



**HAL**  
open science

## Identifying pesticide mixtures at country-wide scale

Milena Cairo, Anne-Christine Monnet, Stéphane Robin, Emmanuelle Porcher,  
Colin Fontaine

► **To cite this version:**

Milena Cairo, Anne-Christine Monnet, Stéphane Robin, Emmanuelle Porcher, Colin Fontaine. Identifying pesticide mixtures at country-wide scale. 2023. hal-03815557v3

**HAL Id: hal-03815557**

**<https://hal.science/hal-03815557v3>**

Preprint submitted on 27 Mar 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



# Identifying pesticide mixtures at country-wide scale

Milena CAIRO<sup>1\*</sup>, Anne-Christine MONNET<sup>1</sup>, Stéphane ROBIN<sup>1,2</sup>, Emmanuelle PORCHER<sup>1</sup>, Colin FONTAINE<sup>1</sup>

<sup>1</sup> 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

<sup>2</sup> Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, F-75005 Paris, France

\*Corresponding author: Milena Cairo, [milena.cairo1@mnhn.fr](mailto:milena.cairo1@mnhn.fr)

## 1 ABSTRACT

2

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

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

14 We found that the 5,642 French postcodes can be clustered into a small number  
15 of postcode groups (ca. 20), characterized by a specific pattern of pesticide purchases,  
16 i.e. pesticide mixtures. Substances defining mixtures can be sorted into “core”  
17 substances highly probable in most postcode groups and “discriminating” substances,  
18 which are specific to and highly probable in some postcode groups only, thus playing  
19 a key role in the identity of pesticide mixtures. We found 12 core substances: two

20 insecticides (deltamethrin and lambda-cyhalothrin), six herbicides (glyphosate,  
21 diflufenican, fluroxypyr, MCPA, 2,4-d, triclopyr) and four fungicides (fludioxonil,  
22 tebuconazole, difenoconazole, thiram). The number of discriminating substances per  
23 postcode group ranged from 2 to 74. These differences in substance purchases  
24 seemed related to differences in crop composition but also potentially to regional  
25 effects.

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

32

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

35

36

## 37 INTRODUCTION

38 Since the mid-20<sup>th</sup> century, pesticides have become of common use in agriculture and  
39 their effects on both the environment and human health are a growing concern. For  
40 example, systemic pesticides are known to affect a broad range of organisms, from  
41 invertebrates, both terrestrial and aquatic, to amphibians or birds (Humann-Guillemint  
42 et al., 2019; Mahmood et al., 2016; Yang et al., 2008), thereby questioning the  
43 sustainability of agroecosystem functioning and related services (Deguines et al.,  
44 2014; Dudley et al., 2017; Furlan et al., 2018; Geiger et al., 2010). Pesticides are also  
45 identified as a concern for human health, with numerous pesticide poisonings reported  
46 across developing countries (Boedeker et al., 2020) and recent evidence of  
47 relationships between diseases such as Parkinson's or cancers and exposure to  
48 organophosphate insecticides (Sheahan et al., 2017; Tassin de Montaigu and  
49 Goulson, 2020).

50 The effect of pesticides on biodiversity are usually demonstrated with a focus on  
51 a single substance or a limited set of substances in general (e.g. thiamethoxam,  
52 clothianidin, imidacloprid, thiacloprid or glyphosate (Botías et al., 2015; Busse et al.,

53 2001; Rundlöf et al., 2015; Van Bruggen et al., 2018). Yet, wild organisms are exposed  
54 to complex mixtures (Dudley et al., 2017), owing to the diversity of substances  
55 available and used in farmlands. Hence, studying substance mixtures is considered a  
56 central task for environmental risk assessment (Lydy et al., 2004a), notably because  
57 the effects of pesticide cocktails can strongly exceed the additive effects of single  
58 compounds (Bopp et al., 2016; Junghans et al., 2006). Laboratory experiments  
59 demonstrate synergetic interactions among substances within mixtures, affecting the  
60 effect of the cocktails in non-additive ways (Cedergreen, 2014; Hernández et al., 2017;  
61 Heys et al., 2016). While the importance of studying the effects of cocktails beyond  
62 those of single substances was highlighted as soon as the late sixties (Keplinger and  
63 Deichmann, 1967), and their evaluation is mandatory in the European Union since  
64 2009 (EC No 1107/2009), few attempts to do so exist outside laboratories (Gibbons et  
65 al., 2015).

66 Studies examining the effects of substance cocktails use two approaches:  
67 bottom-up or top-down (Altenburger et al., 2013; Hernández et al., 2017; Relyea,  
68 2009). The bottom-up approach aims at testing all possible mixture compositions,  
69 starting from pairs of substances to more complex combinations. This method makes  
70 it challenging to consider more than a handful of substances. For example, ten  
71 substances represent 45 possible pairs and over a thousand possible combinations of  
72 three or more substances (Lydy et al., 2004a). Moreover, such approach might be  
73 more suited to experiments in controlled rather than natural environments, as the latter  
74 are recognized as strongly contaminated (Tang et al., 2021), making the control of  
75 mixture composition difficult. The top-down approach proposes to compare the effect  
76 of cocktails, starting from potentially frequent mixtures including a high number of  
77 substances, but at the cost of not testing all combinations. In addition, the few existing  
78 field studies generally focused on the effects of pesticide cocktails composed of a  
79 restricted number of substances, on specific crops or on restricted spatial extent,  
80 thereby limiting a broad understanding of cocktail effects (e.g. Brittain et al., 2010;  
81 Hallmann et al., 2014; Millot et al., 2017, but see Schreiner et al., 2016 & (Fritsch et  
82 al., 2022). The top-down approach makes it critical to identify relevant mixture  
83 compositions, i.e. those actually occurring in the fields. The number of actual mixtures  
84 encountered in agroecosystems should be much lower than the number of possible  
85 combinations of substances because each substance is often intended for a limited set  
86 of crops only and because agricultural production is regionally specialised on particular

87 crops. Such regional specialisation implies that existing mixtures are likely to be  
88 spatially structured. However, we still miss an overall picture of the pesticide mixture  
89 composition and its spatial structure over large spatial extents.

90

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

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

114

## 115 METHODS

### 116 *1.1 Pesticide data*

117 Data on active substances were obtained from the French national bank of  
118 pesticide sales (BNV-d; <https://bnvd.ineris.fr>). The BNV-d database registers active  
119 substances under mandatory reporting. The seller indicates the amount of each active

120 substance purchased and the postcode of the buyer in the database. This database  
121 thus indicates the quantity of active substances purchased at the spatial resolution of  
122 the postcode of the buyer. Postcode are the third level of administrative division in  
123 France, lower than the European Union NUTS3 level (administrative departments) and  
124 range from 0.17 km<sup>2</sup> to 614.39 km<sup>2</sup> in metropolitan France (median = 62.79 km<sup>2</sup>, Q1 =  
125 19.59 km<sup>2</sup>, Q3 =140.36 km<sup>2</sup>). Substances are identified with their generic name and a  
126 unique identifier, the Chemical Abstracts Service number. We modified generic names  
127 when synonyms were found. We only retained substances with a license fee (i.e. under  
128 compulsory reporting) because we can expect thorough reporting for these.

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

138 The data provides the total mass (in g) bought per substance with mandatory  
139 reporting, of which in 2017 there were 279. We analysed these quantitative data at the  
140 postcode level, assuming that substances purchased in a given postcode would be  
141 used within the same postcode or in close vicinity. Given the spatial extent of farms,  
142 pesticides may not always be spread exactly in the postcode where farmers are  
143 domiciled, but are unlikely to be used beyond the neighbouring postcodes, with one  
144 exception that we discarded. Using specific postcodes (CEDEX) that enable the  
145 identification of private companies, we discarded the data related to the national  
146 railroad company (SNCF): SNCF is a major buyer with central purchasing bodies that  
147 do not use the substances within the postcode of purchase. We converted all remaining  
148 CEDEX codes to their corresponding regular postcodes. We were thus left with 5,642  
149 postcodes with information about the quantities (in g) of 279 active substances  
150 purchased in 2017. We classified these substances into fungicides, herbicides,  
151 insecticides following the Pesticide Properties Data Base (PPDB) (Lewis et al., 2016)  
152 and the European commission pesticide database  
153 ([ec.europa.eu/food/plant/pesticides/eu-pesticides-database/active-substances](https://ec.europa.eu/food/plant/pesticides/eu-pesticides-database/active-substances)).

154 There were also 32 substances with other target groups (e.g. rodents or molluscs;  
155 Table S1 for a complete list) that we classified as “other targets”.

156 To relate the use of active substances to the area of arable land in postcodes, we  
157 extracted the total area of cropland from the 2017 French Land Parcel Identification  
158 System (LPIS, “Registre Parcellaire Graphique”, Agence de Services et de Paiements,  
159 2015). This database is a geographic information system developed under the  
160 European Council Regulation No 153/2000, for which the farmers provide annual  
161 information about their fields and crop rotation. We grouped the 16 categories of  
162 cropland types used in LPIS into 11 sub-groups (Figure S9) (Cantelaube and Carles,  
163 2010; Levavasseur et al., 2016). We summed the area of all types of cropland but  
164 meadows to obtain the total crop area per postcode.

165

## 166 1.2 Model-based Clustering

### 167 1.2.1 Input data

168

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

$$174 Y_{ij} = \log\left(\frac{\text{quantity of substance } j \text{ bought in postcode } i}{\text{cropland area of postcode } i}\right)$$

175

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

177

### 178 1.2.2 Model

179 We aimed to provide a clustering of the postcodes according to the quantity of  
180 the various substances bought. Mixture models (McLahan and Peel, 2000) provide a  
181 classical framework to achieve such a clustering. To avoid any confusion with  
182 “pesticide mixtures” we will use “Model-based Clustering” when referring to the  
183 statistical “mixture models”. The model we consider assumes that the  $n$  postcodes are  
184 spread into  $K$  groups and that the respective use of the different substances depends

185 on the group they belong to. Mixture models or model-based clustering precisely aim  
 186 at recovering this unobserved group structure from the observed data.

187

### 188 **1.2.2.1.1 Groups definition**

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

$$192 \quad \pi_k = \Pr\{Z_i = k\}. \quad (1)$$

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

195

### 196 **1.2.2.1.2 Emission distribution**

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

$$201 \quad \gamma_{kj} = \Pr\{X_{ij} = 1 | Z_i = k\}, \quad (2)$$

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

$$204 \quad (Y_{ij} | X_{ij} = 1, Z_i = k) \sim \mathcal{N}(\mu_{kj}, \sigma_{kj}^2). \quad (3)$$

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

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

$$212 \quad f_{jk}(x_{ij}, y_{ij}) = x_{ij} \gamma_{kj} \phi(y_{ij}; \mu_{kj}, \sigma_{kj}^2) + (1 - x_{ij})(1 - \gamma_{kj})$$

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



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

218

### 219 **1.2.3 Inference**

220

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

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

232

#### 233 **1.2.3.1.1 Expectation-Maximization algorithm**

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

240 The EM algorithm alternates the steps 'E' (for expectation) and 'M' (for  
241 maximization) until convergence. It can be shown that the likelihood of the data  
242  $\log p(Y; \theta)$  increases after each EM step. The reader may refer to Dempster et al.  
243 (1977) or McLahan and Peel (2000) for a formal justification of the procedure.

244

#### 245 **1.2.3.1.2 E step**

246 This step aimed at recovering the relevant information to evaluate the objective  
247 function. In the case of model-based clustering, the E steps only amounts to evaluating

248 the conditional probability  $\tau_{ik}$  for the postcode  $i$  to belong to group  $k$  given the data  
 249 observed for the postcode and the estimate of the parameter  $\theta_{ik}$  after iteration  $h - 1$ :

$$250 \quad \tau_{ik}^{(h-1)} = \Pr\{Z_i = k | \{(X_{ij}, Y_{ij})\}_{1 \leq j \leq p}; \theta^{(h-1)}\}$$

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

$$254 \quad \hat{\tau}_{ik} = \hat{\pi}_k \prod_{j=1}^p \hat{f}_{jk}(x_{ij}, y_{ij}) / (\sum_{\ell=1}^K \hat{\pi}_\ell \prod_{j=1}^p \hat{f}_{j\ell}(x_{ij}, y_{ij})).$$

255

### 256 **1.2.3.1.3 M step**

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

$$260 \quad \mathbb{E}[\log p(Y, Z; \theta) | Y; \theta^{(h)}] = \sum_{i=1}^n \sum_{k=1}^K \hat{\tau}_{ik} (\log \pi_k + \sum_{j=1}^p \log f_{kj}(x_{ij}, y_{ij})).$$

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

$$264 \quad \hat{N}_k = \sum_{i=1}^n \hat{\tau}_{ik}, \hat{M}_{kj} = \sum_{i=1}^n \hat{\tau}_{ik} x_{ij}.$$

265  $\hat{N}_k$  is the current estimate of the number of entities belonging to group  $k$ ;  $\hat{M}_{kj}$  is the  
 266 current estimate of the number of entities from group  $k$  where substance  $j$  is bought.

267 For the proportions and probability of use, we get the following updates:

$$268 \quad \hat{\pi}_k = \hat{N}_k / n, \hat{\gamma}_{kj} = \hat{M}_{kj} / \hat{N}_k.$$

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

$$270 \quad \hat{\mu}_{kj} = \frac{1}{\hat{M}_{jk}} \sum_{i=1}^n \hat{\tau}_{ik} x_{ij} y_{ij} \hat{\sigma}_{kj}^2 = \left( \frac{1}{\hat{M}_{jk}} \sum_{i=1}^n \hat{\tau}_{ik} x_{ij} y_{ij}^2 \right) - (\hat{\mu}_{kj})^2.$$

271 Similar estimates of  $\sigma_j^2$  and  $\sigma^2$  can be derived for the models with constrained  
 272 variances.

273

### 274 **1.2.4 Model selection**

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

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

280 As indicated above, the number of independent parameters is:

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

284

### 285 1.2.5 Estimated parameters

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

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

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

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

299

## 300 1.3 Analyses on estimated parameters

### 301 1.3.1 Spatial structure of postcode groups

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

311

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

321

### 322 *1.3.2 Searching for the drivers of the substance composition of groups*

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

335

### 336 *1.3.3 Test of the temporal robustness of the model-based clustering*

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

344 We used postcode probabilities to be in group  $k$  (i.e.  $\hat{t}_{ik}$ ) to compare results from  
345 the two model-based clusterings, with the 2017 groups as a reference. We compared  
346 each 2017 group with all 2015-2018 groups by calculating the proportion of postcodes  
347 in each 2017 group that belong to each 2015-2018 group. We thus obtained a matrix  
348 with the percentage of postcodes from 2017 groups that were found in the various  
349 2015-2018 groups (Gelbard et al., 2007).

350

## 351 RESULTS

352

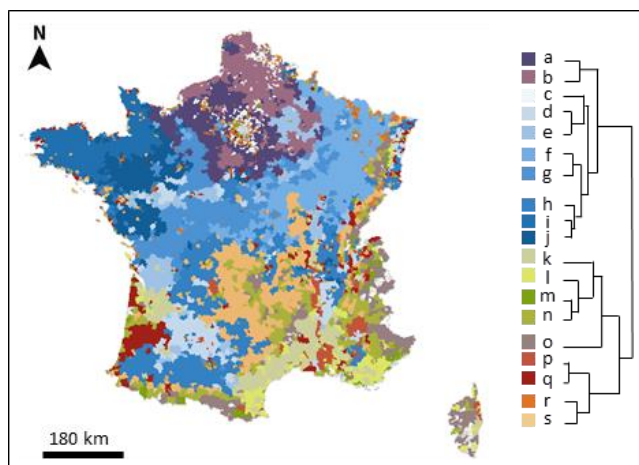
### 353 1.4 *The model-based clustering yields a small number of groups of postcodes*

354 The model-based clustering with unconstrained variances had the highest BIC  
355 and classified the 5,642 postcodes into 19 groups on the basis of 2017 purchase data  
356 for 279 active substances (Figure S2). Most postcodes were unambiguously attributed  
357 to a single of these groups, as shown by the bimodal distribution of the probability for  
358 a postcode  $i$  to belong to group  $k$ , with most values close to 0 or 1 (Figure S3). Only  
359 13 out of 5,642 postcodes had a maximum probability to be in a group lower than 0.7.

360

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

371



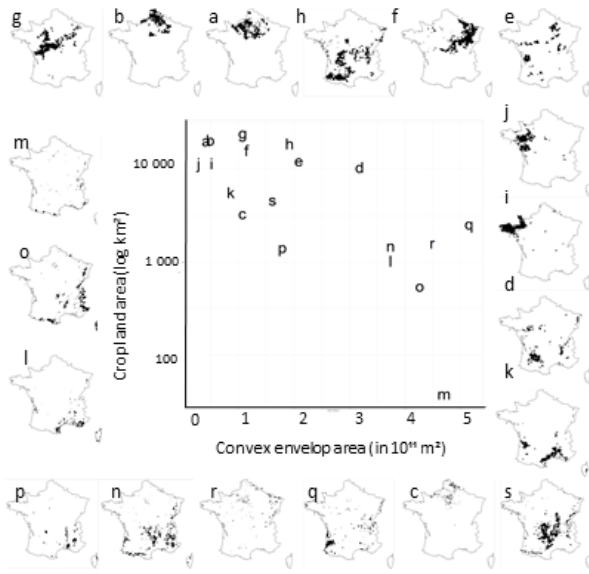
372

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

377

378 Postcode groups corresponded to specific geographical and/or agricultural  
 379 regions. For example, group *i* corresponded mostly to Brittany (the western peninsula)  
 380 and group *b* was predominantly located in Northern France. Groups *e* and *d* were more  
 381 scattered across the country but overlapped almost perfectly with wine regions (*Figure*  
 382 *2*). Note that a couple of groups were composed of a limited number of postcodes  
 383 spatially scattered across France (e.g. groups *m* and *o* *Figure 2*). In particular, group  
 384 *m* represented less than 39 km<sup>2</sup> of cropland and is generally discarded in the following.

385 The groups identified by the model-based clustering were relatively robust to a  
 386 change in the temporal range of the data, as shown by the results of the clustering on  
 387 the 2015-2018 data (*Figure S7*). This second clustering yielded 24 groups and the  
 388 percentage of shared postcodes between the 2017 groups and their most similar 2015-  
 389 2018 groups varied between 41% and 80% (median = 62%, Q1 = 53%, Q3 = 66%).  
 390 For example, groups in Normandie (group *a* vs. group 15) or part of the Languedoc  
 391 region (group *k* vs. 10) were stable over time (*Figure S7*). The higher number of groups  
 392 obtained with the 2015-2018 model-based clustering (24 vs. 19) was often due to the  
 393 split of some 2017 groups into two 2015-2018 groups. For example, for 2017 group *i*,  
 394 there was 53% similarity with 2015-2018 group 16 and 40% similarity with group 20  
 395 (*Figure S7*). Because of this temporal consistency in the clustering, we only present in  
 396 the following the analyses on the 2017 dataset, which is thought to be more accurate  
 397 (see 1.1).



398

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

406

### 407 1.5 Substance composition of postcode groups: core and discriminating substances

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

421





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

429

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

444 Discriminating substances are defined as substances with medium to high mean  
445 probability of purchase, mechanically associated with a large variance across groups  
446 in this probability (Figure S5). Because of their contrasting probability of purchase  
447 across groups, discriminating substances were likely to contribute greatly to the  
448 formation of groups. We used the arbitrary range of average probabilities from 0.5 to  
449 0.85 to define discriminating substances. Using these thresholds, we found a set of 84  
450 discriminating substances, including 45 herbicides, 25 fungicides, 10 insecticides and  
451 4 with other targets (Supplementary information 2). In the following, we focus on  
452 discriminating substances that are highly probable ( $\hat{\gamma}_{kj} > 0.85$ ) in at least one postcode  
453 group, i.e. substances that are likely major components of pesticide mixtures occurring  
454 in a given group. We found seven widespread discriminating substances purchased  
455 with a probability higher than 0.85 in at least 12 out of 19 groups: azoxystrobin,  
456 boscalid, cypermethrin, mesotrione, metsulfuron-methyl, pendimethalin and  
457 prothioconazole. These substances are very close to core substances. Conversely,

458 four substances were highly specific, being purchased with high probability ( $> 0.85$ ) in  
459 less than four groups (e.g. metribuzin in groups *d* and *b*). Within a group, the number  
460 of discriminating substances with high probability of purchase ( $> 0.85$ ) varied strongly  
461 among groups, from 2 for group *r* to 80 for group *g* (mean =  $43 \pm 27$ ). This cross-group  
462 variation in the number of highly probable discriminating substances has implication  
463 for the composition and complexity of pesticide mixtures in French agroecosystems:  
464 from relatively “simple” (12 core substances and 11 discriminating substances in group  
465 *q*) to highly complex (12 core substances and 74 discriminating substances in group  
466 *g*).

467

468 The 156 remaining substances, with a low average probability to be purchased  
469 ( $< 0.5$ ), also had a role in group identification, but were seldom purchased and will not  
470 be described further (Figure 3).

471

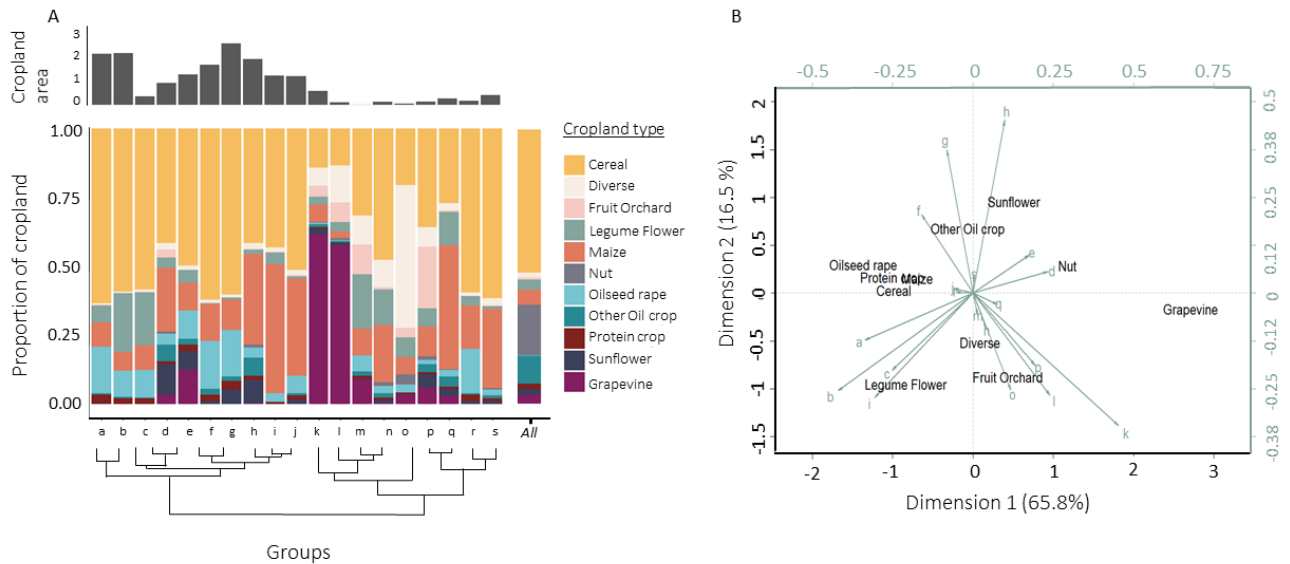
472 *1.6 Postcode groups differ in terms of crop composition, but active substance purchase may*  
473 *not be solely driven by crop identity*

474

475 Groups of postcodes, which by construction are composed of different mixtures  
476 of substances, also differed in terms of proportions of cropland grown with various  
477 crops, such that groups with close pesticide composition sometimes, but not always,  
478 also exhibited similar crop usage (Figure 4). The possible relations between pesticide  
479 composition and crop composition can be visualized either on Figure 4, where crop  
480 composition of groups similar in terms of pesticides purchases are plotted next to each  
481 other, or on the biplot of the log ratio analysis (Figure 5), in which groups with similar  
482 crop composition are plotted next to each other. For example, groups *k* and *l*,  
483 characterized by a large proportion of vineyards, were close to each other both in the  
484 log-ratio analysis, which is indicative of similar crop compositions (Figure 5) and in the  
485 hierarchical clustering, which is indicative of similar pesticide purchases (Figure 4).  
486 The same was true for groups *b*, *c* and *i*, and, to a lesser extent, *a*, characterized by  
487 an appreciable proposition of crops from the legume/flower category. However, some  
488 groups such as *h* and *g* were different in terms of substances (not in the same sub-  
489 group, Figure 4) while exhibiting comparable proportions of crop types (Figure 4).  
490 Alternatively, some groups that were closely related in terms of substance purchases,  
491 such as groups *i* and *h*, could be characterized by dissimilar crop compositions. The

492 latter patterns may suggest regionalisation of substance use, such that neighbouring  
 493 regions tend to use similar products or substances even with variations in crops grown  
 494 (e.g. *i* and *h*).

495



496  
 497 *Figure 4: A. Distribution of crop type area across groups. The top grey histogram shows the*  
 498 *distribution of total cropland area across groups (in 10<sup>4</sup> km<sup>2</sup>). The dendrogram was obtained*  
 499 *using an agglomerative hierarchical clustering on the basis of Ward's method among groups*  
 500 *(see 2.2.1). B. Biplot of the log ratio analysis relating the proportion of crop types in each group.*  
 501 *Only groups identified as spatially coherent are displayed (see 3.2). For readability, the groups*  
 502 *and crop types are displayed on two different scales: black for crop types, green for groups. The*  
 503 *size of arrows corresponds to the contribution of each group. Groups that appear close to each*  
 504 *other on the biplot have similar crop composition, which can be inferred from the contribution*  
 505 *of each crop type to the axes.*

506

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

515

## 516 DISCUSSION

517

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

532

533 *1.7 Significance of the identification of highly probable active substances, and of mixtures of*  
534 *active substances characteristic of postcode groups, for the study of the impacts of*  
535 *pesticides in the environment*

536

537 The identification of active substances that are purchased with high probability in  
538 all (core substances) or a subset (discriminating substances) of postcode groups might  
539 contribute to reducing the potential street light effect, whereby most research efforts  
540 focus on molecules that are either easy to study (Hendrix, 2017) or that were  
541 popularized by previous studies (Tsvetkov and Zayed, 2021). Unsurprisingly, most  
542 core substances identified here are already well-known, widely-used substances.  
543 Glyphosate is the most widely used broad-spectrum herbicide (Jatinder Pal Kaur Gill  
544 et al. 2017; Myers et al. 2016), with associated concerns regarding pervasive direct  
545 and indirect effects (Van Bruggen et al., 2018). Tebuconazole and difenoconazole, two  
546 triazole fungicides, are widely used and studied (Zubrod et al., 2019). Deltamethrin and  
547 lambda-cyhalothrin, two pyrethroids impacting nervous systems (Ray and Fry, 2006;  
548 Soderlund and Bloomquist, 1989), are known to have adverse effects on a large range  
549 of non-target species such as fish, birds and amphibians (Ali et al. 2011). Yet, a

550 preliminary literature search on these 12 core substances suggests that the research  
551 effort on their adverse effects on biodiversity is still highly variable. For core herbicides,  
552 a simple search of the molecule name together with “biodiversity” or “ecotoxicology” in  
553 the abstract of articles on ISI Web of Science yields more than two hundred research  
554 articles for glyphosate and around seventy for 2,4-d, but only 2 to 17 articles for  
555 diflufenican, fluroxypyr, MCPA, triclopyr and pendimethalin. For core insecticides, the  
556 same search returns ca. 40 articles for lambda-cyhalothrin and deltamethrin. The four  
557 core fungicides were no exception, with a number of research articles below ten for  
558 thiram, fludioxonil and difenoconazole and around thirty for tebuconazole. Ultimately,  
559 our method eases the bottom-up approach in the laboratory by providing a selection  
560 of understudied substances deserving further attention.

561 Studying all possible (combinations of) substances is prohibitive (Wolska et al.,  
562 2007); beyond the identification of single substances, our approach chiefly contributes  
563 to identifying combinations of active substances that are likely to be encountered in  
564 farmland environments, i.e. pesticide mixtures. The model-based clustering identified  
565 a relatively small number of postcode groups (19 to 24 depending on the temporal  
566 coverage of pesticide data). Each group is characterized by a specific combination of  
567 purchases of active substances and can be interpreted as a potential mixture of  
568 pesticides occurring in the location of the postcodes, under the assumption that all  
569 purchased substances are used within the buying area during the year of purchase  
570 (see “Limitations and perspectives” below). Among the 279 active substances  
571 considered in these analyses, we highlighted the core substances included in most  
572 mixtures and the discriminating substances specific to particular mixtures. Within each  
573 postcode group, both types of substances might be a good starting shortlist of  
574 substances within which one can investigate potential interactive effects on  
575 biodiversity. Indeed, these substances are purchased with high probability in at least  
576 some large groups of postcodes, hence are potentially part of widespread mixtures.  
577 Although this list is much shorter than the total list of authorized active substances, it  
578 still contains 12 core substances, plus 2 to 80 discriminating substances depending on  
579 the postcode group. Since our approach to identifying core and discriminating  
580 substances was based on probability of purchase only, this shortlist of substances  
581 could be narrowed down further by selecting active substances bought in large  
582 quantities (see also “Limitations and perspectives”) or with high toxicity. The  
583 appreciable number of core and discriminating substances composing mixtures is

584 anyway consistent with surveys showing that active substances are rarely found alone  
585 in the environment (Silva et al., 2019). It also further substantiates the need for a  
586 broader assessment of the synergistic effects of pesticides on biodiversity, often  
587 completed on a limited set of substances only (Schreiner et al., 2016; Silva et al.,  
588 2019). For core substances, for example, some cocktail effects have already been  
589 studied but mostly on pairs of substances (Brodeur et al., 2014; Peluso et al., 2022)  
590 and more rarely for cocktails of three or more substances (Cedergreen, 2014; Glinski  
591 et al., 2018; Van Meter et al., 2018). Focusing on the reasonable number of relatively  
592 complex mixtures identified by the present approach would contribute to improve our  
593 understanding of the synergistic effects of realistic cocktails on organisms.

594

595

## 596 *1.8 Limitations & perspectives*

### 597 *1.8.1 Limited spatio-temporal resolution of the BNV-d data*

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

617 identified (glyphosate and tebuconazole) or to discriminant substances (boscalid and  
618 epoxiconazole) with a high probability of being used over half of the postcode groups.

619

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

636

### 637 *1.8.2 Going beyond the use of purchase probabilities and arbitrary thresholds to identify the* 638 *substances of interest for risk assessment*

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

648 In addition, most of our interpretation of pesticide mixture composition relies on  
649 the estimated purchase probabilities, but these mixtures were also identified using

650 information on the mean and variance of purchased amounts within postcodes, hence  
651 mixtures differ for these variables as well. For example, glyphosate, a core substance  
652 with high purchase probability in all postcode groups, was bought in contrasting  
653 quantities across postcode groups: the average amount was 53.9 kg/km<sup>2</sup> and ranged  
654 from 7.8 kg/km<sup>2</sup> in group *p* to 146 kg/km<sup>2</sup> in group *i*. Although the purchase probability  
655 was positively correlated to the mean purchased quantity and negatively to its  
656 variance, the correlation is not strong, and further analysis is needed to fully uncover  
657 variation in substance quantities within the mixtures we identified.

658

### 659 *1.8.3 Taking into account the yearly variation in pesticide use*

660 Our analysis appeared relatively robust to the time period of the pesticide  
661 purchase data, as suggested by the comparison of postcode groups obtained with the  
662 2017 and the 2015-2018 datasets. This strong correlation between the 2017 and the  
663 2015-2018 analysis is not entirely surprising because of the presence of the 2017 data  
664 in both analyses. Yet, adding three years of data into the analysis did not affect much  
665 the composition of postcode groups, which suggests relatively stable patterns of  
666 pesticide purchase in France over a short time period. Nonetheless, we observed  
667 some differences, mainly due to the split of some groups, which were also expected  
668 due to climatic variation, changes in legislation on pesticide use (Urruty et al., 2016) or  
669 changes in crop areas (Levavasseur et al., 2016). A better integration of the temporal  
670 dynamics of pesticide purchases in the characterisation of pesticide mixtures is needed  
671 if we are to monitor pesticide mixtures across France. This can be achieved by applying  
672 the model-based clustering to each year of data separately. Investigating the spatial  
673 stability of groups and mixture compositions across years would contribute to either  
674 estimate annual mixtures or to find temporarily stable mixtures. Finding recurrent  
675 mixtures could facilitate risk assessment over years. Indeed, this could provide key  
676 information on the frequency of mixtures encountered by organisms as repeated  
677 contact might increase risks (Stuligross and Williams, 2021).

678

679



680 1.9 *Postcode groups are related to the crop they grow, as well as to other regional factors,*  
681 *but the underlying mechanisms remain to be fully identified*

682 Although no spatial information was included in the model-based clustering  
683 analysis, the postcode groups exhibited a strong spatial structure, in which most  
684 groups are strongly aggregated and only a few small groups are scattered across  
685 France. Such spatial structure was expected since pesticide use is strongly crop-  
686 dependent. For example, acetamiprid, a substance used to protect fruit trees or  
687 grapevine against aphids, is bought with high probability in groups *l, e* and *d*, with high  
688 proportion of fruit orchards and grapevines. Similarly, cyproconazole, a substance with  
689 a broader spectrum of use, is bought with high probability in several groups with  
690 contrasting crop compositions (*a, b, e, f, g, h, j, k, l, n, o, q, r* Figure 4). However,  
691 deviations from this pattern were found: some adjacent postcode groups can have  
692 different sets of crops but similar substance purchases or some spatially distant  
693 postcode groups can have similar sets of crops but different substance purchases.  
694 This observation suggests that local conditions, such as climate or pests, or some  
695 regional patterns in the pesticide market and/or distribution, can drive the purchase of  
696 active substances more than the set of crops grown (Silva et al., 2019; Storck et al.,  
697 2017). Hence, the differences among postcode groups were related to a combination  
698 of crop identity effects and other regional effects that will need additional analysis to  
699 be identified. A straightforward perspective for the model-based clustering approach  
700 would thus be to incorporate environmental covariates in the model, and evaluate how  
701 clusters are modified.

702

## 703 CONCLUSION

704

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

## 713 Acknowledgement

714

715 This project was funded and supported by ANSES (grant agreement 2019-CRB-  
716 03\_PV19) via the tax on sales of plant protection products. The proceeds of this tax  
717 are assigned to ANSES to finance the establishment of the system for monitoring the  
718 adverse effects of plant protection products, called ‘phytopharmacovigilance’ (PPV),  
719 established by the French Act on the future of agriculture of 13 October 2014. We wish  
720 to thank the steering committee of the project: Fabrizio Botta, Sandrine Charles, Marc  
721 Girondot, Olivier Le Gall, Thomas Quintaine, and Lynda Saibi-Yedjer. Milena Cairo  
722 was supported by ANR project VITIBIRD (ANR-20-CE34-0008) while working on this  
723 project. This work also benefitted from the support of the project ECONET (ANR-18-  
724 CE02-0010) and of the “Chaire Modélisation Mathématique et Biodiversité”.

725

## 726 Conflict of interest

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

728

## 729 SUPPLEMENTARY MATERIALS

730

731 Supplementary materials to this article can be found online at

732 <https://doi.org/10.5281/zenodo.7693149>

733

734

## 735 REFERENCES

736

737 Ali, S. F., Shieh, B. H., Alehaideb, Z., Khan, M. Z., Louie, A., Fageh, N., & Law, F. C.  
738 (2011). A review on the effects of some selected pyrethroids and related  
739 agrochemicals on aquatic vertebrate biodiversity. *Canadian Journal of Pure &*  
740 *Applied Sciences*, 5(2), 1455-1464.

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

744 Boedeker, W., Watts, M., Clausing, P., Marquez, E., 2020. The global distribution of  
745 acute unintentional pesticide poisoning: estimations based on a systematic  
746 review. *BMC Public Health* 20, 1–19. [https://doi.org/10.1186/s12889-020-09939-](https://doi.org/10.1186/s12889-020-09939-0)  
747 0

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

749 Richarz, A.-N., Triebe, J., Worth, A., 2016. Review of case studies on the human  
750 and environmental risk assessment of chemical mixtures.  
751 <https://doi.org/10.2788/272583>

752 Botías, C., David, A., Horwood, J., Abdul-Sada, A., Nicholls, E., Hill, E., Goulson, D.,  
753 2015. Neonicotinoid Residues in Wildflowers, a Potential Route of Chronic  
754 Exposure for Bees. *Environ. Sci. Technol.* 49, 12731–12740.  
755 <https://doi.org/10.1021/acs.est.5b03459>

756 Brittain, C.A., Vighi, M., Bommarco, R., Settele, J., Potts, S.G., 2010. Impacts of a  
757 pesticide on pollinator species richness at different spatial scales. *Basic Appl.*  
758 *Ecol.* 11, 106–115. <https://doi.org/10.1016/j.baae.2009.11.007>

759 Brodeur, J.C., Poliserpi, M.B., D’Andrea, M.F., Sánchez, M., 2014. Synergy between  
760 glyphosate- and cypermethrin-based pesticides during acute exposures in  
761 tadpoles of the common South American Toad *Rhinella arenarum*.  
762 *Chemosphere* 112, 70–76. <https://doi.org/10.1016/j.chemosphere.2014.02.065>

763 Busse, M.D., Ratcliff, A.W., Shestak, C.J., Powers, R.F., 2001. Glyphosate toxicity  
764 and the effects of long-term vegetation control on soil microbial communities.  
765 *Soil Biol. Biochem.* 33, 1777–1789. [https://doi.org/10.1016/S0038-](https://doi.org/10.1016/S0038-0717(01)00103-1)  
766 [0717\(01\)00103-1](https://doi.org/10.1016/S0038-0717(01)00103-1)

767 Cantelaube, P., Carles, M., 2010. Le registre parcellaire graphique : des donn é es g  
768 é ographiques pour d é crire la couverture du sol agricole.

769 Cedergreen, N., 2014. Quantifying synergy: A systematic review of mixture toxicity  
770 studies within environmental toxicology. *PLoS One* 9.  
771 <https://doi.org/10.1371/journal.pone.0096580>

772 Deguines, N., Jono, C., Baude, M., Henry, M., Julliard, R., Fontaine, C., 2014. Large-  
773 scale trade-off between agricultural intensification and crop pollination services.  
774 *Front. Ecol. Environ.* 12, 212–217. <https://doi.org/10.1890/130054>

775 Dempster;A.P, Laird, N., Rubin, D., 1977. Maximum Likelihood from Incomplete  
776 data via the EM Algorithm.

777 Dudley, N., Attwood, S.J., Goulson, D., Jarvis, D., Bharucha, Z.P., Pretty, J., 2017.  
778 How should conservationists respond to pesticides as a driver of biodiversity  
779 loss in agroecosystems? *Biol. Conserv.* 209, 449–453.  
780 <https://doi.org/10.1016/j.biocon.2017.03.012>

781 Fritsch, C., Appenzeller, B., Burkart, L., Coeurdassier, M., Scheifler, R., Raoul, F.,  
782 Driget, V., Powolny, T., Gagnaison, C., Rieffel, D., Afonso, E., Goydadin, A.C.,  
783 Hardy, E.M., Palazzi, P., Schaeffer, C., Gaba, S., Bretagnolle, V., Bertrand, C.,  
784 2022. Pervasive exposure of wild small mammals to legacy and currently used  
785 pesticide mixtures in arable landscapes. *Sci. Rep.* 1–22.  
786 <https://doi.org/10.1038/s41598-022-19959-y>

787 Furlan, L., Pozzebon, A., Duso, C., Simon-Delso, N., Sánchez-Bayo, F., Marchand,  
788 P.A., Codato, F., Bijleveld van Lexmond, M., Bonmatin, J.M., 2018. An update of  
789 the Worldwide Integrated Assessment (WIA) on systemic insecticides. Part 3:  
790 alternatives to systemic insecticides. *Environ. Sci. Pollut. Res.* 1–23.  
791 <https://doi.org/10.1007/s11356-017-1052-5>

792 Geiger, F., Bengtsson, J., Berendse, F., Weisser, W.W., Emmerson, M., Morales,  
793 M.B., Ceryngier, P., Liira, J., Tscharntke, T., Winqvist, C., Eggers, S.,  
794 Bommarco, R., Pärt, T., Bretagnolle, V., Plantegenest, M., Clement, L.W.,  
795 Dennis, C., Palmer, C., Oñate, J.J., Guerrero, I., Hawro, V., Aavik, T., Thies, C.,  
796 Flohre, A., Hänke, S., Fischer, C., Goedhart, P.W., Inchausti, P., 2010.  
797 Persistent negative effects of pesticides on biodiversity and biological control  
798 potential on European farmland. *Basic Appl. Ecol.* 11, 97–105.

799 <https://doi.org/10.1016/j.baae.2009.12.001>

800 Gelbard, R., Goldman, O., Spiegler, I., 2007. Investigating diversity of clustering  
801 methods: An empirical comparison. *Data Knowl. Eng.* 63, 155–166.  
802 <https://doi.org/10.1016/j.datak.2007.01.002>

803 Gibbons, D., Morrissey, C., Mineau, P., 2015. A review of the direct and indirect  
804 effects of neonicotinoids and fipronil on vertebrate wildlife. *Environ. Sci. Pollut.*  
805 *Res.* 22, 103–118. <https://doi.org/10.1007/s11356-014-3180-5>

806 Glinski, D.A., Purucker, S.T., Van Meter, R.J., Black, M.C., Henderson, W.M., 2018.  
807 Endogenous and exogenous biomarker analysis in terrestrial phase amphibians  
808 (*Lithobates sphenoccephala*) following dermal exposure to pesticide mixtures.  
809 *Env. chem* 60, 1–24. <https://doi.org/10.1071/EN18163>.

810 Greenacre, M., 2019. Variable Selection in Compositional Data Analysis Using  
811 Pairwise Logratios. *Math. Geosci.* 51, 649–682. [https://doi.org/10.1007/s11004-](https://doi.org/10.1007/s11004-018-9754-x)  
812 [018-9754-x](https://doi.org/10.1007/s11004-018-9754-x)

813 Hallmann, C.A., Foppen, R.P.B., Van Turnhout, C.A.M., De Kroon, H., Jongejans, E.,  
814 2014. Declines in insectivorous birds are associated with high neonicotinoid  
815 concentrations. *Nature* 511, 341–343. <https://doi.org/10.1038/nature13531>

816 Hendrix, C.S., 2017. The streetlight effect in climate change research on Africa. *Glob.*  
817 *Environ. Chang.* 43, 137–147. <https://doi.org/10.1016/j.gloenvcha.2017.01.009>

818 Hernández, A.F., Gil, F., Lacasaña, M., 2017. Toxicological interactions of pesticide  
819 mixtures: an update. *Arch. Toxicol.* 91, 3211–3223.  
820 <https://doi.org/10.1007/s00204-017-2043-5>

821 Heys, K.A., Shore, R.F., Pereira, M.G., Jones, K.C., Martin, F.L., 2016. Risk  
822 assessment of environmental mixture effects. *RSC Adv.* 6, 47844–47857.  
823 <https://doi.org/10.1039/c6ra05406d>

824 Humann-Guillemint, Ségolène, Binkowski, Ł.J., Jenni, L., Hilke, G., Glauser, G.,  
825 Helfenstein, F., 2019. A nation-wide survey of neonicotinoid insecticides in  
826 agricultural land with implications for agri-environment schemes. *J. Appl. Ecol.*  
827 56, 1502–1514. <https://doi.org/10.1111/1365-2664.13392>

828 Humann-Guillemint, S., Tassin de Montaigu, C., Sire, J., Grünig, S., Gning, O.,  
829 Glauser, G., Vallat, A., Helfenstein, F., 2019. A sublethal dose of the  
830 neonicotinoid insecticide acetamiprid reduces sperm density in a songbird.  
831 *Environ. Res.* 177, 108589. <https://doi.org/10.1016/j.envres.2019.108589>

832 Junghans, M., Backhaus, T., Faust, M., Scholze, M., Grimme, L.H., 2006. Application  
833 and validation of approaches for the predictive hazard assessment of realistic  
834 pesticide mixtures. *Aquat. Toxicol.* 76, 93–110.  
835 <https://doi.org/10.1016/j.aquatox.2005.10.001>

836 Keplinger, M.L., Deichmann, W.B., 1967. Acute toxicity of combinations of pesticides.  
837 *Toxicol. Appl. Pharmacol.* 10, 586–595. [https://doi.org/10.1016/0041-](https://doi.org/10.1016/0041-008X(67)90097-X)  
838 [008X\(67\)90097-X](https://doi.org/10.1016/0041-008X(67)90097-X)

839 Levavasseur, F., Martin, P., Bouty, C., Barbottin, A., Bretagnolle, V., Théron, O.,  
840 Scheurer, O., Piskiewicz, N., 2016. RPG Explorer: A new tool to ease the  
841 analysis of agricultural landscape dynamics with the Land Parcel Identification  
842 System. *Comput. Electron. Agric.* 127, 541–552.  
843 <https://doi.org/10.1016/j.compag.2016.07.015>

844 Lewis, K.A., Tzilivakis, J., Warner, D.J., Green, A., 2016. An international database  
845 for pesticide risk assessments and management. *Hum. Ecol. risk Assess.* 22,  
846 1050–1064. <https://doi.org/10.1017/CBO9781107415324.004>

847 Lopez-Antia, A., Ortiz-Santaliestra, M.E., Mougeot, F., Mateo, R., 2015. Imidacloprid-  
848 treated seed ingestion has lethal effect on adult partridges and reduces both

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

851 Lydy, M., Belden, J., Wheelock, C., Hammock, B., Denton, D., 2004a. Challenges in  
852 regulating pesticide mixtures. *Ecol. Soc.* 9. [https://doi.org/10.5751/ES-00694-](https://doi.org/10.5751/ES-00694-090601)  
853 [090601](https://doi.org/10.5751/ES-00694-090601)

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

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

859 Millot, F., Decors, A., Mastain, O., Quintaine, T., Berny, P., Vey, D., Lasseur, R., Bro,  
860 E., 2017. Field evidence of bird poisonings by imidacloprid-treated seeds: a  
861 review of incidents reported by the French SAGIR network from 1995 to 2014.  
862 *Environ. Sci. Pollut. Res.* 24, 5469–5485. [https://doi.org/10.1007/s11356-016-](https://doi.org/10.1007/s11356-016-8272-y)  
863 [8272-y](https://doi.org/10.1007/s11356-016-8272-y)

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

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

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

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

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

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

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

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

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

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

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

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

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

899 Toxicity assessment on honey bee larvae of a repeated exposition of a systemic  
900 fungicide, boscalid. *Bull. Insectology* 70, 83–90.

901 Soderlund, D.M., Bloomquist, J.R., 1989. Neurotoxic actions of pyrethroid  
902 insecticides. *Annu. Rev. Entomol.* 34, 77–96.  
903 <https://doi.org/10.1146/annurev.en.34.010189.000453>

904 Storck, V., Karpouzias, D.G., Martin-Laurent, F., 2017. Towards a better pesticide  
905 policy for the European Union. *Sci. Total Environ.* 575, 1027–1033.  
906 <https://doi.org/10.1016/j.scitotenv.2016.09.167>

907 Stuligross, C., Williams, N.M., 2021. Past insecticide exposure reduces bee  
908 reproduction and population growth rate. *Proc. Natl. Acad. Sci. U. S. A.* 118, 1–  
909 6. <https://doi.org/10.1073/pnas.2109909118>

910 Tang, F.H.M., Lenzen, M., McBratney, A., Maggi, F., 2021. Risk of pesticide pollution  
911 at the global scale. *Nat. Geosci.* 14, 206–210. [https://doi.org/10.1038/s41561-](https://doi.org/10.1038/s41561-021-00712-5)  
912 [021-00712-5](https://doi.org/10.1038/s41561-021-00712-5)

913 Tassinde Montaigu, C., Goulson, D., 2020. Identifying agricultural pesticides that may  
914 pose a risk for birds. *PeerJ*.

915 Tsvetkov, N., Zayed, A., 2021. Searching beyond the streetlight: Neonicotinoid  
916 exposure alters the neurogenomic state of worker honey bees. *Ecol. Evol.* 11,  
917 18733–18742. <https://doi.org/10.1002/ece3.8480>

918 Urruty, N., Deveaud, T., Guyomard, H., Boiffin, J., 2016. Impacts of agricultural land  
919 use changes on pesticide use in French agriculture. *Eur. J. Agron.* 80, 113–123.  
920 <https://doi.org/10.1016/j.eja.2016.07.004>

921 Van Bruggen, A.H.C., He, M.M., Shin, K., Mai, V., Jeong, K.C., Finckh, M.R., Morris,  
922 J.G., 2018. Environmental and health effects of the herbicide glyphosate. *Sci.*  
923 *Total Environ.* 616–617, 255–268.  
924 <https://doi.org/10.1016/j.scitotenv.2017.10.309>

925 Van Meter, R.J., Glinski, D.A., Purucker, S.T., Henderson, W.M., 2018. Influence of  
926 exposure to pesticide mixtures on the metabolomic profile in post-metamorphic  
927 green frogs (*Lithobates clamitans*). *Sci. Total Environ.* 624, 1348–1359.  
928 <https://doi.org/10.1016/j.scitotenv.2017.12.175>

929 Wolska, L., Sagajdakow, A., Kuczyńska, A., Namieśnik, J., 2007. Application of  
930 ecotoxicological studies in integrated environmental monitoring: Possibilities and  
931 problems. *TrAC - Trends Anal. Chem.* 26, 332–344.  
932 <https://doi.org/10.1016/j.trac.2006.11.012>

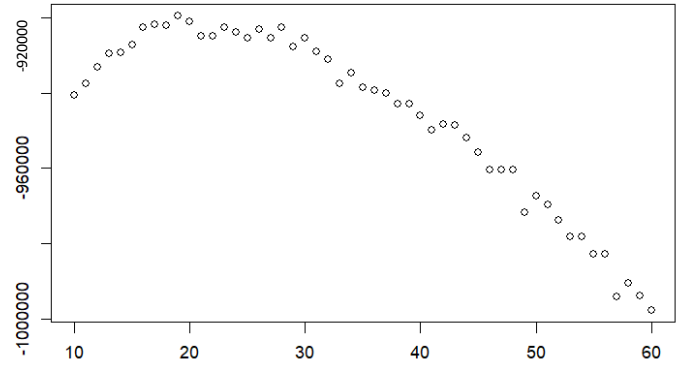
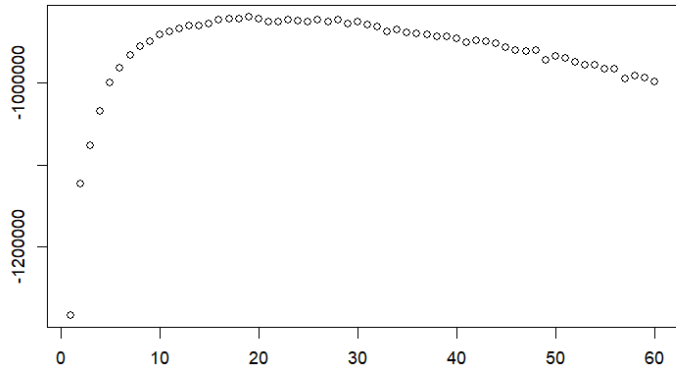
933 Yang, E.C., Chuang, Y.C., Chen, Y.L., Chang, L.H., 2008. Abnormal foraging  
934 behavior induced by sublethal dosage of imidacloprid in the honey bee  
935 (*Hymenoptera: Apidae*). *J. Econ. Entomol.* 101, 1743–1748.  
936 <https://doi.org/10.1603/0022-0493-101.6.1743>

937 Zubrod, J.P., Bundschuh, M., Arts, G., Brühl, C.A., Imfeld, G., Knäbel, A.,  
938 Payraudeau, S., Rasmussen, J.J., Rohr, J., Scharmüller, A., Smalling, K.,  
939 Stehle, S., Schulz, R., Schäfer, R.B., 2019. Fungicides: An Overlooked Pesticide  
940 Class? *Environ. Sci. Technol.* 53, 3347–3365.  
941 <https://doi.org/10.1021/acs.est.8b04392>

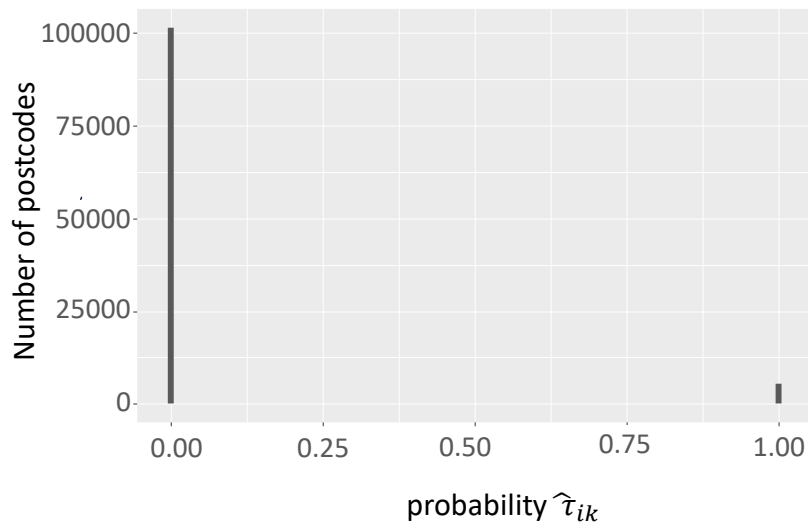
942

# APPENDIX

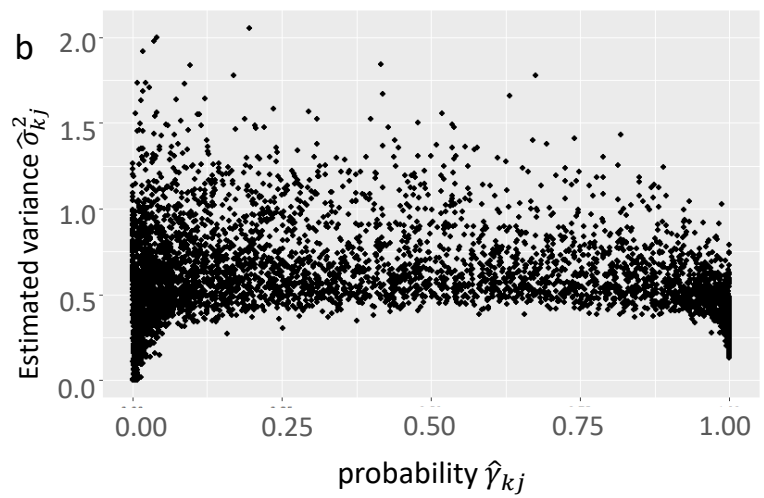
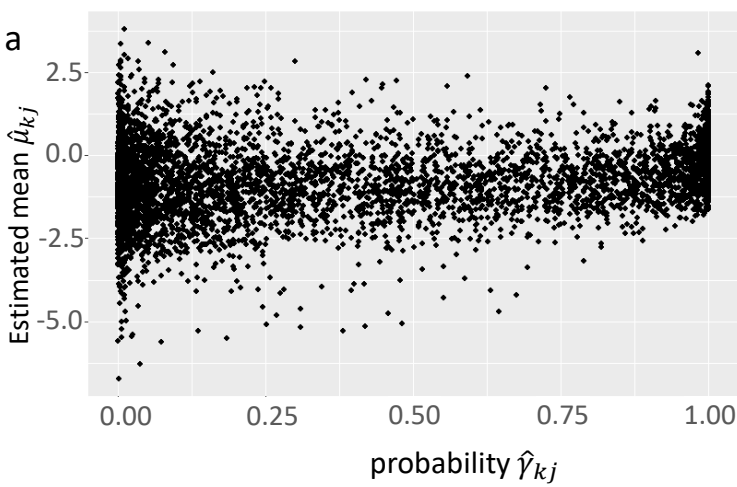
Bayesian Information Criterion



**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

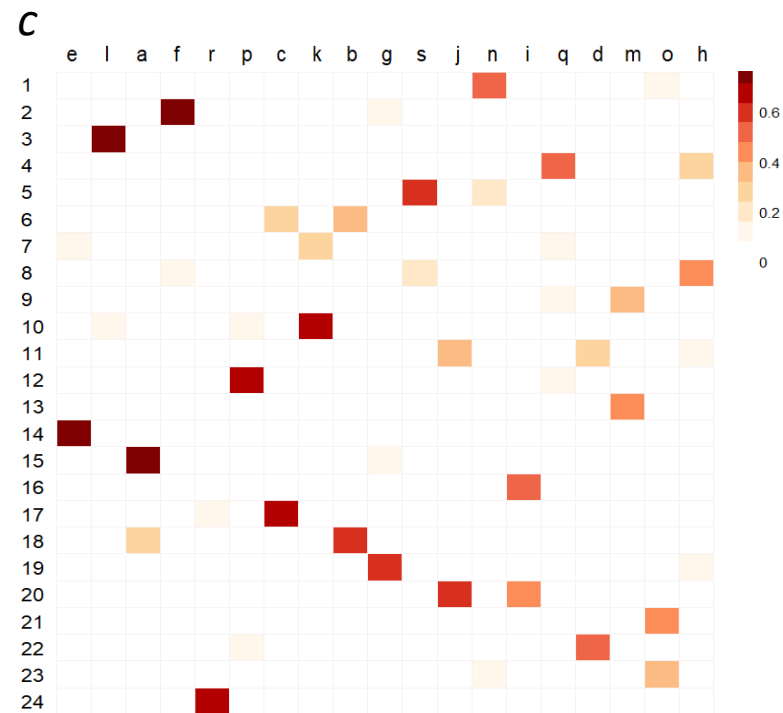
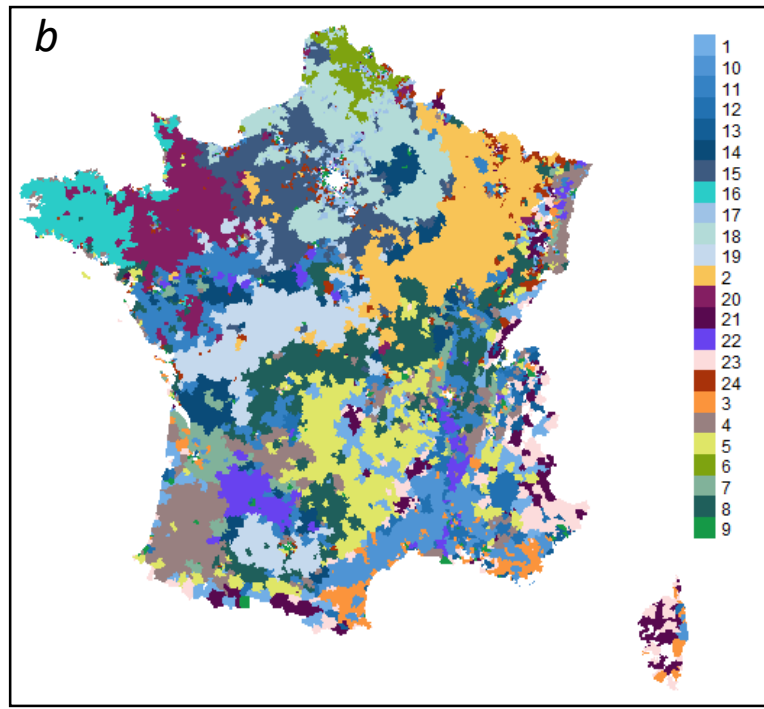
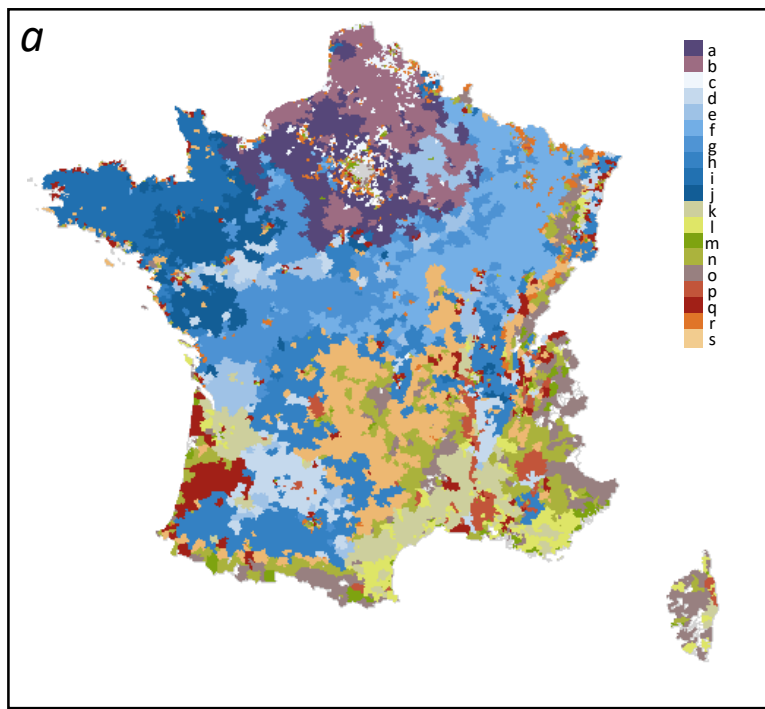


**Figure S2:** Distribution of  $\hat{\tau}_{ik}$ , the probability of postcode  $i$  to be in group  $k$



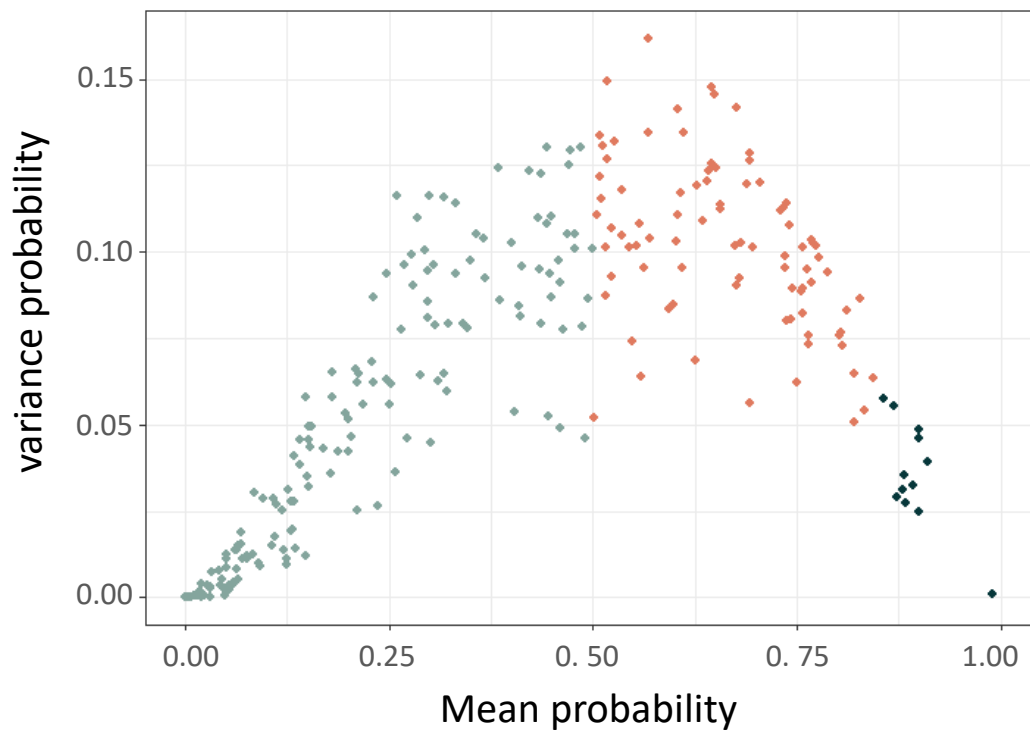
**Figure S3:** Estimated mean ( $\hat{\mu}_{kj}$ , panel a) and variance  $\hat{\sigma}_{kj}^2$ , panel b) of substance quantities purchased in a group as a function of the probability of a substance to be in a group  $\hat{\gamma}_{kj}$ .





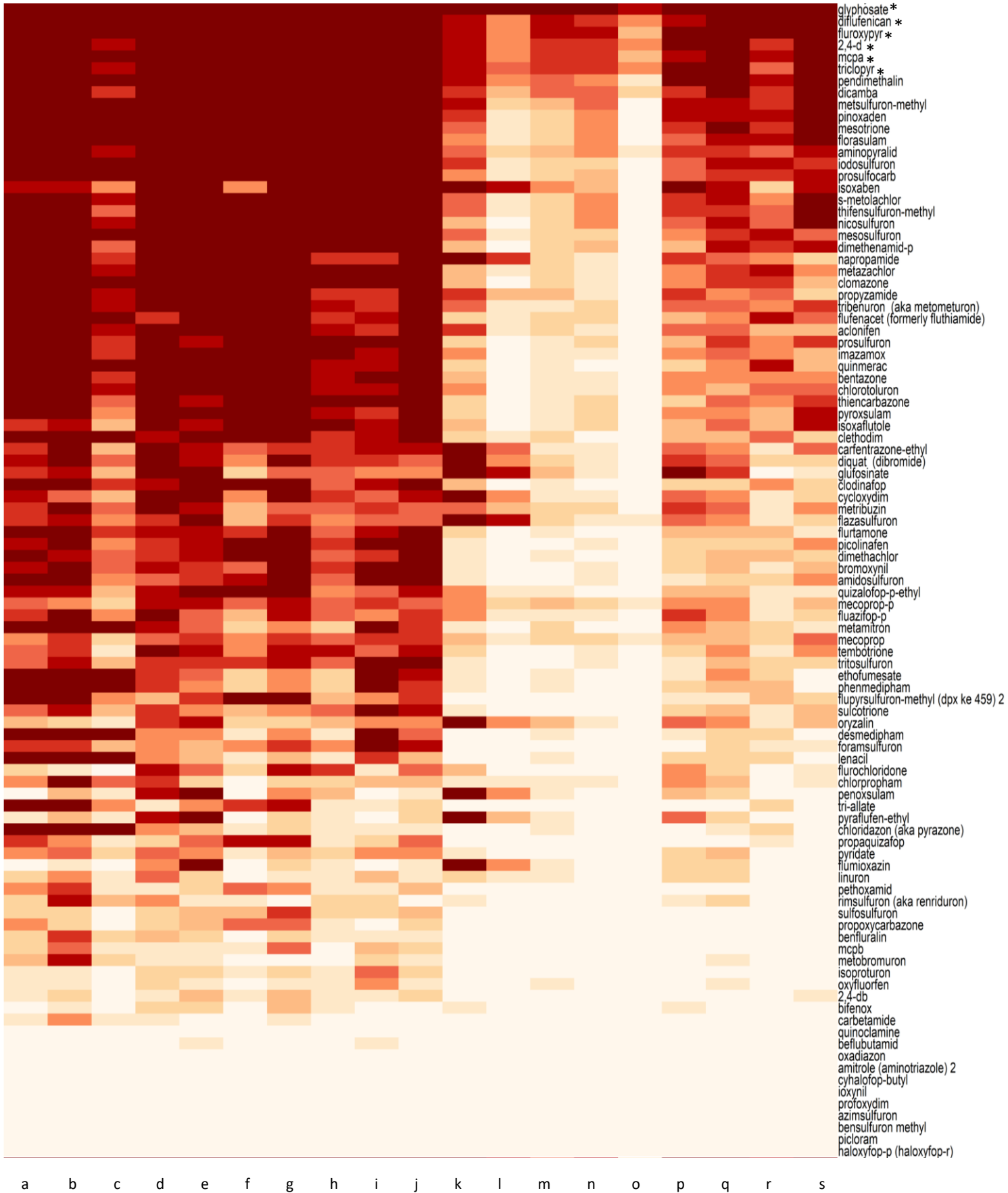
**Figure S4:** Differences and similarities in the clustering of postcodes produced by the mixture model with only 2017 substance purchase data (a) or 2015-2018 data (b). Postcode within a group share the same colour.

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.

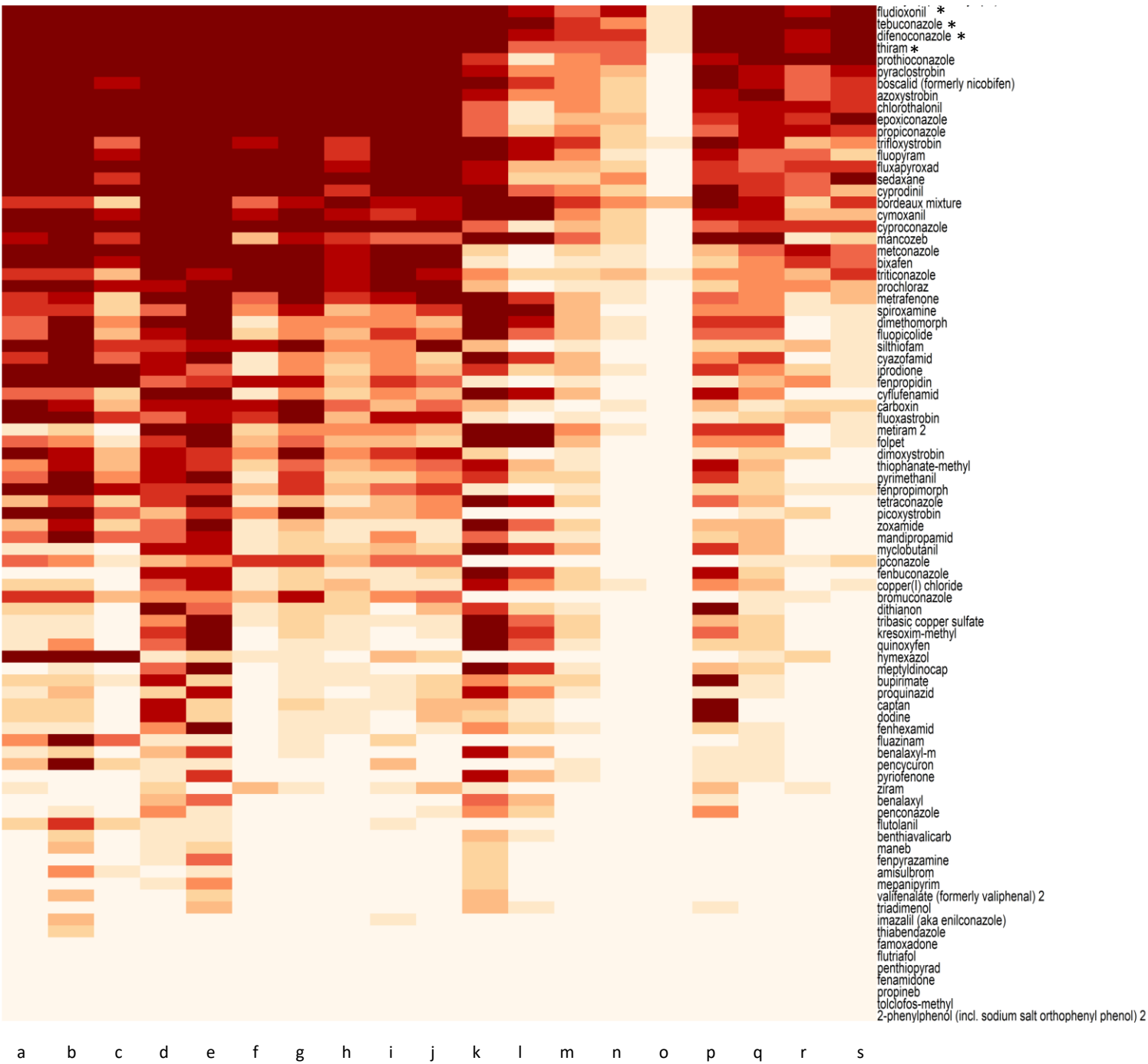


**Figure S5:** Variance of probabilities of substances to be in a group as a function of their mean probability to be in a group. Colours were set to show other (grey), discriminant (orange) and core (black) substances.

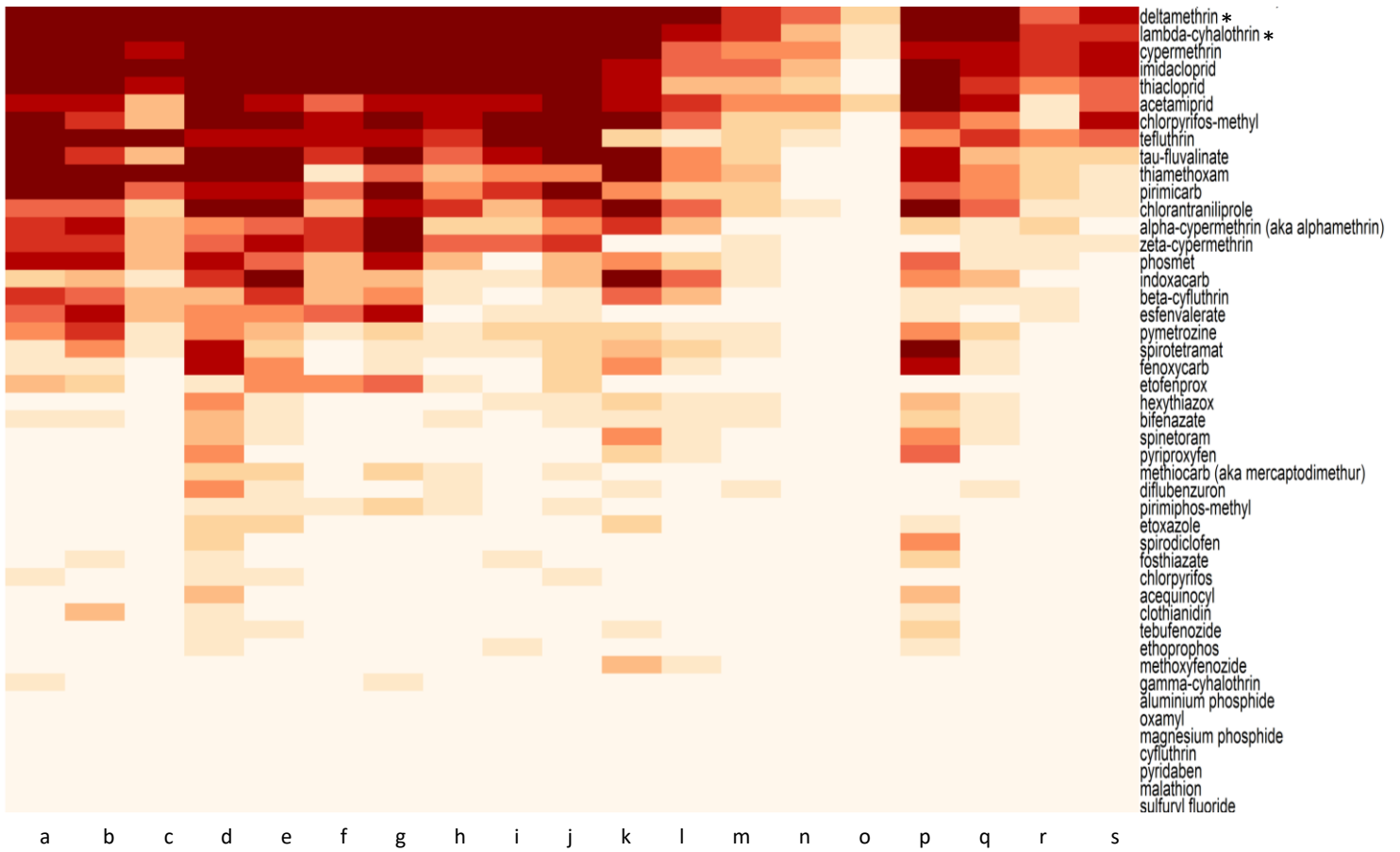
# Herbicides



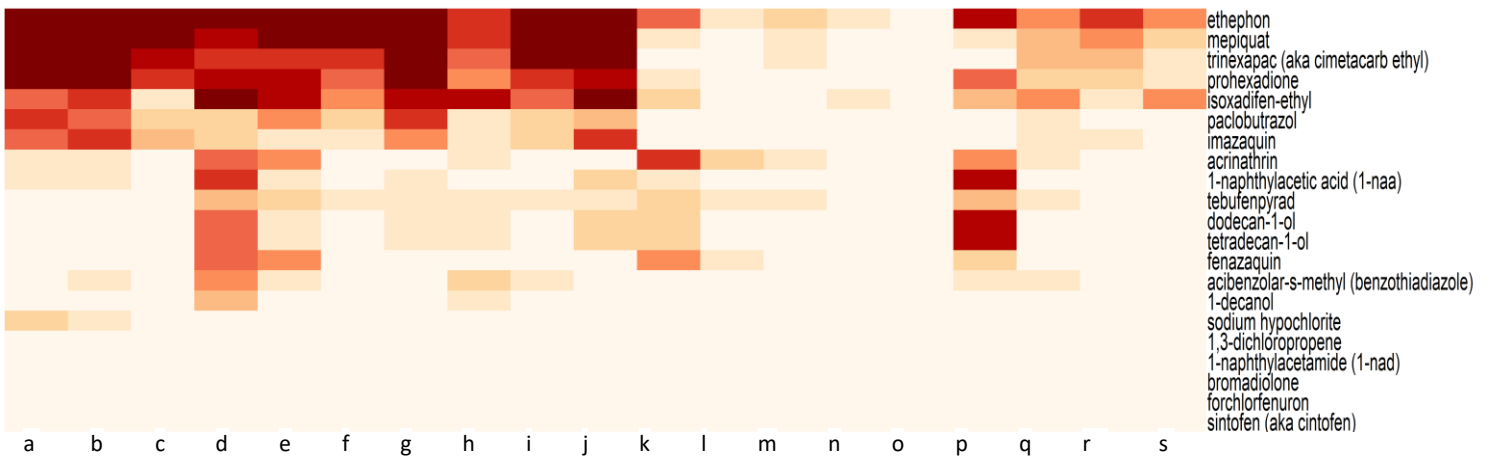
# Fungicides



## 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