664.13392
Humann-Guilleminot, S., Tassin de Montaigu, C., Sire, J., Grünig, S., Gning, O., Glauser, G., Vallat, A., Helfenstein, F., 2019. A sublethal dose of the neonicotinoid insecticide acetamiprid reduces sperm density in a songbird. Environ. Res. 177, 108589. https://doi.org/10.1016/j.envres.2019.108589
Junghans, M., Backhaus, T., Faust, M., Scholze, M., Grimme, L.H., 2006. Application and validation of approaches for the predictive hazard assessment of realistic pesticide mixtures. Aquat. Toxicol. 76, 93–110. https://doi.org/10.1016/j.aquatox.2005.10.001
Keplinger, M.L., Deichmann, W.B., 1967. Acute toxicity of combinations of pesticides. Toxicol. Appl. Pharmacol. 10, 586–595. https://doi.org/10.1016/0041-008X(67)90097-X
Levavasseur, F., Martin, P., Bouty, C., Barbottin, A., Bretagnolle, V., Thérond, O., Scheurer, O., Piskiewicz, N., 2016. RPG Explorer: A new tool to ease the analysis of agricultural landscape dynamics with the Land Parcel Identification System. Comput. Electron. Agric. 127, 541–552. https://doi.org/10.1016/j.compag.2016.07.015
Lewis, K.A., Tzilivakis, J., Warner, D.J., Green, A., 2016. An international database for pesticide risk assessments and management. Hum. Ecol. risk Assess. 22, 1050–1064. https://doi.org/10.1017/CBO9781107415324.004
Lopez-Antia, A., Ortiz-Santaliestra, M.E., Mougeot, F., Mateo, R., 2015. Imidacloprid-treated seed ingestion has lethal effect on adult partridges and reduces both

breeding investment and offspring immunity. Environ. Res. 136, 97–107. https://doi.org/10.1016/j.envres.2014.10.023

Lydy, M., Belden, J., Wheelock, C., Hammock, B., Denton, D., 2004a. Challenges in regulating pesticide mixtures. Ecol. Soc. 9. https://doi.org/10.5751/ES-00694-090601

Lydy, M., Belden, J., Wheelock, C., Hammock, B., Denton, D., 2004b. Challenges in Regulating Pesticide Mixtures. Ecol. Soc. 53, 1689–1699.

Mahmood, I., Sameen, R.I., Shazadi, K., Alvina, G., Hakeem, K.R., 2016. Effects of Pesticides on Environment. Plant, Soil Microbes Vol. 1 Implic. Crop Sci. 1–366. https://doi.org/10.1007/978-3-319-27455-3

Millot, F., Decors, A., Mastain, O., Quintaine, T., Berny, P., Vey, D., Lasseur, R., Bro, E., 2017. Field evidence of bird poisonings by imidacloprid-treated seeds: a review of incidents reported by the French SAGIR network from 1995 to 2014. Environ. Sci. Pollut. Res. 24, 5469–5485. https://doi.org/10.1007/s11356-016-8272-y

Navarro, J., Hadjikakou, M., Ridoutt, B., Parry, H., Bryan, B.A., 2021. Pesticide toxicity hazard of agriculture: regional and commodity hotspots in Australia. Environ. Sci. Technol. 55, 1290–1300. https://doi.org/10.1021/acs.est.0c05717

Oksanen, J., Simpson, G.L., 2022. Package ‘ vegan .’

Peluso, J., Furió Lanuza, A., Pérez Coll, C.S., Aronzon, C.M., 2022. Synergistic effects of glyphosate- and 2,4-D-based pesticides mixtures on Rhinella arenarum larvae. Environ. Sci. Pollut. Res. 29, 14443–14452. https://doi.org/10.1007/s11356-021-16784-0

Qian, L., Qi, S., Cao, F., Zhang, J., Zhao, F., Li, C., Wang, C., 2018. Toxic effects of boscalid on the growth, photosynthesis, antioxidant system and metabolism of Chlorella vulgaris. Environ. Pollut. 242, 171–181. https://doi.org/10.1016/j.envpol.2018.06.055

Ramalanjaona, L., 2020. Mise à jour du calcul des coefficients de répartition spatiale des données de la BNVd Note méthodologique 95.

Ray, D.E., Fry, J.R., 2006. A reassessment of the neurotoxicity of pyrethroid insecticides. Pharmacol. Ther. 111, 174–193. https://doi.org/10.1016/j.pharmthera.2005.10.003

Relyea, R.A., 2009. A cocktail of contaminants: How mixtures of pesticides at low concentrations affect aquatic communities. Oecologia 159, 363–376. https://doi.org/10.1007/s00442-008-1213-9

Rundlöf, M., Andersson, G.K.S., Bommarco, R., Fries, I., Hederström, V., Herbertsson, L., Jonsson, O., Klatt, B.K., Pedersen, T.R., Yourstone, J., Smith, H.G., 2015. Seed coating with a neonicotinoid insecticide negatively affects wild bees. Nature 521, 77–80. https://doi.org/10.1038/nature14420

Schreiner, V.C., Szöcs, E., Bhowmik, A.K., Vijver, M.G., Schäfer, R.B., 2016. Pesticide mixtures in streams of several European countries and the USA. Sci. Total Environ. 573, 680–689. https://doi.org/10.1016/j.scitotenv.2016.08.163

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6, 461–464.

Sheahan, M., Barrett, C.B., Goldvale, C., 2017. Human health and pesticide use in Sub-Saharan Africa. Agric. Econ. (United Kingdom) 48, 27–41. https://doi.org/10.1111/agec.12384

Silva, V., Mol, H.G.J., Zomer, P., Tienstra, M., Ritsema, C.J., Geissen, V., 2019. Pesticide residues in European agricultural soils – A hidden reality unfolded. Sci. Total Environ. 653, 1532–1545. https://doi.org/10.1016/j.scitotenv.2018.10.441

Simon-Delso, N., San Martin, G., Bruneau, E., Hautier, L., Medrzycki, P., 2017.

Toxicity assessment on honey bee larvae of a repeated exposition of a systemic fungicide, boscalid. Bull. Insectology 70, 83–90.
Soderlund, D.M., Bloomquist, J.R., 1989. Neurotoxic actions of pyrethroid insecticides. Annu. Rev. Entomol. 34, 77–96. https://doi.org/10.1146/annurev.en.34.010189.000453
Storck, V., Karpouzas, D.G., Martin-Laurent, F., 2017. Towards a better pesticide policy for the European Union. Sci. Total Environ. 575, 1027–1033. https://doi.org/10.1016/j.scitotenv.2016.09.167
Stuligross, C., Williams, N.M., 2021. Past insecticide exposure reduces bee reproduction and population growth rate. Proc. Natl. Acad. Sci. U. S. A. 118, 1–6. https://doi.org/10.1073/pnas.2109909118
Tang, F.H.M., Lenzen, M., McBratney, A., Maggi, F., 2021. Risk of pesticide pollution at the global scale. Nat. Geosci. 14, 206–210. https://doi.org/10.1038/s41561-021-00712-5
Tassinde Montaigu, C., Goulson, D., 2020. Identifying agricultural pesticides that may pose a risk for birds. PeerJ.
Tsvetkov, N., Zayed, A., 2021. Searching beyond the streetlight: Neonicotinoid exposure alters the neurogenomic state of worker honey bees. Ecol. Evol. 11, 18733–18742. https://doi.org/10.1002/ece3.8480
Urruty, N., Deveaud, T., Guyomard, H., Boiffin, J., 2016. Impacts of agricultural land use changes on pesticide use in French agriculture. Eur. J. Agron. 80, 113–123. https://doi.org/10.1016/j.eja.2016.07.004
Van Bruggen, A.H.C., He, M.M., Shin, K., Mai, V., Jeong, K.C., Finckh, M.R., Morris, J.G., 2018. Environmental and health effects of the herbicide glyphosate. Sci. Total Environ. 616–617, 255–268. https://doi.org/10.1016/j.scitotenv.2017.10.309
Van Meter, R.J., Glinski, D.A., Purucker, S.T., Henderson, W.M., 2018. Influence of exposure to pesticide mixtures on the metabolomic profile in post-metamorphic green frogs (Lithobates clamitans). Sci. Total Environ. 624, 1348–1359. https://doi.org/10.1016/j.scitotenv.2017.12.175
Wolska, L., Sagajdakow, A., Kuczyńska, A., Namieśnik, J., 2007. Application of ecotoxicological studies in integrated environmental monitoring: Possibilities and problems. TrAC - Trends Anal. Chem. 26, 332–344. https://doi.org/10.1016/j.trac.2006.11.012
Yang, E.C., Chuang, Y.C., Chen, Y.L., Chang, L.H., 2008. Abnormal foraging behavior induced by sublethal dosage of imidacloprid in the honey bee (Hymenoptera: Apidae). J. Econ. Entomol. 101, 1743–1748. https://doi.org/10.1603/0022-0493-101.6.1743
Zubrod, J.P., Bundschuh, M., Arts, G., Brühl, C.A., Imfeld, G., Knäbel, A., Payraudeau, S., Rasmussen, J.J., Rohr, J., Scharmüller, A., Smalling, K., Stehle, S., Schulz, R., Schäfer, R.B., 2019. Fungicides: An Overlooked Pesticide Class? Environ. Sci. Technol. 53, 3347–3365. https://doi.org/10.1021/acs.est.8b04392

APPENDIX

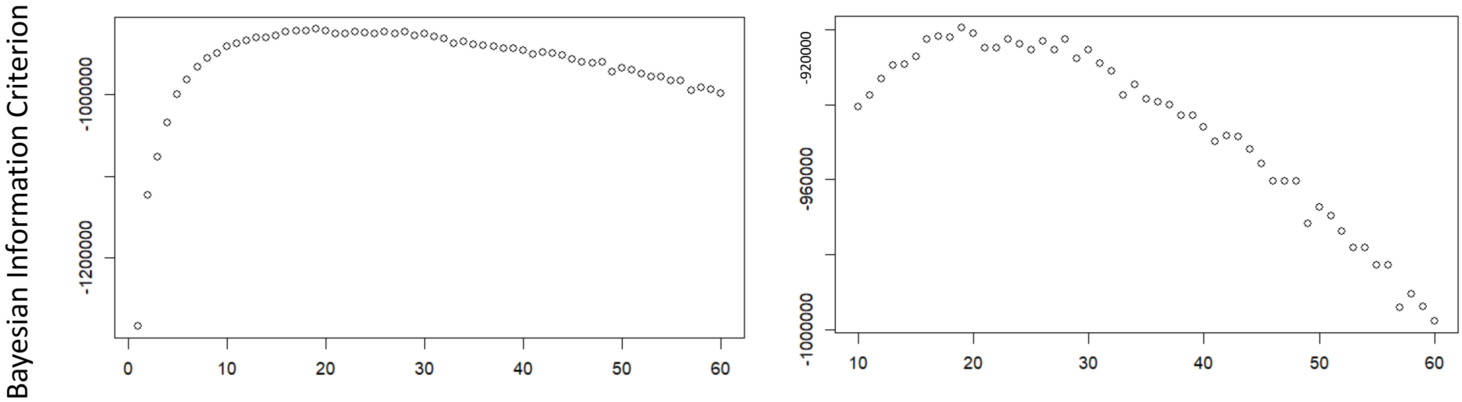


Figure S1 *: Values of BIC as a function of the number of groups in the EM algorithm. Panel a shows the full range of number of groups tested (from 1 to 40). Panel b is a closeup around the maximum BIC value*

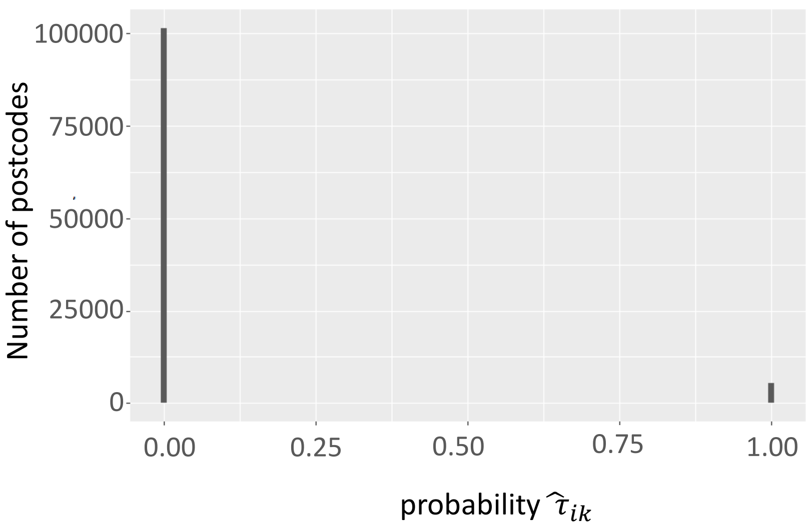


***Figure S2**: Distribution of $\hat{\tau}_{ik}$, the probability of postcode i to be in group k*

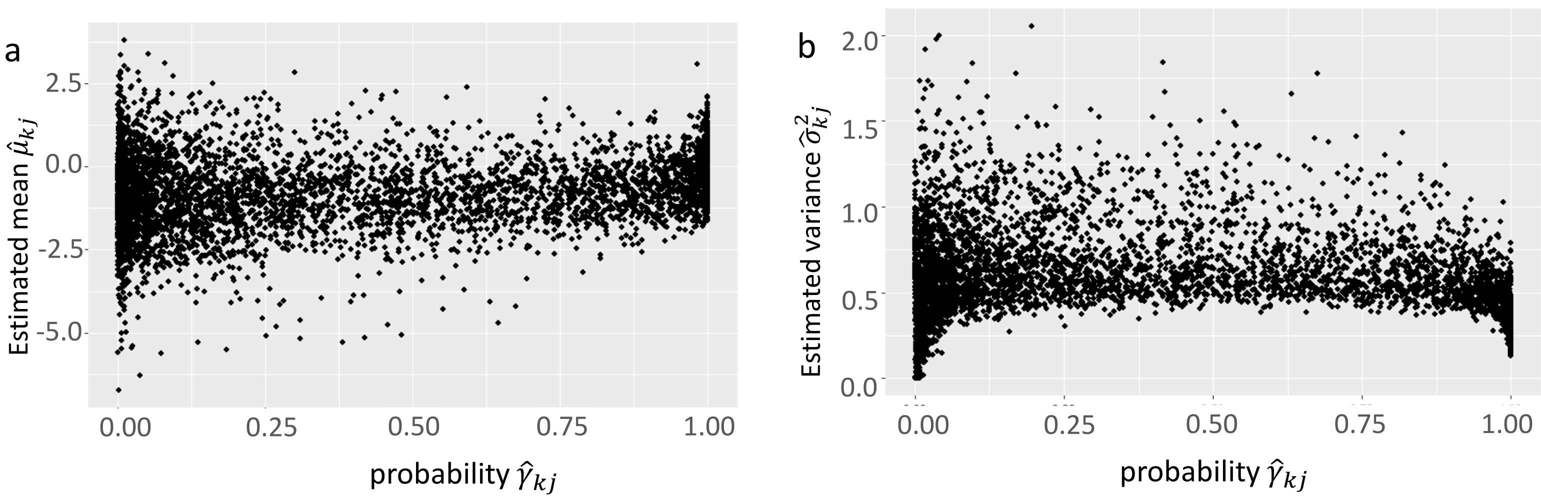
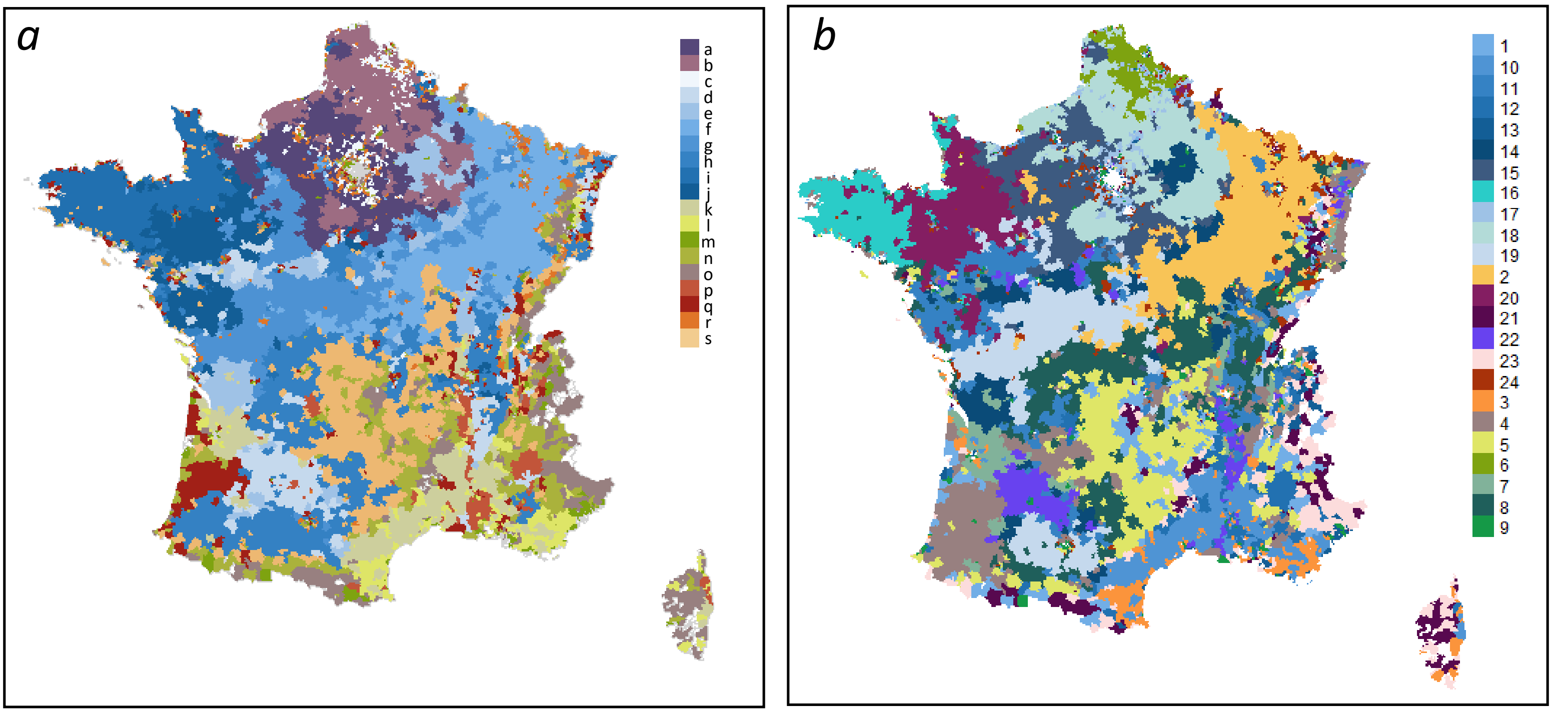


***Figure S3**: Estimated mean ($\hat{\mu}_{kj}$, panel a) and variance $\hat{\sigma}^2_{kj}$, panel b) of substance quantities purchased in a group as a function of the probability of a substance to be in a group $\hat{\gamma}_{kj}$.*



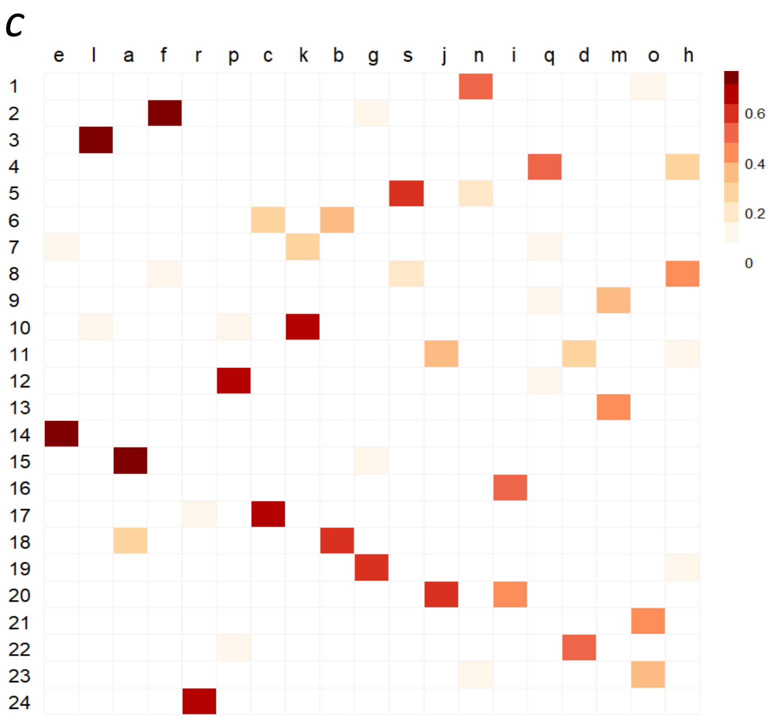


***Figure S4**: Differences and similarities in the clustering of postcodes produced by the mixture model with only 2017 substance purchase data (a) or 2015-2018 data (b). Postcode within a group share the same colour.*
Panel (c) shows proximity of the 2017 groups with 2015-2018 groups on a heatmap, expressed as the percentage postcodes from 2017 groups that were found in the various 2015-2018 groups. The graph should be read vertically: for example, 2017 group i is split mostly into 2015-2018 groups 16 (53%) and 20 (40%) In contrast, 79% postcodes of 2017 group e are found in 2015-2018 group 14.

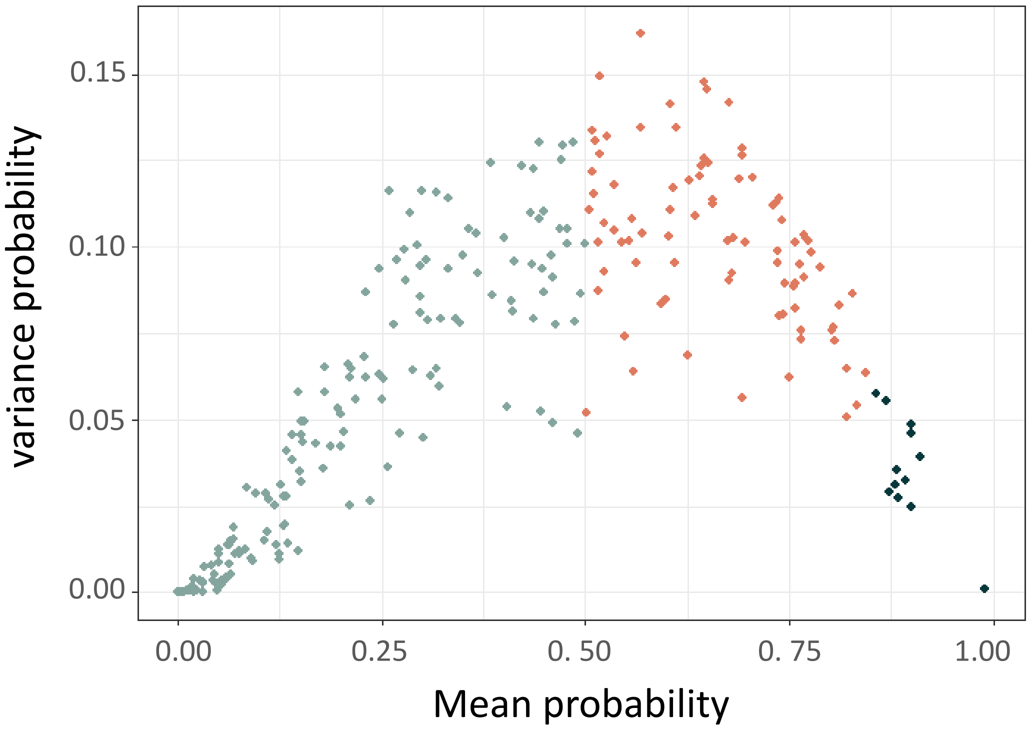
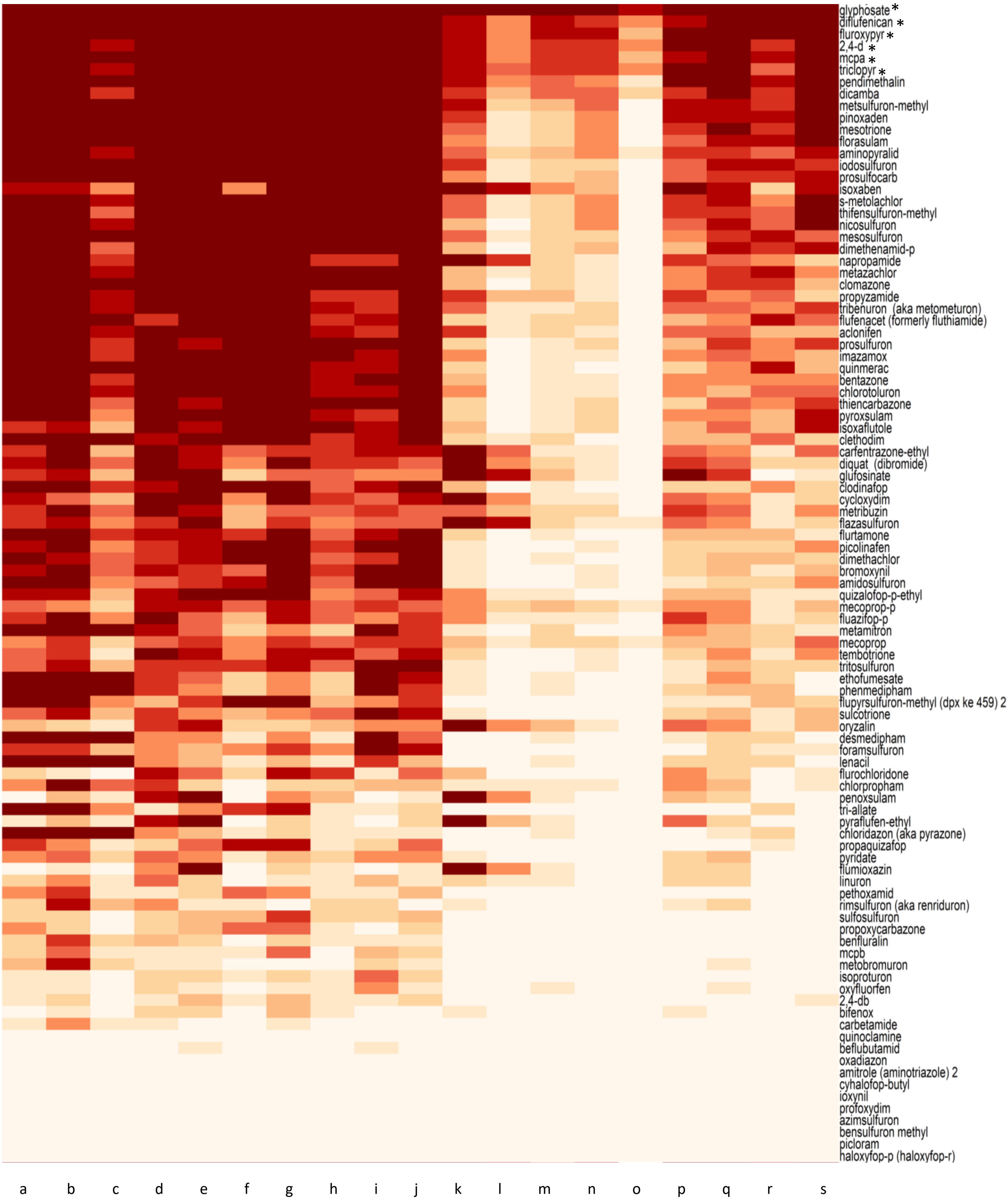
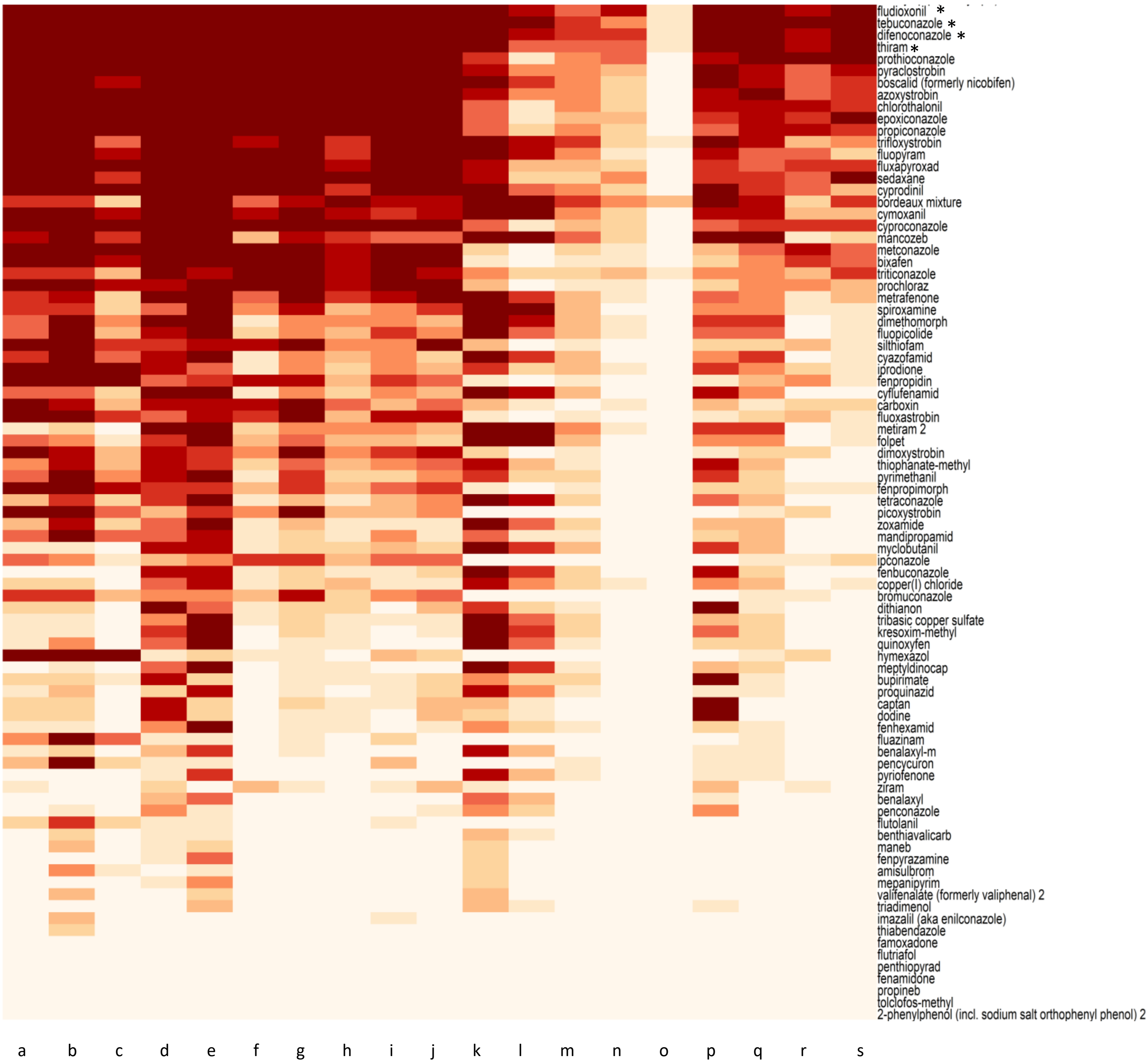


Figure S5:*Variance of probabilities of substances to be in a group as a function of their mean probability to be in a group. Colours were set to show other (grey) , discriminant (orange) and core (black) substances.*

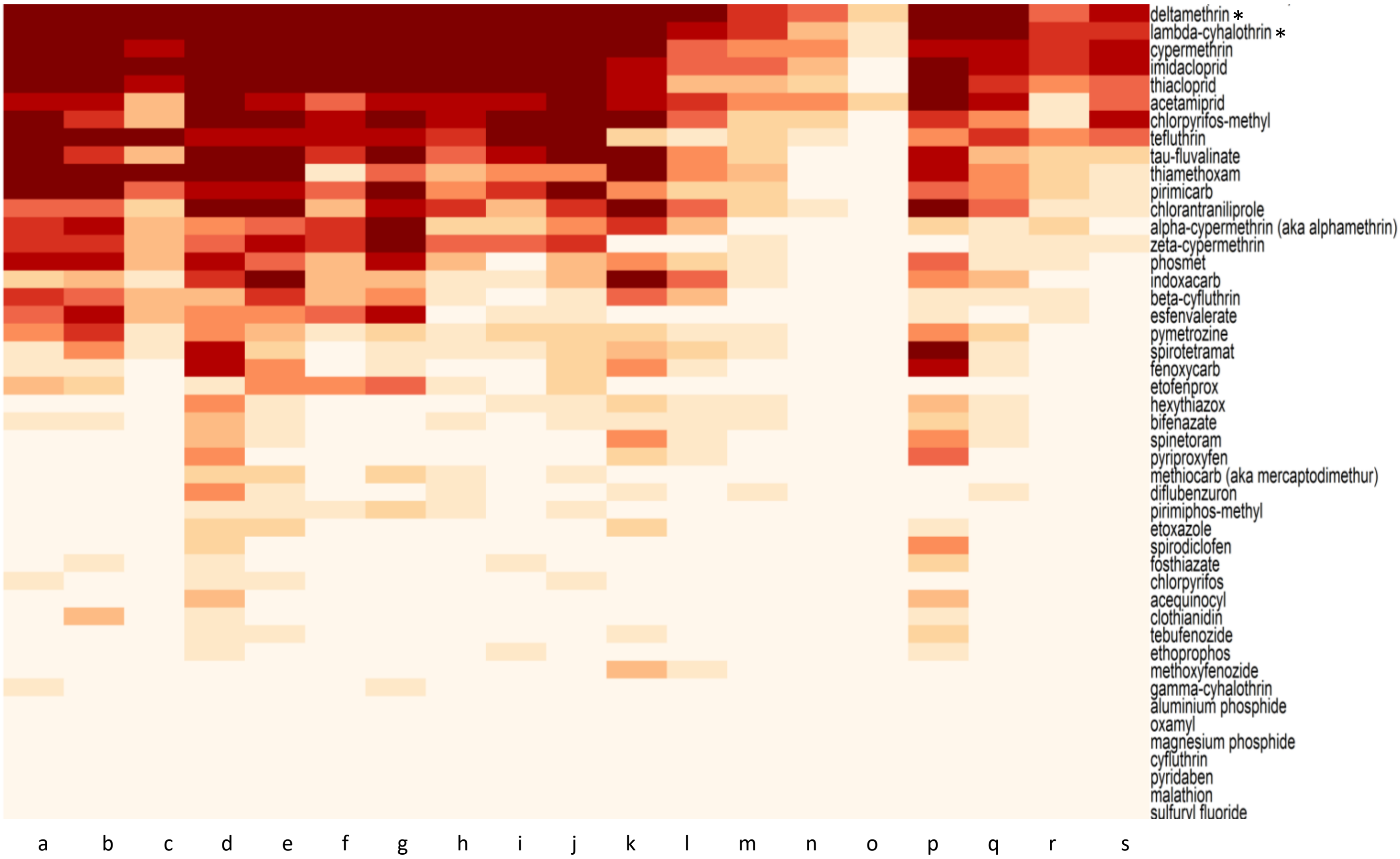
Herbicides


glyphosate *
diflufenican *
fluroxypyr *
2,4-d *
mcpa *
triclopyr *
pendimethalin
dicamba
metsulfuron-methyl
pinoxaden
mesotrione
florasulam
aminopyralid
iodosulfuron
prosulfocarb
isoxaben
s-metolachlor
thifensulfuron-methyl
nicosulfuron
mesosulfuron
dimethenamid-p
napropamide
metazachlor
clomazone
propyzamide
tribenuron (aka metometuron)
flufenacet (formerly fluthiamide)
aclonifen
prosulfuron
imazamox
quinmerac
bentazone
chlorotoluron
thiencarbazone
pyroxsulam
isoxaflutole
clethodim
carfentrazone-ethyl
diquat (dibromide)
glufosinate
clodinafop
cycloxydim
metribuzin
flazasulfuron
flurtamone
picolinafen
dimethachlor
bromoxynil
amidosulfuron
quizalofop-p-ethyl
mecoprop-p
fluazifop-p
metamitron
mecoprop
tembotrione
tritosulfuron
ethofumesate
phenmedipham
flupyrsulfuron-methyl (dpx ke 459) 2
sulcotrione
oryzalin
desmedipham
foramsulfuron
lenacil
flurochloridone
chlorpropham
penoxsulam
tri-allate
pyraflufen-ethyl
chloridazon (aka pyrazone)
propaquizafop
pyridate
flumioxazin
linuron
pethoxamid
rimsulfuron (aka renriduron)
sulfosulfuron
propoxycarbazone
benfluralin
mcpb
metobromuron
isoproturon
oxyfluorfen
2,4-db
bifenox
carbetamide
quinoclamine
beflubutamid
oxadiazon
amitrole (aminotriazole) 2
cyhalofop-butyl
ioxynil
profoxydim
azimsulfuron
bensulfuron methyl
picloram
haloxyfop-p (haloxyfop-r)
a
b
c
d
e
f
g
h
i
j
k
l
m
n
o
p
q
r
s

Fungicides


fludioxonil *
tebuconazole *
difenoconazole *
thiram *
prothioconazole
pyraclostrobin
boscalid (formerly nicobifen)
azoxystrobin
chlorothalonil
epoxiconazole
propiconazole
trifloxystrobin
fluopyram
fluxapyroxad
sedaxane
cyprodinil
bordeaux mixture
cymoxanil
cyproconazole
mancozeb
metconazole
bixafen
triticonazole
prochloraz
metrafenone
spiroxamine
dimethomorph
fluopicolide
silthiofam
cyazofamid
iprodione
fenpropidin
cyflufenamid
carboxin
fluoxastrobin
metiram 2
folpet
dimoxystrobin
thiophanate-methyl
pyrimethanil
fenpropimorph
tetraconazole
picoxystrobin
zoxamide
mandipropamid
myclobutanil
ipconazole
fenbuconazole
copper(I) chloride
bromuconazole
dithianon
tribasic copper sulfate
kresoxim-methyl
quinoxyfen
hymexazol
meptyldinocap
bupirimate
proquinazid
captan
dodine
fenhexamid
fluazinam
benalaxyl-m
pencycuron
pyriofenone
ziram
benalaxyl
penconazole
flutolanil
benthiavalicarb
maneb
fenpyrazamine
amisulbrom
mepanipyrim
valifenalate (formerly valiphenal) 2
triadimenol
imazalil (aka enilconazole)
thiabendazole
famoxadone
flutriafol
penthiopyrad
fenamidone
propineb
tolclofos-methyl
2-phenylphenol (incl. sodium salt orthophenyl phenol) 2
a
b
c
d
e
f
g
h
i
j
k
l
m
n
o
p
q
r
s

Insecticides



Other targets

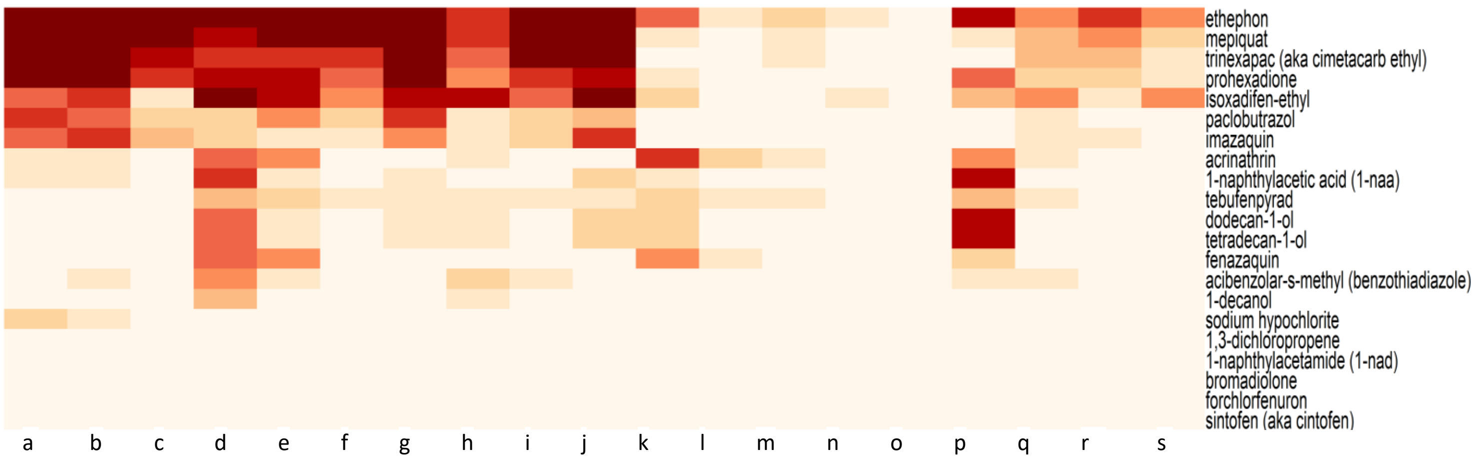


Figure S6: Heatmap of probability $\hat{\gamma}_{kj}$, that substance j is used in postcode k. Groups were obtained from a mixture models optimized by maximum likelihood with an iterative method: Expectation Maximization. Groups were ordered by similar composition of substance purchases. Substances belong to four categories: herbicides, fungicides, insecticides and other targets. Within each category of substances, substances were ordered in increasing number of groups in which they were used. Asterisks (*) highlight core substances.

Table S1: Complete list of targets associated with the “other targets” category

Targets or actions	Number of substances
Acaricide	5
Algicide	1
Attractant	2
Bactericide	1
Nematicide	1
Plant activator	1
Plant growth regulator	11
Rodenticide	2
Safener	1

Table S2 : Correspondence table of crop categories from the LPIS and aggregated crop categories used in the analyses

CATEGORY FROM LPIS	CATEGORY USED
Common wheat	Cereals
Barley	Cereals
Other cereals	Cereals
Miscellaneous	Miscellaneous
Arboriculture	Orchard
Olive trees	Orchard
Fruit Orchard	Orchard
Legume flower	Legume flower
Maize	Maize
Nut	Nut
Other oil crops	Other oil crops
Protein crops	Protein crops
Rapeseed oil	Rapeseed oil
Sunflower	Sunflower
Grapevine	Grapevine