



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-03815557v2

HAL Id: hal-03815557

<https://hal.science/hal-03815557v2>

Preprint submitted on 3 Mar 2023 (v2), last revised 27 Mar 2023 (v3)

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 CAIRO¹, Anne-Christine MONNET¹, Stéphane ROBIN^{1,2}, Emmanuelle PORCHER¹, Colin FONTAINE¹

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

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

34 INTRODUCTION

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

47 The effect of pesticides on biodiversity are usually demonstrated with a focus on
48 a single substance or a limited set of substances in general (e.g. thiamethoxam,
49 clothianidin, imidacloprid, thiacloprid or glyphosate (Botías et al., 2015; Busse et al.,
50 2001; Rundlöf et al., 2015; Van Bruggen et al., 2018). Yet, wild organisms are exposed
51 to complex mixtures (Dudley et al., 2017), owing to the diversity of substances
52 available and used in farmlands. Hence, studying substance mixtures is considered a
53 central task for environmental risk assessment (Lydy et al., 2004a), notably because
54 the effects of pesticide cocktails can strongly exceed the additive effects of single
55 compounds (Bopp et al., 2016; Junghans et al., 2006). Laboratory experiments
56 demonstrate synergetic interactions among substances within mixtures, affecting the
57 effect of the cocktails in non-additive ways (Cedergreen, 2014; Hernández et al., 2017;
58 Heys et al., 2016). While the importance of studying the effects of cocktails beyond
59 those of single substances was highlighted as soon as the late sixties (Keplinger and
60 Deichmann, 1967), and their evaluation is mandatory in the European Union since
61 2009 (EC No 1107/2009), few attempts to do so exist outside laboratories (Gibbons et
62 al., 2015).

63 Studies examining the effects of substance cocktails use two approaches:
64 bottom-up or top-down (Altenburger et al., 2013; Hernández et al., 2017; Relyea,
65 2009). The bottom-up approach aims at testing all possible mixture compositions,
66 starting from pairs of substances to more complex combinations. This method makes

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

87
88 Here, we introduce a new statistical method to identify relevant pesticide mixtures, i.e.
89 actual combinations of substances potentially co-occurring in agroecosystems, across
90 Metropolitan France. We overcame the general problem of limited availability of data
91 on temporal and spatial use of pesticides (Navarro et al., 2021) by taking advantage
92 of the recent publication of an up-to-date database on pesticide purchases in France,
93 the French national bank of pesticide sales database
94 (<https://www.data.gouv.fr/fr/datasets/ventes-de-pesticides-par-departement/>). This
95 database has registered mandatory reporting of quantities of active substances
96 purchased in France since 2013 (law n°2006-1772) at a relatively fine spatial grain
97 (postcode of the buyer). France is also the seventh largest user of pesticides in the
98 world (FAO 2020) and has a wide range of agricultural types (Urruty et al., 2016), which
99 makes it a well-suited case country to identify pesticide mixtures encountered in the
100 field by wild organisms, as well as their spatial variation.

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

111

112 METHODS

113 1.1 Pesticide data

114 Data on active substances were obtained from the French national bank of
115 pesticide sales (BNV-d; <https://bnvd.ineris.fr>). The BNV-d database registers active
116 substances under mandatory reporting. The seller indicates the amount of each active
117 substance purchased and the postcode of the buyer in the database. This database
118 thus indicates the quantity of active substances purchased at the spatial resolution of
119 the postcode of the buyer. Postcode are the third level of administrative division in
120 France, lower than the European Union NUTS3 level (administrative departments) and
121 range from 0.17 km² to 614.39 km² in metropolitan France (median = 62.79 km², Q1 =
122 19.59 km², Q3 =140.36 km²). Substances are identified with their generic name and a
123 unique identifier, the Chemical Abstracts Service number. We modified generic names
124 when synonyms were found. We only retained substances with a license fee (i.e. under
125 compulsory reporting) because we can expect thorough reporting for these.

126 The years registered in the database ranged from 2013 to 2020. We discarded
127 the year 2013 because of incomplete data during the first reporting year, and the two
128 latest years of the time series (2019 and 2020) because additions and changes in the
129 database are allowed for two years after reporting. Also, note that the legislation has
130 kept changing until 2016, with consequences for the mandatory nature of reporting for
131 some substances or treatments. In particular, until 2016 the geographical information
132 associated with seed coating substances was that of the seed coating company, not

133 of the buyer. Hence, 2017 can be considered the most accurate and thorough year
134 within the period 2013-2020.

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

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

162

163 *1.2 Model-based Clustering*

164 *1.2.1 Input data*

165

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

$$171 \quad Y_{ij} = \log\left(\frac{\text{quantity of substance } j \text{ bought in postcode } i}{\text{cropland area of postcode } i}\right)$$

172
 173 (Y_{ij} is NA when substance j is not bought in postcode i).

174

175 **1.2.2 Model**

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

184

185 **1.2.2.1 Groups definition**

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

$$189 \quad \pi_k = \Pr\{Z_i = k\}. \quad (1)$$

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

192

193 **1.2.2.1.2 Emission distribution**

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

198 $\gamma_{kj} = \Pr\{X_{ij} = 1|Z_i = k\}, \quad (2)$

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

201 $(Y_{ij}|X_{ij} = 1, Z_i = k) \sim \mathcal{N}(\mu_{kj}, \sigma_{kj}^2).$ (3)

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

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

209 $f_{jk}(x_{ij}, y_{ij}) = x_{ij}\gamma_{kj}\phi(y_{ij}; \mu_{kj}, \sigma_{kj}^2) + (1 - x_{ij})(1 - \gamma_{kj})$

210 denoting by $\phi(\cdot; \mu, \sigma^2)$ the probability density function of the Gaussian distribution
 211 $\mathcal{N}(\mu, \sigma^2)$.

212 To avoid over-parametrization, we also considered models with constrained variance,
 213 assuming either that the variance depends on the substance but not on the group:
 214 $\sigma_{kj}^2 \equiv \sigma_j^2$, or that the variance is the same for all substances in all groups: $\sigma_{kj}^2 \equiv \sigma^2$.

215

216 1.2.3 Inference

217

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

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

229

230 **1.2.3.1.1 Expectation-Maximization algorithm**

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

237 The EM algorithm alternates the steps 'E' (for expectation) and 'M' (for
238 maximization) until convergence. It can be shown that the likelihood of the data
239 $\log p(Y; \theta)$ increases after each EM step. The reader may refer to Dempster et al.
240 (1977) or McLahan and Peel (2000) for a formal justification of the procedure.

241

242 **1.2.3.1.2 E step**

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

$$247 \tau_{ik}^{(h-1)} = \Pr\{Z_i = k | \{(X_{ij}, Y_{ij})\}_{1 \leq j \leq p}; \theta^{(h-1)}\}$$

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

$$251 \hat{\tau}_{ik} = \hat{\pi}_k \prod_{j=1}^p \hat{f}_{jk}(x_{ij}, y_{ij}) / (\sum_{\ell=1}^K \hat{\pi}_\ell \prod_{j=1}^p \hat{f}_{j\ell}(x_{ij}, y_{ij})).$$

252

253 **1.2.3.1.3 M step**

254 The M step updates the parameter estimate by maximizing
255 $\mathbb{E}[\log p(Y, Z; \theta) | Y; \theta^{(h-1)}]$ with respect to θ . The objective function can be calculated
256 using the conditional probabilities τ_{ik} s

$$257 \mathbb{E}[\log p(Y, Z; \theta) | Y; \theta^{(h)}] = \sum_{i=1}^n \sum_{k=1}^K \hat{\tau}_{ik} (\log \pi_k + \sum_{j=1}^p \log f_{kj}(x_{ij}, y_{ij})).$$

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

$$261 \hat{N}_k = \sum_{i=1}^n \hat{\tau}_{ik}, \hat{M}_{kj} = \sum_{i=1}^n \hat{\tau}_{ik} x_{ij}.$$

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

$$265 \quad \hat{\pi}_k = \hat{N}_k/n, \hat{\gamma}_{kj} = \hat{M}_{kj}/\hat{N}_k.$$

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

$$267 \quad \hat{\mu}_{kj} = \frac{1}{\hat{M}_{jk}} \sum_{i=1}^n \hat{t}_{ik} x_{ij} y_{ij} \hat{\sigma}_{kj}^2 = \left(\frac{1}{\hat{M}_{jk}} \sum_{i=1}^n \hat{t}_{ik} x_{ij} y_{ij}^2 \right) - (\hat{\mu}_k)^2.$$

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

270

271 1.2.4 Model selection

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

$$276 \quad BIC = \log p(Y; \hat{\theta}) - \frac{n}{2} \log(\# \text{independent parameters}).$$

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

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

281

282 1.2.5 Estimated parameters

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

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

290 $\hat{\gamma}_{kj}$ was the probability of a substance j to be used in a postcode of group k . We used
 291 this probability to study the composition of active substances in each group k .

292 $\hat{\mu}_{kj}$ and $\hat{\sigma}_{kj}^2$ were the estimated mean and variance of the log-quantity of substance j
293 per square meter of cropland purchased in a postcode from group k . These quantities
294 were used to refine our understanding of the substance composition of postcode
295 groups.

296

297 *1.3 Analyses on estimated parameters*

298 *1.3.1 Spatial structure of postcode groups*

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

308

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

318

319 *1.3.2 Searching for the drivers of the substance composition of groups*

320 We tried to identify some of the possible drivers of the substance composition of
321 groups using two complementary approaches. First, we tested whether the groups
322 obtained with the model-based clustering, which by construction differ in terms of
323 active substances purchased, also differed in terms of crop composition. To compare
324 the proportion of area covered with different crops among groups, we performed a log-

325 ratio analysis (LRA). This approach was implemented in the R package *easyCODA*
326 (Greenacre, 2019, R Core Team 2021). Second, we used Mantel tests (Mantel &
327 Valand 1970) to estimate the correlations between three distance matrices among
328 postcode groups: distances in the composition of substances purchased in the group
329 (see above), distances in crop composition, and geographic distances. We used a
330 spearman method and used 9999 permutations, computed with the *vegan* package
331 (Oksanen and Simpson, 2022)

332

333 *1.3.3 Test of the temporal robustness of the model-based clustering*

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

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

347

348 RESULTS

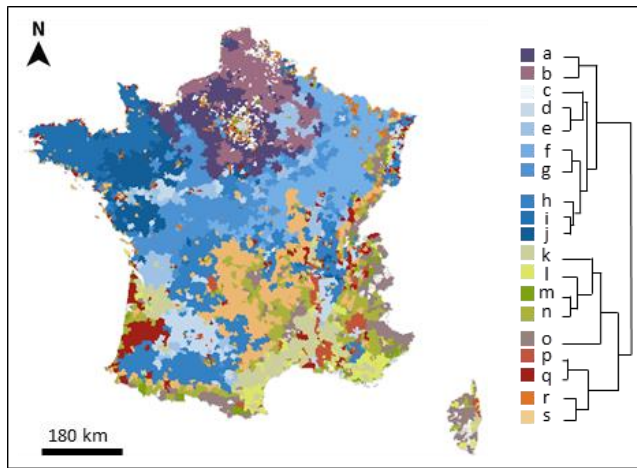
349

350 *1.4 The model-based clustering yields a small number of groups of postcodes*

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

357

358 Most groups of postcodes identified by the model-based clustering were spatially
 359 aggregated, albeit of contrasting sizes (Figure 1). The number of postcodes per group
 360 ranged from 135 to 493 (median = 270, Q1 = 215.5, Q3 = 378.5), which translated into
 361 a cropland area per group ranging from 38.7 km² to 24,184 km² (median = 5,573.7
 362 km², Q1 = 1,547.55 km², Q3 = 13,959 km²). The cropland area of groups was
 363 negatively related to the area of the convex envelop encompassing it, such that groups
 364 with the largest cropland area tended to be the most spatially clustered (Figure 2).
 365 Such a spatial clustering of postcodes purchasing similar pesticide substances was
 366 expected as agricultural practices are spatially structured (see below) but keep in mind
 367 that the model-based clustering did not incorporate spatial information.
 368

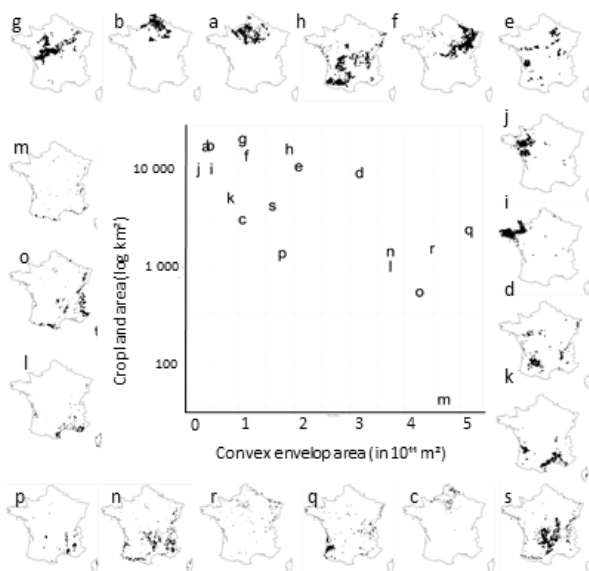


369
 370 *Figure 1: Map of France split into postcode groups obtained from the model-based clustering*
 371 *on the basis of active substances purchased within postcodes in 2017. Postcodes within a group*
 372 *share the same colour. The dendrogram was obtained using an agglomerative hierarchical*
 373 *clustering.*

374
 375 Postcode groups corresponded to specific geographical and/or agricultural
 376 regions. For example, group *i* corresponded mostly to Brittany (the western peninsula)
 377 and group *b* was predominantly located in Northern France. Groups *e* and *d* were more
 378 scattered across the country but overlapped almost perfectly with wine regions (*Figure*
 379 *2*). Note that a couple of groups were composed of a limited number of postcodes
 380 spatially scattered across France (e.g. groups *m* and *o* *Figure 2*). In particular, group
 381 *m* represented less than 39 km² of cropland and is generally discarded in the following.

382 The groups identified by the model-based clustering were relatively robust to a
 383 change in the temporal range of the data, as shown by the results of the clustering on

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



395
 396 *Figure 2: Relationship between cropland area (log scale) and convex area, a proxy for spatial*
 397 *extent, of groups. The spatial distribution of each group is plotted around the relationship, with*
 398 *one map of France per group, in which postcodes forming each group are highlighted in black.*
 399 *Groups are ordered clockwise from top left in decreasing cropland area. Note that the focus on*
 400 *cropland area (not total area) in a postcode makes some groups with little cropland (e.g.*
 401 *mountain areas, *q* or *m*) appear with a relatively large black area on the maps, although they*
 402 *are ranked low in terms of cropland area.*

403

404 1.5 Substance composition of postcode groups: core and discriminating substances

405 Postcode groups differed in terms of the composition of substances purchased
 406 (Figure 3), as expected from the clustering algorithm, but may also share common
 407 substances. Group composition was inferred, and can be characterised by, (1) the
 408 probability of a substance to be purchased in a postcode from a given group (\hat{y}_{kj}), and,

409 if the substance is purchased, (2) the estimated mean quantity purchased ($\hat{\mu}_{kj}$) as well
410 as (3) the estimated variance in the latter quantity (σ_{jk}^2). In the following, for the sake
411 of simplicity, we chose to focus on the probability of substances to be purchased,
412 knowing that this probability was positively related with the estimated mean quantity
413 (Figure S4 & Figure S6, $r = 0.2$) and negatively related with the estimated variance
414 (Figure S4, $r = -0.07$). For a given substance, this probability can also vary substantially
415 across groups, and we used this variability to distinguish two main types of substances
416 with interest for the definition of postcode groups and for the identification of relevant
417 pesticide mixtures : core substances and discriminating substances (Figure 4).

418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435

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

443

444 Core substances, defined as substances with a high average and low variance
445 of probability to be purchased across groups, were by definition found in most groups;
446 they were widespread molecules that were likely to form the backbone of mixtures
447 encountered by living organisms in farmland. Using an arbitrary threshold value of
448 mean purchase probability of 0.85, we found 12 such core substances with high
449 probabilities (Figure 3 & Figure S5): two pyrethroid insecticides (deltamethrin, lambda-
450 cyhalothrin), six herbicides of different chemical families (glyphosate, diflufenicanil,
451 fluroxypyr, MCPA, 2,4-d, triclopyr) and four fungicides (fludioxonil, tebuconazole,
452 difenoconazole and thiram). Because they were found with high probability in most
453 groups, these substances were unlikely to weight strongly in the definition of postcode
454 groups, although they can contribute via differences in the mean quantities used
455 across groups. For example, the average estimated amount of glyphosate purchased
456 ranged from 19 to 928 kg/ m² of cropland (median = 44, Q1 = 38, Q3 = 35) among
457 groups.

458 Discriminating substances are defined as substances with medium to high mean
459 probability of purchase, mechanically associated with a large variance across groups
460 in this probability (Figure S5). Because of their contrasting probability of purchase
461 across groups, discriminating substances were likely to contribute greatly to the
462 formation of groups. We used the arbitrary range of average probabilities from 0.5 to
463 0.85 to define discriminating substances. Using these thresholds, we found a set of 84
464 discriminating substances, including 45 herbicides, 25 fungicides, 10 insecticides and
465 4 with other targets (Supplementary information 2). In the following, we focus on
466 discriminating substances that are highly probable ($\hat{\gamma}_{kj} > 0.85$) in at least one postcode
467 group, i.e. substances that are likely major components of pesticide mixtures occurring
468 in a given group. We found seven widespread discriminating substances purchased
469 with a probability higher than 0.85 in at least 12 out of 19 groups: azoxystrobin,
470 boscalid, cypermethrin, mesotrione, metsulfuron-methyl, pendimethalin and
471 prothioconazole. These substances are very close to core substances. Conversely,

472 four substances were highly specific, being purchased with high probability (> 0.85) in
473 less than four groups (e.g. metribuzin in groups *d* and *b*). Within a group, the number
474 of discriminating substances with high probability of purchase (> 0.85) varied strongly
475 among groups, from 2 for group *r* to 80 for group *g* (mean = 43 ± 27). This cross-group
476 variation in the number of highly probable discriminating substances has implication
477 for the composition and complexity of pesticide mixtures in French agroecosystems:
478 from relatively “simple” (12 core substances and 11 discriminating substances in group
479 *q*) to highly complex (12 core substances and 74 discriminating substances in group
480 *g*).

481

482 The 156 remaining substances, with a low average probability to be purchased
483 (< 0.5), also had a role in group identification, but were seldom purchased and will not
484 be described further (Figure 3).

485

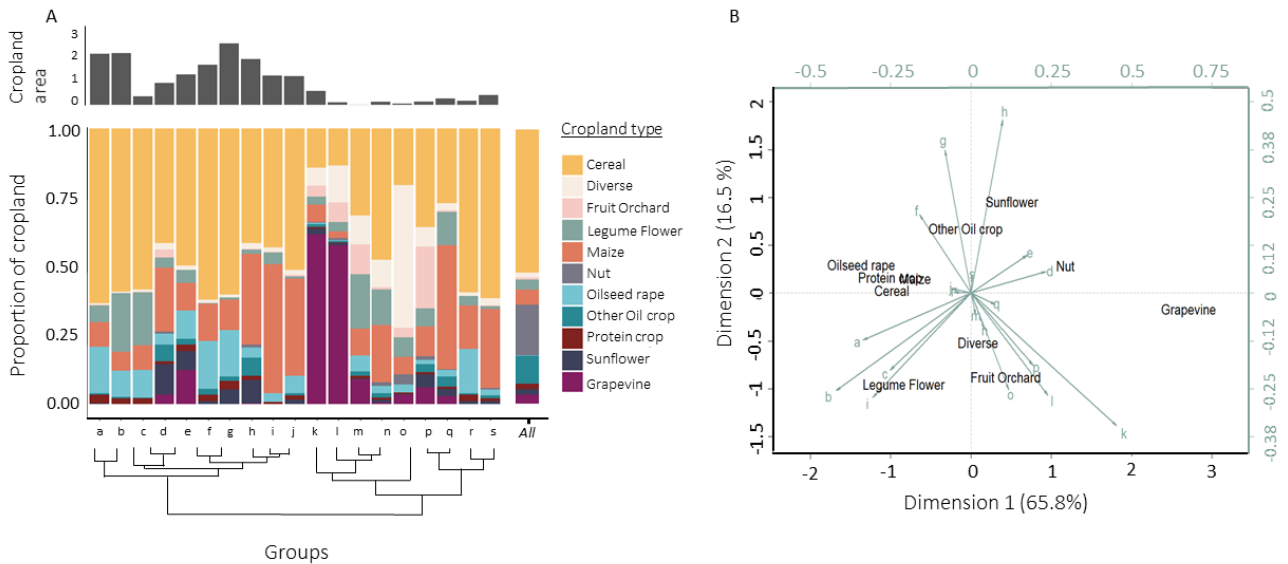
486 *1.6 Postcode groups differ in terms of crop composition, but active substance purchase may*
487 *not be solely driven by crop identity*

488

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

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

509



510
 511 *Figure 4: A. Distribution of crop type area across groups. The top grey histogram shows the*
 512 *distribution of total cropland area across groups (in 10⁴ km²). The dendrogram was obtained*
 513 *using an agglomerative hierarchical clustering on the basis of Ward's method among groups*
 514 *(see 2.2.1). B. Biplot of the log ratio analysis relating the proportion of crop types in each group.*
 515 *Only groups identified as spatially coherent are displayed (see 3.2). For readability, the groups*
 516 *and crop types are displayed on two different scales: black for crop types, green for groups. The*
 517 *size of arrows corresponds to the contribution of each group. Groups that appear close to each*
 518 *other on the biplot have similar crop composition, which can be inferred from the contribution*
 519 *of each crop type to the axes.*

520

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

529

530 DISCUSSION

531

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

546

547 *1.7 Significance of the identification of highly probable active substances, and of mixtures of* 548 *active substances characteristic of postcode groups, for the study of the impacts of* 549 *pesticides in the environment*

550

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

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

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

598 anyway consistent with surveys showing that active substances are rarely found alone
599 in the environment (Silva et al., 2019). It also further substantiates the need for a
600 broader assessment of the synergistic effects of pesticides on biodiversity, often
601 completed on a limited set of substances only (Schreiner et al., 2016; Silva et al.,
602 2019). For core substances, for example, some cocktail effects have already been
603 studied but mostly on pairs of substances (Brodeur et al., 2014; Peluso et al., 2022)
604 and more rarely for cocktails of three or more substances (Cedergreen, 2014; Glinski
605 et al., 2018; Van Meter et al., 2018). Focusing on the reasonable number of relatively
606 complex mixtures identified by the present approach would contribute to improve our
607 understanding of the synergistic effects of realistic cocktails on organisms.

608

609

610 *1.8 Limitations & perspectives*

611 *1.8.1 Limited spatio-temporal resolution of the BNV-d data*

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

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

633

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

650

651 *1.8.2 Going beyond the use of purchase probabilities and arbitrary thresholds to identify the* 652 *substances of interest for risk assessment*

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

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

664 information on the mean and variance of purchased amounts within postcodes, hence
665 mixtures differ for these variables as well. For example, glyphosate, a core substance
666 with high purchase probability in all postcode groups, was bought in contrasting
667 quantities across postcode groups: the average amount was 53.9 kg/km² and ranged
668 from 7.8 kg/km² in group *p* to 146 kg/km² in group *i*. Although the purchase probability
669 was positively correlated to the mean purchased quantity and negatively to its
670 variance, the correlation is not strong, and further analysis is needed to fully uncover
671 variation in substance quantities within the mixtures we identified.

672

673 *1.8.3 Taking into account the yearly variation in pesticide use*

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

692

693

694 *1.9 Postcode groups are related to the crop they grow, as well as to other regional factors,*
695 *but the underlying mechanisms remain to be fully identified*

696 Although no spatial information was included in the model-based clustering
697 analysis, the postcode groups exhibited a strong spatial structure, in which most
698 groups are strongly aggregated and only a few small groups are scattered across
699 France. Such spatial structure was expected since pesticide use is strongly crop-
700 dependent. For example, acetamiprid, a substance used to protect fruit trees or
701 grapevine against aphids, is bought with high probability in groups *l, e* and *d*, with high
702 proportion of fruit orchards and grapevines. Similarly, cyproconazole, a substance with
703 a broader spectrum of use, is bought with high probability in several groups with
704 contrasting crop compositions (*a, b, e, f, g, h, j, k, l, n, o, q, r* Figure 4). However,
705 deviations from this pattern were found: some adjacent postcode groups can have
706 different sets of crops but similar substance purchases or some spatially distant
707 postcode groups can have similar sets of crops but different substance purchases.
708 This observation suggests that local conditions, such as climate or pests, or some
709 regional patterns in the pesticide market and/or distribution, can drive the purchase of
710 active substances more than the set of crops grown (Silva et al., 2019; Storck et al.,
711 2017). Hence, the differences among postcode groups were related to a combination
712 of crop identity effects and other regional effects that will need additional analysis to
713 be identified. A straightforward perspective for the model-based clustering approach
714 would thus be to incorporate environmental covariates in the model, and evaluate how
715 clusters are modified.

716

717 CONCLUSION

718

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

727 Acknowledgement

728

729 This project was funded and supported by ANSES (grant agreement 2019-CRB-
730 03_PV19) via the tax on sales of plant protection products. The proceeds of this tax
731 are assigned to ANSES to finance the establishment of the system for monitoring the
732 adverse effects of plant protection products, called ‘phytopharmacovigilance’ (PPV),
733 established by the French Act on the future of agriculture of 13 October 2014. We wish
734 to thank the steering committee of the project: Fabrizio Botta, Sandrine Charles, Marc
735 Girondot, Olivier Le Gall, Thomas Quintaine, and Lynda Saibi-Yedjer. Milena Cairo
736 was supported by ANR project VITIBIRD (ANR-20-CE34-0008) while working on this
737 project. This work also benefitted from the support of the project ECONET (ANR-18-
738 CE02-0010) and of the “Chaire Modélisation Mathématique et Biodiversité”.

739

740 Conflict of interest

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

742

743 SUPPLEMENTARY MATERIALS

744

745 Supplementary materials to this article can be found online at

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

747

748

749 REFERENCES

750

751 Ali, S. F., Shieh, B. H., Alehaideb, Z., Khan, M. Z., Louie, A., Fageh, N., & Law, F. C.
752 (2011). A review on the effects of some selected pyrethroids and related
753 agrochemicals on aquatic vertebrate biodiversity. *Canadian Journal of Pure &*
754 *Applied Sciences*, 5(2), 1455-1464.

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

758 Boedeker, W., Watts, M., Clausing, P., Marquez, E., 2020. The global distribution of
759 acute unintentional pesticide poisoning: estimations based on a systematic
760 review. *BMC Public Health* 20, 1–19. [https://doi.org/10.1186/s12889-020-09939-](https://doi.org/10.1186/s12889-020-09939-0)
761 [0](https://doi.org/10.1186/s12889-020-09939-0)

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

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

766 Botías, C., David, A., Horwood, J., Abdul-Sada, A., Nicholls, E., Hill, E., Goulson, D.,
767 2015. Neonicotinoid Residues in Wildflowers, a Potential Route of Chronic
768 Exposure for Bees. *Environ. Sci. Technol.* 49, 12731–12740.
769 <https://doi.org/10.1021/acs.est.5b03459>

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

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

777 Busse, M.D., Ratcliff, A.W., Shestak, C.J., Powers, R.F., 2001. Glyphosate toxicity
778 and the effects of long-term vegetation control on soil microbial communities.
779 *Soil Biol. Biochem.* 33, 1777–1789. [https://doi.org/10.1016/S0038-](https://doi.org/10.1016/S0038-0717(01)00103-1)
780 [0717\(01\)00103-1](https://doi.org/10.1016/S0038-0717(01)00103-1)

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

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

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

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

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

795 Fritsch, C., Appenzeller, B., Burkart, L., Coeurdassier, M., Scheifler, R., Raoul, F.,
796 Driget, V., Powolny, T., Gagnaison, C., Rieffel, D., Afonso, E., Goydadin, A.C.,
797 Hardy, E.M., Palazzi, P., Schaeffer, C., Gaba, S., Bretagnolle, V., Bertrand, C.,
798 2022. Pervasive exposure of wild small mammals to legacy and currently used
799 pesticide mixtures in arable landscapes. *Sci. Rep.* 1–22.
800 <https://doi.org/10.1038/s41598-022-19959-y>

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

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

813 <https://doi.org/10.1016/j.baae.2009.12.001>
814 Gelbard, R., Goldman, O., Spiegler, I., 2007. Investigating diversity of clustering
815 methods: An empirical comparison. *Data Knowl. Eng.* 63, 155–166.
816 <https://doi.org/10.1016/j.datak.2007.01.002>
817 Gibbons, D., Morrissey, C., Mineau, P., 2015. A review of the direct and indirect
818 effects of neonicotinoids and fipronil on vertebrate wildlife. *Environ. Sci. Pollut.*
819 *Res.* 22, 103–118. <https://doi.org/10.1007/s11356-014-3180-5>
820 Glinski, D.A., Purucker, S.T., Van Meter, R.J., Black, M.C., Henderson, W.M., 2018.
821 Endogenous and exogenous biomarker analysis in terrestrial phase amphibians
822 (*Lithobates sphenoccephala*) following dermal exposure to pesticide mixtures.
823 *Env. chem* 60, 1–24. <https://doi.org/10.1071/EN18163>.
824 Greenacre, M., 2019. Variable Selection in Compositional Data Analysis Using
825 Pairwise Logratios. *Math. Geosci.* 51, 649–682. [https://doi.org/10.1007/s11004-](https://doi.org/10.1007/s11004-018-9754-x)
826 [018-9754-x](https://doi.org/10.1007/s11004-018-9754-x)
827 Hallmann, C.A., Foppen, R.P.B., Van Turnhout, C.A.M., De Kroon, H., Jongejans, E.,
828 2014. Declines in insectivorous birds are associated with high neonicotinoid
829 concentrations. *Nature* 511, 341–343. <https://doi.org/10.1038/nature13531>
830 Hendrix, C.S., 2017. The streetlight effect in climate change research on Africa. *Glob.*
831 *Environ. Chang.* 43, 137–147. <https://doi.org/10.1016/j.gloenvcha.2017.01.009>
832 Hernández, A.F., Gil, F., Lacasaña, M., 2017. Toxicological interactions of pesticide
833 mixtures: an update. *Arch. Toxicol.* 91, 3211–3223.
834 <https://doi.org/10.1007/s00204-017-2043-5>
835 Heys, K.A., Shore, R.F., Pereira, M.G., Jones, K.C., Martin, F.L., 2016. Risk
836 assessment of environmental mixture effects. *RSC Adv.* 6, 47844–47857.
837 <https://doi.org/10.1039/c6ra05406d>
838 Humann-Guillemint, Ségolène, Binkowski, Ł.J., Jenni, L., Hilke, G., Glauser, G.,
839 Helfenstein, F., 2019. A nation-wide survey of neonicotinoid insecticides in
840 agricultural land with implications for agri-environment schemes. *J. Appl. Ecol.*
841 56, 1502–1514. <https://doi.org/10.1111/1365-2664.13392>
842 Humann-Guillemint, S., Tassin de Montaigne, C., Sire, J., Grünig, S., Gning, O.,
843 Glauser, G., Vallat, A., Helfenstein, F., 2019. A sublethal dose of the
844 neonicotinoid insecticide acetamiprid reduces sperm density in a songbird.
845 *Environ. Res.* 177, 108589. <https://doi.org/10.1016/j.envres.2019.108589>
846 Junghans, M., Backhaus, T., Faust, M., Scholze, M., Grimme, L.H., 2006. Application
847 and validation of approaches for the predictive hazard assessment of realistic
848 pesticide mixtures. *Aquat. Toxicol.* 76, 93–110.
849 <https://doi.org/10.1016/j.aquatox.2005.10.001>
850 Keplinger, M.L., Deichmann, W.B., 1967. Acute toxicity of combinations of pesticides.
851 *Toxicol. Appl. Pharmacol.* 10, 586–595. [https://doi.org/10.1016/0041-](https://doi.org/10.1016/0041-008X(67)90097-X)
852 [008X\(67\)90097-X](https://doi.org/10.1016/0041-008X(67)90097-X)
853 Levavasseur, F., Martin, P., Bouty, C., Barbottin, A., Bretagnolle, V., Théron, O.,
854 Scheurer, O., Piskiewicz, N., 2016. RPG Explorer: A new tool to ease the
855 analysis of agricultural landscape dynamics with the Land Parcel Identification
856 System. *Comput. Electron. Agric.* 127, 541–552.
857 <https://doi.org/10.1016/j.compag.2016.07.015>
858 Lewis, K.A., Tzilivakis, J., Warner, D.J., Green, A., 2016. An international database
859 for pesticide risk assessments and management. *Hum. Ecol. risk Assess.* 22,
860 1050–1064. <https://doi.org/10.1017/CBO9781107415324.004>
861 Lopez-Antia, A., Ortiz-Santaliestra, M.E., Mougeot, F., Mateo, R., 2015. Imidacloprid-
862 treated seed ingestion has lethal effect on adult partridges and reduces both

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

865 Lydy, M., Belden, J., Wheelock, C., Hammock, B., Denton, D., 2004a. Challenges in
866 regulating pesticide mixtures. *Ecol. Soc.* 9. [https://doi.org/10.5751/ES-00694-](https://doi.org/10.5751/ES-00694-090601)
867 [090601](https://doi.org/10.5751/ES-00694-090601)

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

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

873 Millot, F., Decors, A., Mastain, O., Quintaine, T., Berny, P., Vey, D., Lasseur, R., Bro,
874 E., 2017. Field evidence of bird poisonings by imidacloprid-treated seeds: a
875 review of incidents reported by the French SAGIR network from 1995 to 2014.
876 *Environ. Sci. Pollut. Res.* 24, 5469–5485. [https://doi.org/10.1007/s11356-016-](https://doi.org/10.1007/s11356-016-8272-y)
877 [8272-y](https://doi.org/10.1007/s11356-016-8272-y)

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

924 Tang, F.H.M., Lenzen, M., McBratney, A., Maggi, F., 2021. Risk of pesticide pollution
925 at the global scale. *Nat. Geosci.* 14, 206–210. [https://doi.org/10.1038/s41561-](https://doi.org/10.1038/s41561-021-00712-5)
926 [021-00712-5](https://doi.org/10.1038/s41561-021-00712-5)

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

929 Tsvetkov, N., Zayed, A., 2021. Searching beyond the streetlight: Neonicotinoid
930 exposure alters the neurogenomic state of worker honey bees. *Ecol. Evol.* 11,
931 18733–18742. <https://doi.org/10.1002/ece3.8480>

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

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

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

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

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

951 Zubrod, J.P., Bundschuh, M., Arts, G., Brühl, C.A., Imfeld, G., Knäbel, A.,
952 Payraudeau, S., Rasmussen, J.J., Rohr, J., Scharmüller, A., Smalling, K.,
953 Stehle, S., Schulz, R., Schäfer, R.B., 2019. Fungicides: An Overlooked Pesticide
954 Class? *Environ. Sci. Technol.* 53, 3347–3365.
955 <https://doi.org/10.1021/acs.est.8b04392>
956

APPENDIX

Bayesian Information Criterion

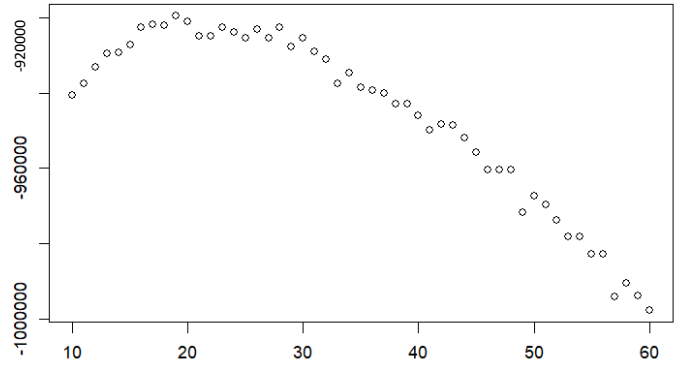
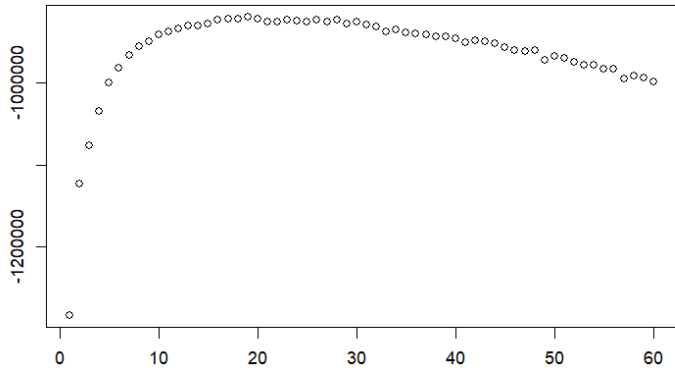


Figure S1: Values of BIC as a function of the number of groups in the EM algorithm. Panel a shows the full range of number of groups tested (from 1 to 40). Panel b is a closeup around the maximum BIC value

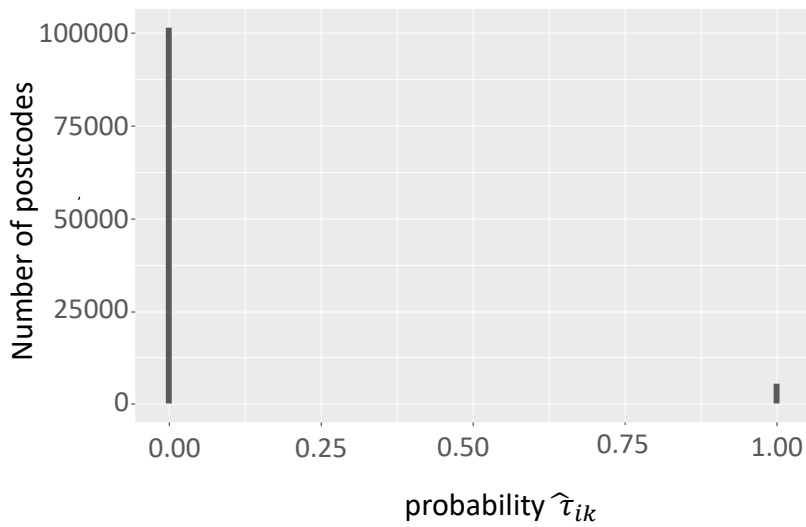


Figure S2: Distribution of $\hat{\tau}_{ik}$, the probability of postcode i to be in group k

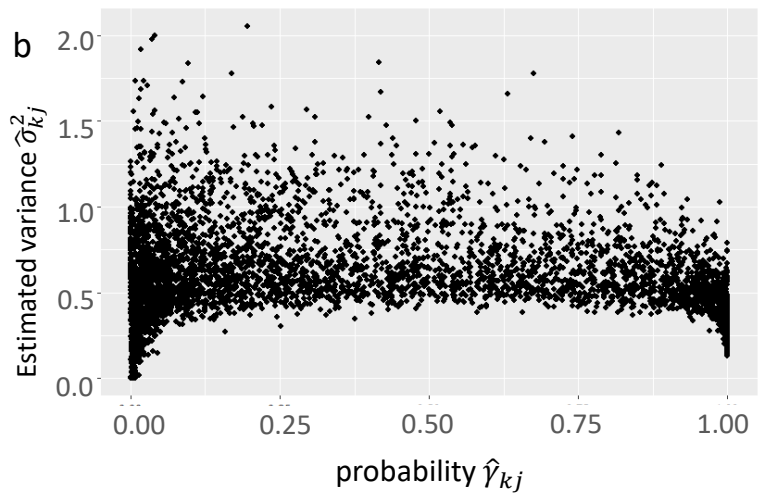
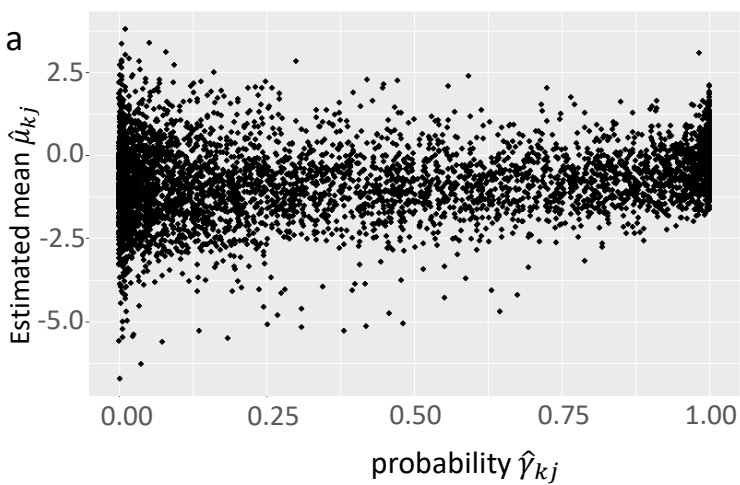


Figure S3: Estimated mean ($\hat{\mu}_{kj}$, panel a) and variance $\hat{\sigma}_{kj}^2$, panel b) of substance quantities purchased in a group as a function of the probability of a substance to be in a group $\hat{\gamma}_{kj}$.

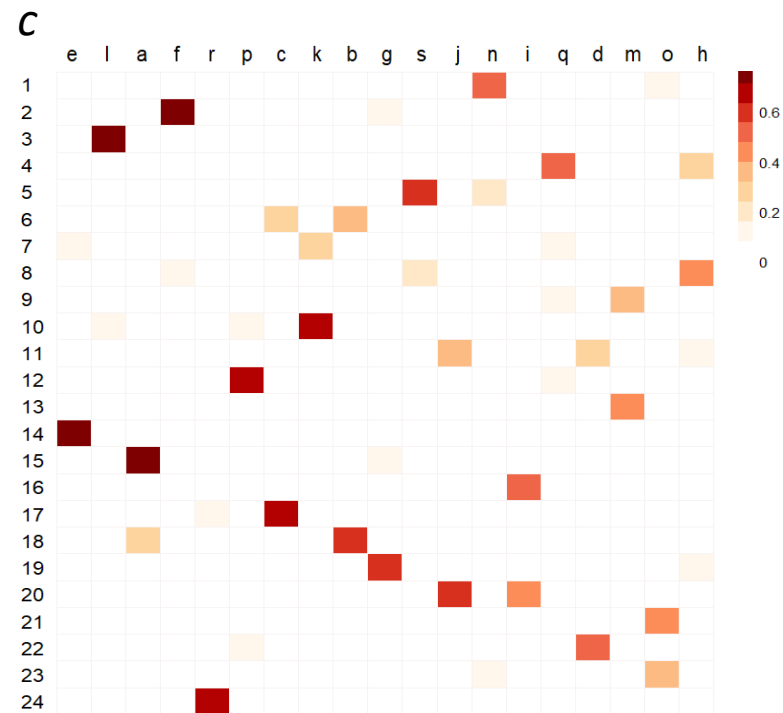
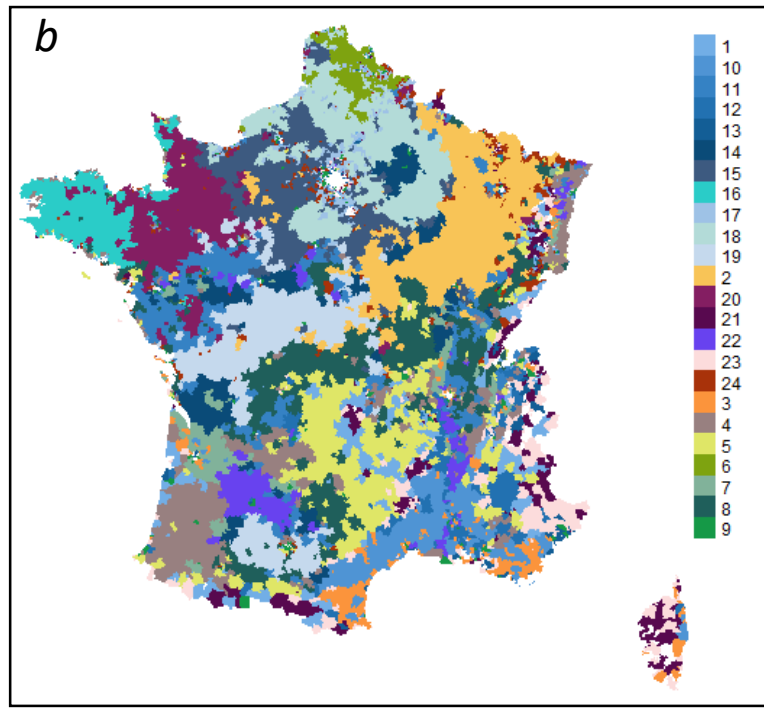
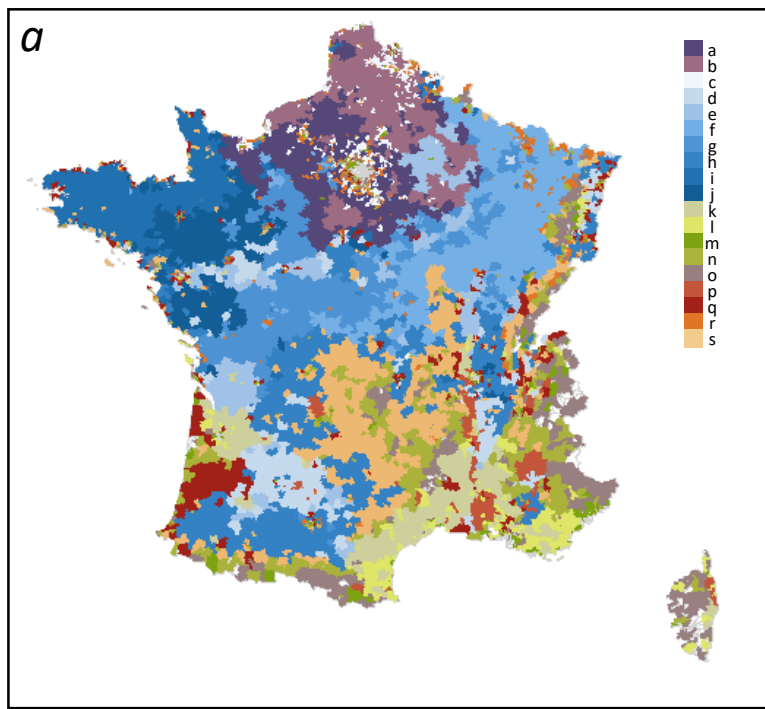


Figure S4: Differences and similarities in the clustering of postcodes produced by the mixture model with only 2017 substance purchase data (a) or 2015-2018 data (b). Postcode within a group share the same colour.

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.

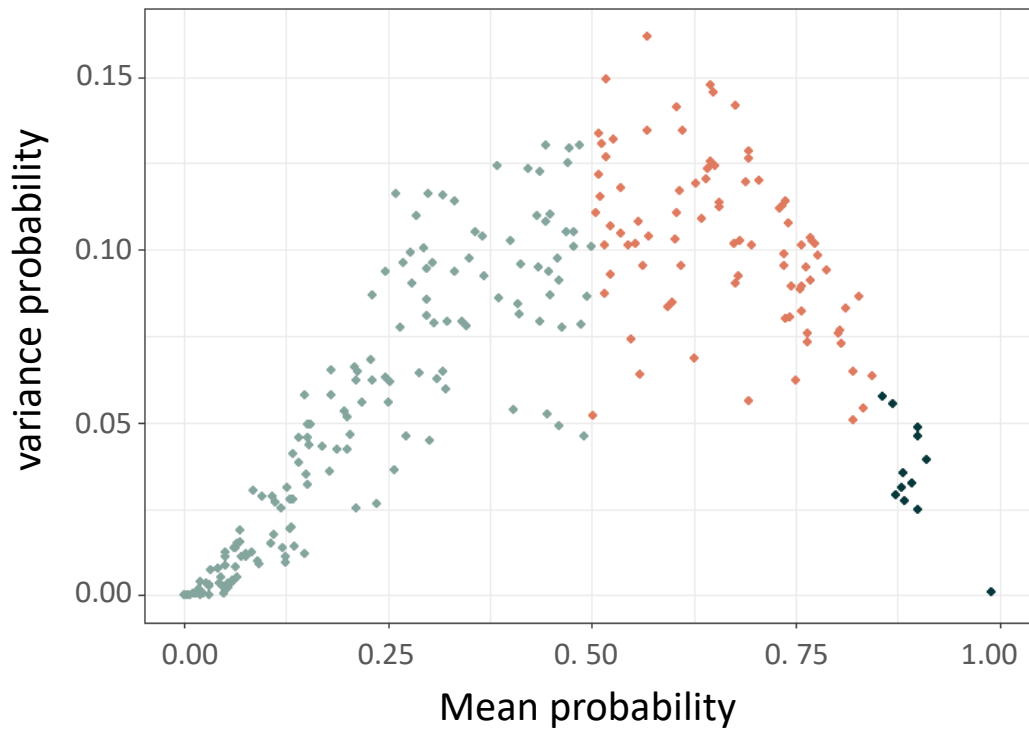
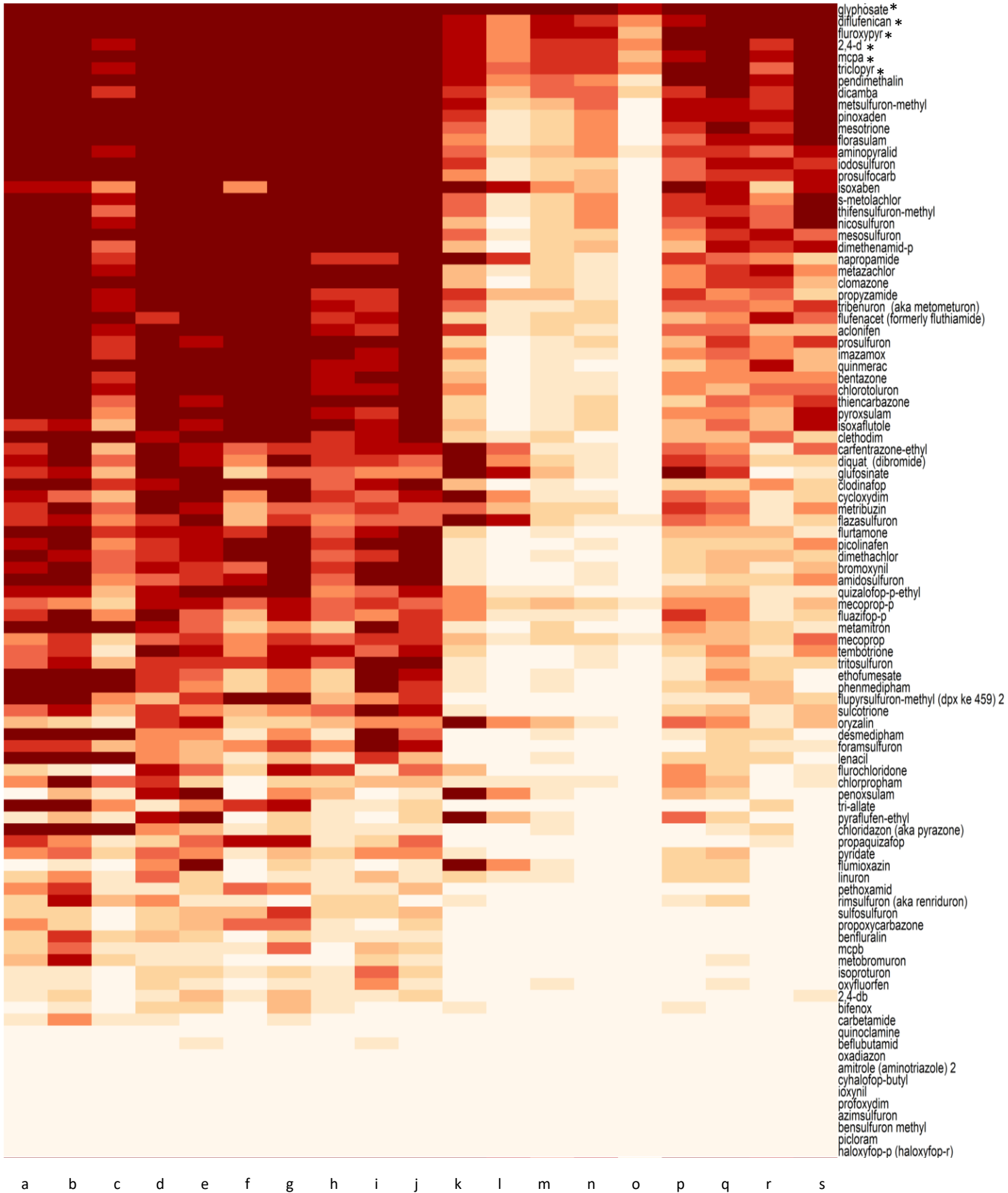
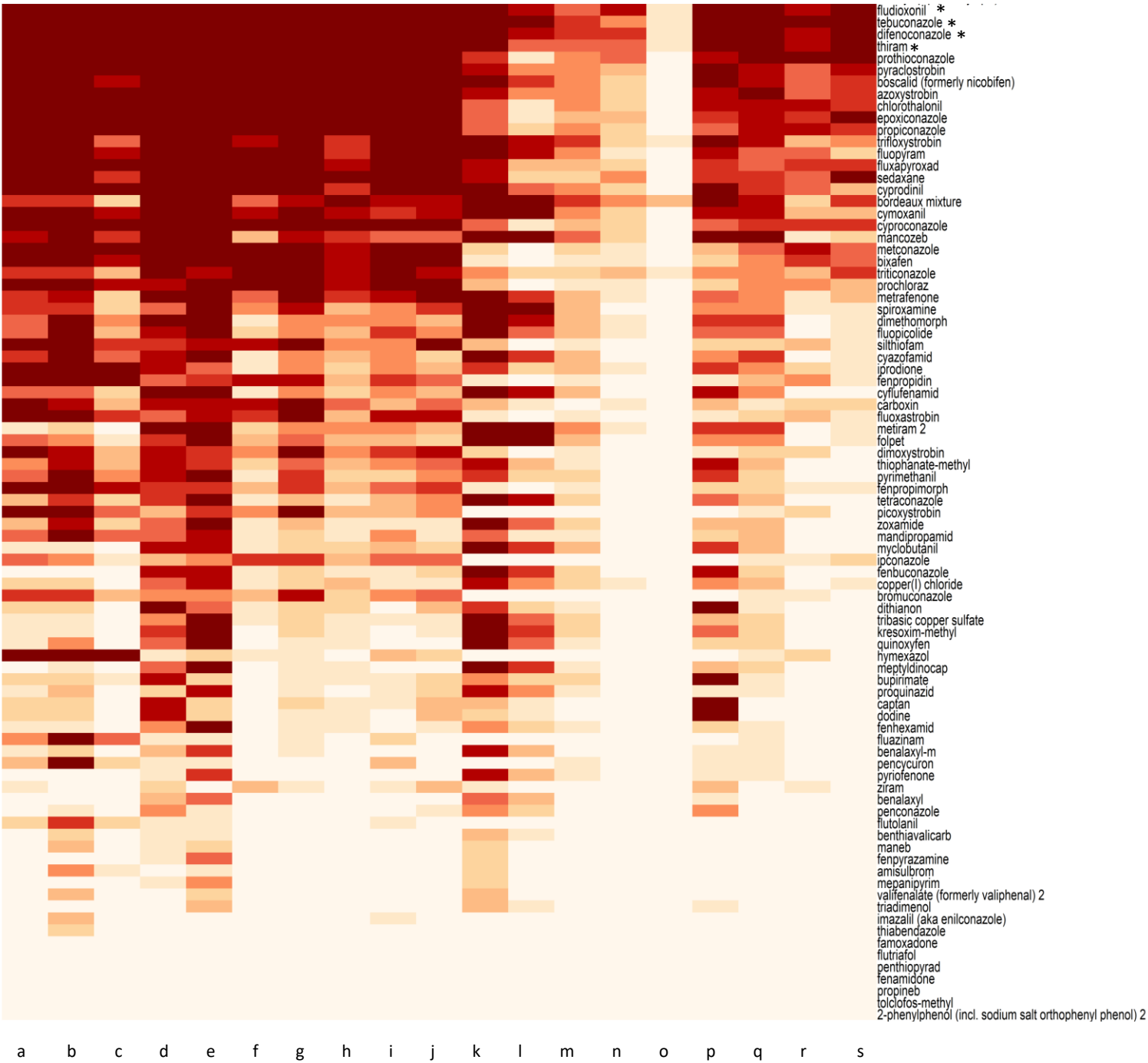


Figure S5: Variance of probabilities of substances to be in a group as a function of their mean probability to be in a group. Colours were set to show other (grey), discriminant (orange) and core (black) substances.

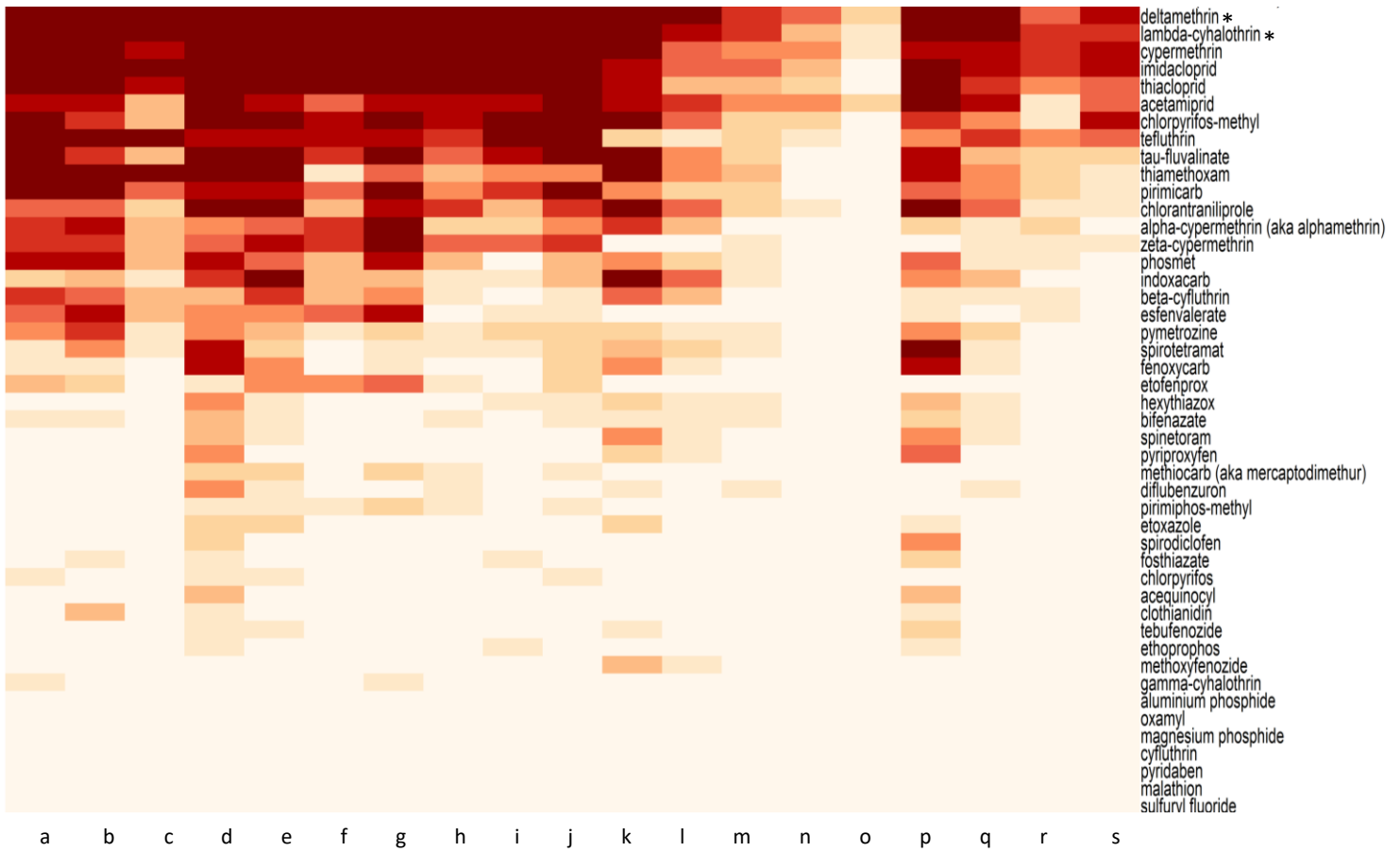
Herbicides



Fungicides



Insecticides



Other targets

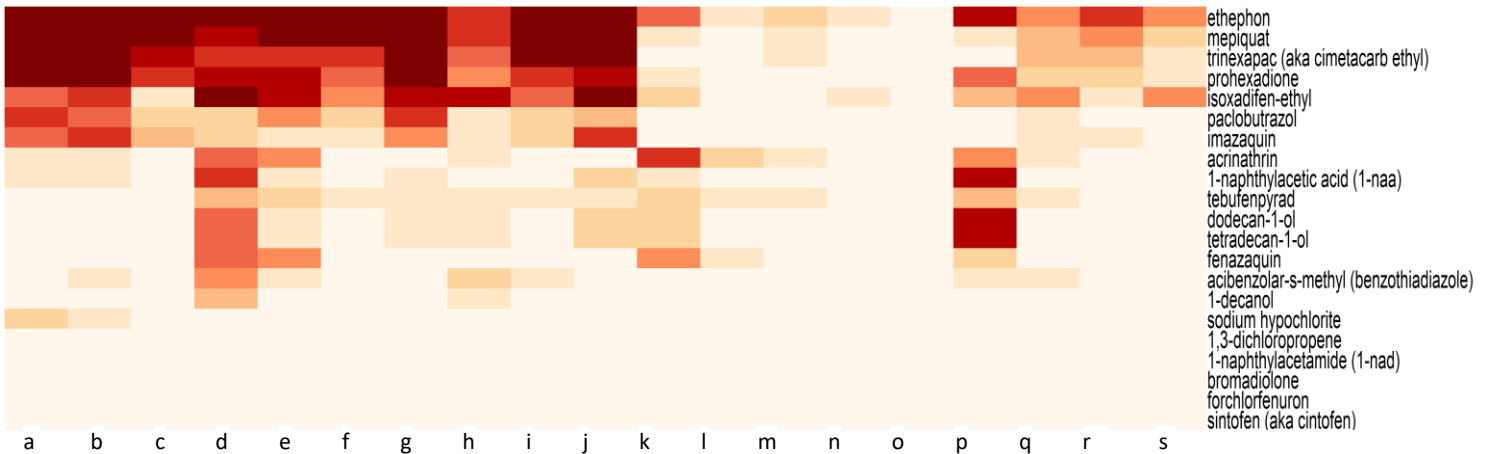


Figure S6: Heatmap of probability $\hat{\gamma}_{kj}$, that substance j is used in postcode k . Groups were obtained from a mixture models optimized by maximum likelihood with an iterative method: Expectation Maximization. Groups were ordered by similar composition of substance purchases. Substances 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