Patterns of cleaning product exposures using a novel clustering approach for data with correlated variables

Matthieu Marbac, Mohammed Sedki, Marie-Christine Boutron-Ruault, Orianne Dumas

Patterns of cleaning product exposures using a novel clustering approach for data with correlated variables

Matthieu Marbac[a], Mohammed Sedki[b], Marie-Christine Boutron-Ruault[c,d], Orianne Dumas[e,f,*]

[a] CREST, Ensai, Campus de KerLan, 35172 BRUZ, France

Address: Ensai, Campus de KerLan, 35172 BRUZ, France

eMail: matthieu.marbac-lourdelle@ensai.fr

[b] Université Paris-Sud and INSERM UMR 1181 B2PHI, Villejuif, France

Address : INSERM UMR 1181, 16 avenue Paul Vaillant Couturier, 94807 Villejuif cedex, France

eMail: mohammed.sedki@u-psud.fr

[c] CESP, Inserm U1018, Université Paris-Sud, UVSQ, Université Paris-Saclay, 94805, Villejuif, France

[d] Gustave Roussy Institute, Villejuif, France

Address: CESP Inserm U1018, Gustave Roussy, Espace Maurice Tubiana, 114 rue Edouard Vaillant, 94804 Villejuif cedex, France.

eMail: Marie-christine.BOUTRON@gustaveroussy.fr

[e] INSERM, VIMA: Aging and chronic diseases. Epidemiological and public health approaches, U1168, F-94807, Villejuif, France

[f] Univ Versailles St-Quentin-en-Yvelines, UMR-S 1168, F-78180, Montigny le Bretonneux, France

* *Corresponding author*

Address: Inserm UMRS 1168, VIMA- Aging and chronic diseases - Epidemiological and public health approaches, 16 avenue Paul Vaillant Couturier, 94807 Villejuif cedex, France

eMail: orianne.dumas@inserm.fr

Word count: 3,733

Abstract

Purpose: Clustering methods may be useful in epidemiology to better characterize exposures and account for their multidimensional aspects. In this context, application of clustering models allowing for highly dependent variables is of particular interest. We aimed to characterize patterns of domestic exposure to cleaning products using a novel clustering model allowing for highly dependent variables.

Methods: To identify domestic cleaning patterns in a large population of French women, we used a mixture model of dependency blocks. This novel approach specifically models within-class dependencies, and is an alternative to the latent class model which assumes conditional independence. Analyses were conducted in 19,398 participants of the E3N study (women aged 61-88 years) who completed a questionnaire regarding household cleaning habits.

Results: Seven classes were identified, which differed for the frequency of cleaning tasks (e.g. dusting/sweeping/hoovering) and use of specific products (e.g. bleach, sprays). The model also grouped the variables into conditionally independent blocks, providing a summary of the main dependencies among the variables.

Conclusions: The mixture model of dependency blocks, a useful alternative to the latent class model may have broader application in epidemiology, in particular in the context of exposome research and growing need for data-reduction methods.

Key words: Cluster Analysis; Classification; Disinfectants; Household cleaning; Latent Class.

List of abbreviations:

E3N: *Etude Epidémiologique auprès des femmes de la Mutuelle Générale de l'Education Nationale* (Epidemiological study among women of a French National health insurance plan covering mostly teachers)

ICL: Integrated Completed Likelihood

INTRODUCTION

Clustering methods are increasingly used in epidemiology to characterize and account for multidimensional aspects of both outcomes and exposures. Clustering models aim at identifying homogeneous groups (classes) of participants based on a large set of characteristics [1–3]. This approach is widely used, for instance to study complex and heterogeneous traits [4] such as mental health disorders [5] or, more recently, asthma [1]. Regarding exposures, cluster analysis has been traditionally used in nutritional epidemiology to derive dietary patterns [6,7]. In air pollution studies, clustering approaches have also been proposed as one solution to the issue of multi-pollutant or highly correlated exposures [8,9]. However, application of clustering approaches to characterize exposures or risk factors for diseases remain scarce [10–12]. A broader use of data-reduction approaches to better characterize environmental exposures is of specific interest, especially in the context of exposome research and the need to take into consideration the multiplicity and correlations of exposures [13–16].

Many people, especially women, are regularly exposed to cleaning products in private homes, and corresponding health hazards are increasingly acknowledged [17,18]. Associations have been reported between professional and domestic cleaning, and respiratory [17,19–22] and cardiovascular [18] health. However, the specific tasks and substances at risk still need to be elucidated. Household cleaning implies various tasks and the use of many chemicals, driven by general habits or behaviors. Identifying domestic cleaning exposure patterns, i.e. aggregating members of a study population into homogeneous clusters with similar characteristics, would help characterize individual exposures and their links with health outcomes. In epidemiological studies, domestic cleaning exposures are usually evaluated by questionnaires that assess frequency of numerous tasks and use of various products [19–21]. Characterization of domestic

cleaning patterns thus requires clustering models allowing for ordinal and possibly highly dependent variables.

Among clustering approaches, finite mixture models [23] achieve the clustering goal in a probabilistic framework. These approaches model the distribution of the observed variables and non-observed partition, and provide a classification probability for each individual. Finite mixture models have several strengths. First, probabilistic tools are available to address the question of how many classes should be selected. In addition, missing data can be managed, assuming that variables are missing at random [24]. Finally, this approach generally requires fewer assumptions than other clustering methods [25]. The classical latent class model [3] is a subgroup of finite mixture model based on the important assumption of conditional independence (i.e., that within each latent class, all variables are statistically independent). This model is a powerful approach to cluster categorical data, and is easily implemented and interpreted. However, it suffers from severe biases when some within-class dependencies occur [26]. The mixture model of dependency blocks, an extension of the latent class model relaxing the conditional independence assumption, has recently been developed [27]. This model groups the observed variables into conditionally independent blocks. The main within-class dependencies are thus reflected by the grouping of the variables into blocks.

Using data from a large subsample of the French E3N study (Etude Epidémiologique auprès des femmes de la Mutuelle Générale de l'Education Nationale), we aimed to identify and characterize domestic cleaning patterns among women. For this purpose, we used a mixture model of dependency blocks, extended to ordinal data having the same number of modalities, to identify both classes (similar patterns of responses across individuals) and blocks (groups of

variables that are correlated within classes). This paper presents this novel approach for the first time in an epidemiological study.

METHODS

Study population

The E3N study, initiated in 1990, is a prospective cohort among women of the Mutuelle Générale de l'Education Nationale (MGEN, a French National health insurance plan covering mostly teachers) [28]. A total of 98,997 women aged 40-65 were included at baseline and have been followed-up approximately every two years. The current analysis uses data from a nested case-control study on asthma (Asthma-E3N) conducted in 2011-2013 [29]. A total of 7,100 women with asthma and 14,200 aged-matched women without asthma were invited to complete a questionnaire regarding respiratory health and environmental exposures. Questionnaires were returned by 19,398 participants (91.8%) [29]. The study protocol was approved by the French Institutional Ethics Committee and all participants gave written informed consent.

The study included detailed standardized questionnaire [19–21] on the frequency of cleaning tasks performed and products used for domestic cleaning. Questions related to three main themes: domestic tasks (ten questions), use of specific cleaning products (seven questions), and use of different types of sprays (seven questions). Women were asked how frequently they did household cleaning and used each cleaning products or spray: never, <1 day/week, 1-3 days/week, or 4-7 days/week.

Finally, three additional variables were of interest in the current study: age, education level (defined as completion or not of at least 3 years of education after high school) and household help (defined as positive or negative answer to the question “does someone help you

for household cleaning, e.g. husband, household employee, or family members?") [21]. These variables were expected to be associated with domestic cleaning habits, and thus to vary across the identified domestic cleaning patterns.

Characterization of domestic cleaning patterns

Participants were classified based on their responses to the 24 questions on cleaning tasks and products used for domestic cleaning (four-level ordinal variables). Dependencies between the 24 variables of interest for the classification were evaluated using the Cramer's V, which measures the dependency between categorical variables [30]. To identify domestic cleaning patterns, we used a mixture model of dependency blocks [27] (see next section). To illustrate the interest of this novel model over classical methods, we also applied latent class models which assume conditional independence. Agreement between the classifications obtained by latent class models and by the mixture model of dependency blocks was evaluated using the ARI (Adjusted Rand Index), an index measuring the proximity between two partitions having possibly different numbers of classes [31]. ARI values close to 1 (maximum) indicate high agreement between the partitions, while values close to 0 indicate absence of agreement. Finally, to evaluate the discriminative properties of the classes produced by the mixture model of dependency blocks, we studied the differences of socio-demographic characteristics across the resulting domestic cleaning classes.

Mixture model of dependency blocks

The mixture model of dependency block has been described in details in a previous publication of one of the authors [27], and further information is provided in the appendix. The

approach specifically models within-class dependencies, and is thus more flexible than the classical latent class model which assumes conditional independence. Briefly, we postulate that the observed population consists of *K* classes (components) of individuals similar to each other based on the variables of interest. To deal with potential within-class dependencies between the variables, the model splits the variables into *B* within-class independent blocks. A specific distribution is used to model variables into blocks by considering their dependencies.

Model interpretation can be done in three steps. First the model evaluates parameters$\pi_1, \ldots, \pi_K$, corresponding to the marginal probability that an individual belongs a given class, reflecting the importance of each class. Second, each class can be summarized by the probability that an individual takes level *l* for the variable *j*, conditionally on belonging to class *k* (often referred to as "posterior probabilities"). In the current study, as each of the *j*=24 variables had *l*=4 levels, we used the posterior mode (i.e., the level with the highest posterior probability) of each variable and its probability to describe the classes. These first two steps of interpretation are common with the latent class model. Finally, for each class *k*, the parameterρ_{kb} reflects the strength of the intra-class dependencies between variables grouped into the same block under each class, and is similar to a correlation coefficient. The parameter ρ_{kb} measures the dependencies within component *k* between all variables of block *b*; it is thus more general than the Cramer's V (dependencies between two categorical variables). Examination of the within-block dependencies provides useful information regarding potential co-linearity between variables.

A mixture model of dependency blocks is defined by a number of classes, a number of blocks and the assignment of the variables into blocks. The model is not specified in advance, but is inferred from the data. We used the ICL (Integrated Complete Likelihood) criterion for model selection because it is especially relevant for the clustering purpose [32]. Indeed, it permits a

trade-off between the model specification and the component overlaps. Thus, it avoids models with high-overlapping classes and results are robust to the model misspecification. Due to the number of competing models, an exhaustive approach computing the ICL for each model cannot be used. Therefore, for each possible number of classes, model selection by maximization of the ICL is achieved by a MCMC method. The stationary distribution of this algorithm is proportional to the ICL. For the selected model, the maximum likelihood estimates are obtained by an EM algorithm [33,34].

In this study, we extended the mixture model of dependency blocks developed for categorical data [27] to ordinal data having the same number of levels. This extension was made by imposing constraints on the maximum dependency distribution to consider the order between the levels of a variable.

The algorithms of model selection and parameter estimation are implemented in the R package ClustOrd downloadable at https://github.com/masedki/ClustOrd.

RESULTS

Description of the 24 variables of interest (domestic cleaning tasks and cleaning products/types of spray used) is presented in Figure 1. The most frequent domestic tasks were using a washing machine and toilet bowl cleaning; the most frequently used cleaning products were "liquid cleaning products" and bleach; and the most frequently used sprays were air-refreshing sprays and windows/mirror sprays.

Strong dependencies were observed between the 24 variables. Indeed, Cramer's V values between the cleaning tasks/products variables ranged from 0.02 to 0.59 with a mean of 0.13 and a standard deviation of 0.12. Although we could expect some within-class dependencies to occur

because of the strong dependencies observed between the 24 variables, we first applied a classical latent class model, which assumes within-class independence. According to the ICL criterion, the best latent class model was a 9-class model (ICL=-291052.8), but the 10-class model obtained a close value (ICL=-291075.9). Thus, the selection of the number of classes remained uncertain with the latent class analysis. Although the number of classes could also be selected by non-statistical methods, e.g., by evaluating the practical usefulness of the resulting latent classes [35], that strategy seemed inappropriate in the context of a complex dataset with strong dependencies across variables, and without *a priori* hypotheses regarding the expected classes.. We expect that the mixture model of dependency blocks, which considers the within-class dependencies, better fits the data distribution and requires fewer components than the classical latent class model, facilitating the interpretation of the clustering results.

Mixture model of dependency blocks

The ICL criterion selected 7 classes. The mean class membership probability range for the seven classes was 0.78-0.93 (overall: 0.83). This model resulted in an allocation of the variables into 10 blocks. The first three blocks ("General cleaning", "Dusting /Sweeping /Hoovering" and "Humid cleaning") were related to essential cleaning tasks, while the next two blocks (“General purpose cleaning products” and “Bleach”) were related to general purpose cleaning products. The four remaining blocks were related to more specific tasks ("Other household tasks", ”Polishing/waxing” and "Windows/mirrors cleaning") or products ("Chemical products" and "Sprays").

The seven classes were summarized using the posterior mode of each variable and its probability, presented for the 24 variables ordered by block (Table 1). For instance, women in the class labelled “Very sparse cleaning” had 49% probability to have the level “<1 day/week”

(mode) for the variable "Cleaning at home", while women in the class labelled "Very Frequent general cleaning" had 90% probability to have the level "4-7 days/week" (mode) for this variable. More detailed results, with posterior probabilities for each level of each variable, by class, are presented in the appendix (Figures E1-E7).

Overall, the classes can be described as follows:

- Very sparse cleaning ($\pi_k = 0.05$): the class grouped women who did household cleaning tasks and used cleaning products very unfrequently.
- Sparse cleaning ($\pi_k = 0.15$): the class grouped women who did household tasks and used cleaning products unfrequently (but more often than the women of the previous class).
- Medium cleaning($\pi_k = 0.10$): the class grouped women who did cleaning tasks and used cleaning products at an intermediate frequency.
- Frequent general cleaning ($\pi_k = 0.28$): the class grouped women who had a high frequency of general cleaning tasks general cleaning, dusting/sweeping/hoovering, humid cleaning), a moderate use of general purpose cleaning products (eg, general purpose products, bleach, window/mirror), and a low use of chemical products and sprays.
- Frequent use of products ($\pi_k = 0.21$): the class grouped women who had a high frequency of cleaning tasks (general cleaning, dusting/sweeping/hoovering, humid cleaning) and use of cleaning products (bleach, polishing/waxing, windows/mirror, chemicals, sprays).
- Very frequent general cleaning ($\pi_k = 0.13$): the class grouped women who had a very high frequency of household tasks (general cleaning, dusting/sweeping/hoovering, humid cleaning), a moderate to high use of general cleaning products (eg, bleach) and a medium use of chemical products and a low use of sprays.

- Very frequent use of products ($\pi_k = 0.08$): the class grouped women who had a very high frequency of cleaning tasks and use of cleaning products, especially bleach, chemical products, and sprays.

The parameterρ_{kb} , which reflects the strength of the intra-class dependencies between variables grouped into the same block under each class, presented in Table 2. Three blocks obtained large values of ρ_{kb} for each class: "Polishing/waxing" block, "Windows/mirrors cleaning" block and "General cleaning habits" block. These high intra-class dependencies support the model choice. For most blocks, the strongest intra-class dependencies between variables were observed for the “very sparse cleaning” class, suggesting that in this class, a low frequency of cleaning tasks and use of products was consistently reported for variables within each block. The lowest intra-class dependencies were often observed in the “frequent use of products” and “very frequent general cleaning” classes, suggesting that frequency of cleaning tasks and use of products may be more heterogeneous within a block for women in these classes.

The ARI between the classifications obtained by the 7-class mixture of dependency blocks and the classification provided by the 7-class (respectively 9-class) traditional latent class model were 0.55 (0.56), indicating that the mixture model of dependency blocks and the traditional latent class model provide different partitions. Thus, modeling the dependencies within the component through the mixture model of dependency blocks both limited the number of classes and impacted the resulting classification.

Discriminative properties of the classes with regards to socio-demographic characteristics

To evaluate the discriminative properties of the classes produced by the selected model, we studied the differences of socio-demographic characteristics across the seven household cleaning classes (Table 3). Generally, women assigned to a class with a lower frequency of cleaning tasks and use of cleaning products (classes named very sparse cleaning and sparse cleaning), compared to women assigned to classes with a higher frequency of cleaning tasks and use of cleaning products (classes named "very frequent general cleaning" and "very frequent use of product"), were older (mean age: 75.5 and 70.8 *vs.* 69.6 and 69.4), had a higher education level (45% and 52% *vs*. 24% and 26%), and had more often household help (98% and 70% *vs*. 17% and 33%), with $p<0.001$ for all pairwise comparisons. In particular, women in the class named "very sparse cleaning" were much older (mean: 75.5 years, standard deviation: 6.9) than other women (mean: 69.8 years, standard deviation: 6.1; $p<0.001$) and almost all of them had household help (98%). The classes named "sparse cleaning" and "medium cleaning" were similar regarding household help (70% and 70%) but could be distinguished by education level (higher level: 52% and 45% respectively, $p<0.001$). Interestingly, women assigned to the class named "very frequent use of products" had more often household help (33%) than the class named "very frequent general cleaning" (17%, $p<0.001$), which had a more moderate use of products.

DISCUSSION

We used a mixture model of dependency blocks to identify household cleaning patterns in a large population of French women. Characterization of cleaning patterns required a clustering model allowing for ordinal and possibly highly dependent variables, and thus not relying on the conditional independence assumption. Our novel approach, which models the within-class dependencies, was a useful alternative to the classical latent class model for this purpose.

Strength of the mixture model of dependency blocks

The main strength of the mixture model of dependency blocks over a classical latent class model is to consider intra-class dependencies, rather than assuming conditional independence. A typical way to circumvent the issues due to the violated assumption of conditional independence is to reduce the number of variables used in the clustering. This preliminary step aims at selecting variables so as to avoid having several highly correlated variables or variables representing similar dimension in the latent class analysis. Usually this selection is based on multiple correspondence analysis results interpretation (data-driven approach) and/or expert knowledge (hypothesis-driven approach), but often involves arbitrary decisions [8]. Furthermore, this selection step may become extremely complicated (if feasible) as the numbers of variables of interest and modalities increases. The lack of objective criteria in this selection step also limits reproducibility.

Our approach replaces this first step by integrating the modeling of within-class dependencies in the clustering model. The model groups similar or correlated variables into conditionally independent blocks. The block distribution is a bi-component mixture of independence and maximum dependency distributions. This specific distribution of the blocks allows summarizing the conditional dependencies of the variables with only one continuous parameter: the proportion of the maximum dependency distribution. Thus, besides dealing with within-class dependency issues, the model provides a useful summary of the data. For instance, the class description can be based on the distribution of participants over 10 blocks, rather than over 24 variables. Examination of the within-block dependencies indicates the highest potential co-linearity between variables. This information is useful, for instance, to determine which

variables should not be studied independently when investigating associations with health outcomes [15,36]; or to improve future questionnaires by avoiding redundancy in the questions.

The ICL criterion can be used for model selection (*i.e.* number of classes and allocation of the variables into blocks). Finally, this approach allows clustering datasets with missing values by assuming that values are missing at random.

Other methods have been proposed for relaxing the conditional independence assumption in finite mixture models. However, these methods either require a large number of parameters which leads to some stability problems [37], or use continuous latent traits variables which renders class interpretation difficult [38]. Moreover, they are not implemented in an R package.

Relevance of the resulting classification

Validity of clustering models is in part evaluated in terms of relevance and interpretability of the resulting classification. The relevance of the resulting blocks can be illustrated by several examples. First, the blocks "Polishing/waxing" and "Windows/mirrors cleaning" were both composed of (i) a question regarding a specific task and (ii) a question regarding a product used specifically for this task, while the variables belonged to different subgroups of questions in the questionnaire. Second, the model grouped most questions about the use of sprays in the "Spray block" (except glass cleaning spray). The identification of this block supports the existence of underlying habits regarding the choice of cleaning product presentation (spray form) regardless of the use (e.g. furniture, floor). This block is of great interest since studies have found associations between the use of spray and asthma risk [19,20] and heart rate variability [18]. Other blocks also appeared to be relevant (e.g., "Chemical products", "Humid cleaning"). Furthermore, it is interesting to note that bleach, the use of which may vary across social groups [39] regardless of other general cleaning habits, resulted in a single-variable block, indicating that

this variable was not correlated with the other ones, conditionally on class. The seven resulting classes also appeared relevant as they could be distinguished by the intensity of general cleaning tasks and the intensity of use of specific products, and for some classes distinction according to the type of products used (e.g., specific chemicals vs. sprays). Finally, the seven classes varied markedly in terms of age, education level, and household help. These differences underline the discriminative properties of the classes, which also support the model validity [6,40]. The study sample (nested case-control study on asthma in a population of elderly women) is not representative of the general French population of elderly women, and thus profiles identified in the current study may differ from profiles that exist in the general population. In future work, the classification resulting from the mixture model of dependency blocks will be used to study associations between domestic cleaning patterns and asthma outcomes.

To our knowledge, this paper presents the first attempt to identify discrete household cleaning patterns using questionnaire data. Clustering methods such as latent class analysis have not been used in this context. Previous smaller studies with a similar household cleaning questionnaire used principal component analysis, another data-reduction method, to identify domestic exposure patterns in women [20,21] (three to four factors were identified, e.g., "essential tasks", "chemical products") or to derive a composite score variable for spray use [18,21]. Similar approaches (principal component analysis or factor analysis) have been used to characterize indoor air or various indoor exposures in households [41–43], or disinfectant exposures among healthcare workers [44]. However, principal component analysis is a variance-based approach, so its use to model ordinal data with few modalities, such as the household cleaning data available in the present study, is less appropriate.

Limits

The mixture model of dependency blocks requires the assumption that the allocation of the variables into blocks is the same for all classes. Relaxing this assumption introduce a lack of identifiability which can prevent the model interpretation. In addition, by considering a grouping of the variables into blocks, the approach implies a computationally intensive step of model selection. Thus, our analysis required 10 days of computation on a 48-(3.00GHz) cores.

Conclusion

Epidemiological research, in particular with the application of the exposome concept, orients towards higher-dimension datasets, increased complexity, and a growing need for methods providing accurate and interpretable data summary. Clustering approaches offer interesting opportunities to account for these multidimensional and complex aspects. The mixture model of dependency blocks presented in this paper, allowing modelling of within-class dependencies, is a useful alternative to the classical latent class model, as demonstrated by an application to the clustering of household cleaning exposures.

Funding: This work was supported by a grant of The Institut pour la Recherche en Santé Publique (IRESP), and of the joint help of Direction Générale de la Santé (DGS), Mission recherche de la Direction de la recherche, des études, de l'évaluation et des statistiques (Mire-DREES), Caisse nationale d'assurance maladie des travailleurs salariés (CNAMTS), Régime Social des Indépendants (RSI) & Caisse nationale de solidarité pour l'autonomie (CNSA). The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013) under REA grant agreement n. PCOFUND-GA-2013-609102, through the PRESTIGE programme coordinated by Campus France. The E3N cohort was funded by the MGEN, the Ligue contre le Cancer, Gustave Roussy Institute, and the Institut National de la Santé et de la Recherche Médicale (Inserm).

Acknowledgements: The authors would like to thank the E3N team and especially M Fangon, M Niravong, LA Hoang, M Valdenaire, S Eltaief, R Gomes, F Wilm, C Kernaleguen, W Tello, C Laplanche, P Gerbouin-Rérolle, R Chaït, G Esselma and F Clavel-Chapelon (Inserm U1018 CESP, Villejuif, France) for the implementation and management of the E3N study. They are indebted to all the participants, without whom the study would not have been possible, for their high involvement in the E3N study. The authors thank Raphaëlle Varraso and Nicole Le Moual (Inserm U1168 VIMA, Villejuif, France) for the E3N-Asthma study implementation (respiratory and household cleaning survey), and for their helpful advice during the conduct of the study and manuscript preparation.

References

[1] Siroux V, Garcia-Aymerich J. The investigation of asthma phenotypes. Curr Opin Allergy Clin Immunol 2011;11:393–9.

[2] Basagaña X, Barrera-Gómez J, Benet M, Antó JM, Garcia-Aymerich J. A framework for multiple imputation in cluster analysis. Am J Epidemiol 2013;177:718–25.

[3] Goodman LA. Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika 1974;61:215–31.

[4] van Smeden M, Naaktgeboren CA, Reitsma JB, Moons KGM, de Groot JAH. Latent Class Models in Diagnostic Studies When There is No Reference Standard--A Systematic Review. Am J Epidemiol 2013;179:423–31.

[5] Böhnke JR, Croudace TJ. Factors of psychological distress: clinical value, measurement substance, and methodological artefacts. Soc Psychiatry Psychiatr Epidemiol 2015;50:515–24.

[6] Siou G Lo, Yasui Y, Csizmadi I, McGregor SE, Robson PJ. Exploring Statistical Approaches to Diminish Subjectivity of Cluster Analysis to Derive Dietary Patterns The Tomorrow Project. Am J Epidemiol 2011;173:956–67.

[7] Varraso R, Garcia-Aymerich J, Monier F, Le Moual N, De Batlle J, Miranda G, et al. Assessment of dietary patterns in nutritional epidemiology: principal component analysis compared with confirmatory factor analysis. Am J Clin Nutr 2012;96:1079–92.

[8] Billionnet C, Sherrill D, Annesi-Maesano I. Estimating the Health Effects of Exposure to Multi-Pollutant Mixture. Ann Epidemiol 2012;22:126–41.

[9] Gass K, Klein M, Chang HH, Flanders WD, Strickland MJ. Classification and regression trees for epidemiologic research: an air pollution example. Environ Heal 2014;13:17.

[10] Virtanen M, Vahtera J, Head J, Dray-Spira R, Okuloff A, Tabak AG, et al. Work Disability among Employees with Diabetes: Latent Class Analysis of Risk Factors in Three Prospective Cohort Studies. PLoS One 2015;10:e0143184.

[11] Leventhal AM, Huh J, Dunton GF. Clustering of modifiable biobehavioral risk factors for chronic disease in US adults: a latent class analysis. Perspect Public Health 2014;134:331–8.

[12] Cheung YK, Yu G, Wall MM, Sacco RL, Elkind MSV, Willey JZ. Patterns of leisure-time physical activity using multivariate finite mixture modeling and cardiovascular risk factors in the Northern Manhattan Study. Ann Epidemiol 2015;25:469–74.

[13] Kauffmann F, Demenais F. Gene-environment interactions in asthma and allergic diseases: challenges and perspectives. J Allergy Clin Immunol 2012;130:1229–40.

[14] Slama R, Vrijheid M. Some challenges of studies aiming to relate the Exposome to human health. Occup Environ Med 2015;72:383–4.

[15] Patel CJ, Ioannidis JPA. Placing epidemiological results in the context of multiplicity and typical correlations of exposures. J Epidemiol Community Health 2014;68:1096–100.

[16] Svingen T, Vinggaard AM. The risk of chemical cocktail effects and how to deal with the issue. Proc Natl Acad Sci U S A 2016;70:322–3.

[17] Siracusa A, De Blay F, Folletti I, Moscato G, Olivieri M, Quirce S, et al. Asthma and exposure to cleaning products - a European Academy of Allergy and Clinical Immunology task force consensus statement. Allergy 2013;68:1532–45.

[18] Mehta AJ, Adam M, Schaffner E, Barthelemy JC, Carballo D, Gaspoz JM, et al. Heart Rate Variability in Association with Frequent Use of Household Sprays and Scented Products in SAPALDIA. Env Heal Perspect 2012;120:958–64.

[19] Zock JP, Plana E, Jarvis D, Antó JM, Kromhout H, Kennedy SM, et al. The use of

household cleaning sprays and adult asthma: an international longitudinal study. Am J Respir Crit Care Med 2007;176:735–41.

[20] Le Moual N, Varraso R, Siroux V, Dumas O, Nadif R, Pin I, et al. Domestic use of cleaning sprays and asthma activity in females. Eur Respir J 2012;40:1381–9. doi:10.1183/09031936.00197611.

[21] Bédard A, Varraso R, Sanchez M, Clavel-Chapelon F, Zock JP, Kauffmann F, et al. Cleaning sprays, household help and asthma among elderly women. Respir Med 2014;108:171–80.

[22] Herr M, Just J, Nikasinovic L, Foucault C, Le Marec AM, Giordanella JP, et al. Influence of host and environmental factors on wheezing severity in infants: Findings from the PARIS birth cohort. Clin Exp Allergy 2012;42:275–83.

[23] McLachlan GJ, Peel D. Finite mixutre models. Wiley-Interscience, New York: Wiley Series in Probability and Statistics: Applied Probability and Statistics; 2000.

[24] Little RJA, Rubin DB. Statistical analysis with missing data. John Wiley & Sons; 2014.

[25] Govaert G. Data analysis. vol. 136. Wiley. com; 2010.

[26] Van Hattum P, Hoijtink H. Market Segmentation Using Brand Strategy Research: Bayesian Inference with Respect to Mixtures of Log-Linear Models. J Classif 2009;26:297–328.

[27] Marbac M, Biernacki C, Vandewalle V. Model-Based Clustering for Conditionally Correlated Categorical Data. J Classif 2015;32:1–31.

[28] Clavel-Chapelon F. Cohort Profile: The French E3N Cohort Study. Int J Epidemiol 2015;44:801–9.

[29] Sanchez M, Varraso R, Bousquet J, Clavel-Chapelon F, Pison C, Kauffmann F, et al. Perceived 10-year change in respiratory health: Reliability and predictive ability. Respir

Med 2015;109:188–99.

[30] Cramér H. Mathematical Methods of Statistics. Princeton University Press; 1946.

[31] Hubert L, Arabie P. Comparing partitions. J Classif 1985;2:193–218.

[32] Biernacki C, Jacques J. Model-based clustering of multivariate ordinal data relying on a stochastic binary search algorithm. Stat Comput 2015:1–15.

[33] Dempster, A.P. and Laird, N.M. and Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B 1977;39:1–38.

[34] McLachlan GJ, Krishnan T. The EM algorithm. Wiley-Interscience, New York: Wiley Series in Probability and Statistics: Applied Probability and Statistics; 1997.

[35] Muthén B, Muthén L. Integrating person-centred and variable-centred analysis: Growth mixture modeling with latent trajectory classes. Alcohol Clin Exp Res 2000;24:882–91.

[36] Agier L, Portengen L, Chadeau-Hyam M, Basagaña X, Giorgis-Allemand L, Siroux V, et al. A Systematic Comparison of Linear Regression-Based Statistical Methods to Assess Exposome-Health Associations. Environ Health Perspect 2016;124:1848–56.

[37] Meila M, Jordan MI. Learning with mixtures of trees. J Mach Learn Res 2001;1:1–48.

[38] Gollini I, Murphy TB. Mixture of latent trait analyzers for model-based clustering of categorical data. Stat Comput 2014;24:569–88.

[39] Zock JP, Plana E, Antó JM, Benke G, Blanc PD, Carosso A, et al. Domestic use of hypochlorite bleach, atopic sensitization, and respiratory symptoms in adults. J Allergy Clin Immunol 2009;124:731–738 e1.

[40] Siroux V, Basagaña X, Boudier A, Pin I, Garcia-Aymerich J, Vesin A, et al. Identifying adult asthma phenotypes using a clustering approach. Eur Respir J 2011;38:310–7.

[41] Qian Z, Zhang J, Korn LR, Wei F, Chapman RS. Factor analysis of household factors: Are they associated with respiratory conditions in Chinese children? Int J Epidemiol

2004;33:582–8.

[42] Arif AA, Shah SM. Association between personal exposure to volatile organic compounds and asthma among US adult population. Int Arch Occup Environ Health 2007;80:711–9.

[43] Billionnet C, Gay E, Kirchner S, Leynaert B, Annesi-Maesano I. Quantitative assessments of indoor air pollution and respiratory health in a population-based sample of French dwellings. Environ Res 2011;111:425–34.

[44] Arif AA, Delclos GL. Association between cleaning-related chemicals and work-related asthma and asthma symptoms among healthcare professionals. Occup Env Med 2012;69:35–40.

Table 1. Description of the 7 classes by the mode of each variable and its probability

Block	Variables composing the block	Very sparse cleaning	Sparse cleaning	Medium cleaning	Frequent general cleaning	Frequent use of products	Very frequent general cleaning	Very Frequent use of products
General cleaning	Cleaning at home	<1 (49)	<1 (75)	<1 (48)	1-3 (79)	1-3 (71)	4-7 (90)	4-7 (51)
	Household cleaning	0 (69)	<1 (82)	<1 (64)	1-3 (78)	1-3 (80)	4-7 (69)	1-3 (53)
Dusting/sweeping/hoovering and rug beating*	Dusting/sweeping/hoovering and rug beating	0 (74)	<1 (88)	<1 (73)	1-3 (82)	1-3 (85)	4-7 (72)	1-3 (58)
Humid cleaning	Mopping	0 (78)	<1 (82)	<1 (75)	1-3 (67)	1-3 (82)	1-3 (53)	1-3 (61)
	Toilet bowl cleaning	0 (41)	<1 (56)	1-3 (52)	1-3 (64)	1-3 (66)	4-7 (57)	4-7 (61)
General purpose cleaning products	Liquid cleaning products	0 (71)	<1 (71)	<1 (68)	<1 (48)	1-3 (60)	<1 (32)	1-3 (48)
	Perfumes	0 (75)	0 (68)	<1 (67)	0 (67)	<1 (58)	0 (52)	1-3 (49)
Bleach*	Bleach	<1 (54)	<1 (74)	<1 (57)	<1 (62)	<1 (54)	<1 (46)	1-3 (49)
Other household tasks	Washing by hand	0 (65)	0 (53)	<1 (49)	0 (50)	<1 (54)	0 (43)	0 (41)
	Washing by machine	1-3 (51)	1-3 (65)	1-3 (74)	1-3 (80)	1-3 (87)	1-3 (73)	1-3 (71)
	Handiwork	0 (66)	<1 (47)	<1 (54)	<1 (44)	<1 (53)	<1 (42)	<1 (41)
Polishing/waxing	Floor/furniture polishing/waxing/shampooing	0 (99)	0 (68)	<1 (59)	<1 (55)	<1 (84)	<1 (71)	<1 (63)
	Polish/waxes	0 (93)	0 (67)	<1 (66)	0 (51)	<1 (85)	<1 (69)	<1 (61)
Windows/mirror cleaning	Windows/mirrors cleaning	0 (92)	<1 (68)	<1 (83)	<1 (87)	<1 (90)	<1 (83)	<1 (67)
	Windows/mirrors sprays	0 (84)	0 (57)	<1 (75)	0 (52)	<1 (73)	<1 (49)	<1 (51)
Chemical products	Ammonia	0 (97)	0 (92)	0 (79)	0 (91)	0 (76)	0 (87)	0 (68)
	Acids	0 (61)	<1 (67)	<1 (71)	<1 (65)	<1 (75)	<1 (63)	<1 (42)
	Stain removers	0 (67)	<1 (51)	<1 (79)	0 (50)	<1 (85)	<1 (59)	<1 (60)
Sprays	Furniture sprays	0 (94)	0 (82)	<1 (56)	0 (81)	<1 (56)	0 (65)	<1 (43)
	Floor cleaning sprays	0 (98)	0 (97)	0 (81)	0 (97)	0 (88)	0 (95)	0 (72)
	Degreasing/oven sprays	0 (87)	0 (80)	0 (49)	0 (80)	0 (51)	0 (69)	<1 (49)
	Air-refreshing sprays	0 (66)	0 (73)	<1 (46)	0 (75)	<1 (45)	0 (67)	<1 (33)
	Insecticide/pesticide/acaricide sprays	0 (66)	0 (61)	<1 (65)	0 (64)	<1 (66)	0 (57)	<1 (59)
	Other sprays	0 (94)	0 (92)	0 (66)	0 (94)	0 (70)	0 (88)	0 (51)

Variables are ordered by block. Data presented as posterior mode, i.e., level with the highest posterior probability (probability, expressed in %), for each variable. 0: never; <1: <1 day/week; 1-3: 1-3 days/week; 4-7: 4-7 days/week..* Single-variable blocks.

Table 2. Intra-class dependencies between variables regrouped into the same block under each class

Block	Variables composing the block	Very sparse cleaning	Sparse cleaning	Medium cleaning	Frequent general cleaning	Frequent use of products	Very frequent general cleaning	Very Frequent use of products
General cleaning	Cleaning at home; Household cleaning	0.40	0.43	0.38	0.26	0.15	0.45	0.39
Dusting/sweeping/hoovering and rug beating*	Dusting/sweeping/hoovering and rug beating	-	-	-	-	-	-	-
Humid cleaning	Mopping; Toilet bowl cleaning	0.23	0.24	0.15	0.06	0.18	0.18	0.19
General purpose cleaning products	Liquid cleaning products; Perfumes	0.28	0.12	0.21	0.13	0.10	0.09	0.11
Bleach*	Bleach	-	-	-	-	-	-	-
Other household tasks	Washing by hand; Washing by machine; Handiwork	0.13	0.02	0.01	0.01	0.02	0.03	0.04
Polishing/waxing	Floor/furniture polishing/waxing/shampooing; Polish/waxes	0.87	0.53	0.51	0.52	0.45	0.40	0.50
Windows/mirror cleaning	Windows/mirrors cleaning; Windows/mirrors sprays	0.64	0.23	0.40	0.25	0.39	0.19	0.41
Chemical products	Ammonia; Acids; Stain removers	0.36	0.10	0.11	0.08	0.04	0.07	0.10
Sprays	Furniture; Floor; Degreasing/Oven; Air-refreshing; Insecticide/Pesticide; Other sprays	0.38	0.18	0.04	0.16	0.02	0.14	0.04

Data presented are values of the parameter ρ_{kb} for each block and each class.* Single-variable blocks.

Table 3. Description of the socio-demographic characteristics according to classes

	Very sparse cleaning	Sparse cleaning	Medium cleaning	Frequent general cleaning	Frequent use of products	Very frequent general cleaning	Very Frequent use of products	P
Age, mean	75.5	70.8	70.3	69.4	68.3	69.6	69.4	<0.001
Household help, %	98	70	70	28	24	17	33	<0.001
Higher education level*, %	45	52	45	37	33	24	26	<0.001

* Completed at least 3 years of education after high school.

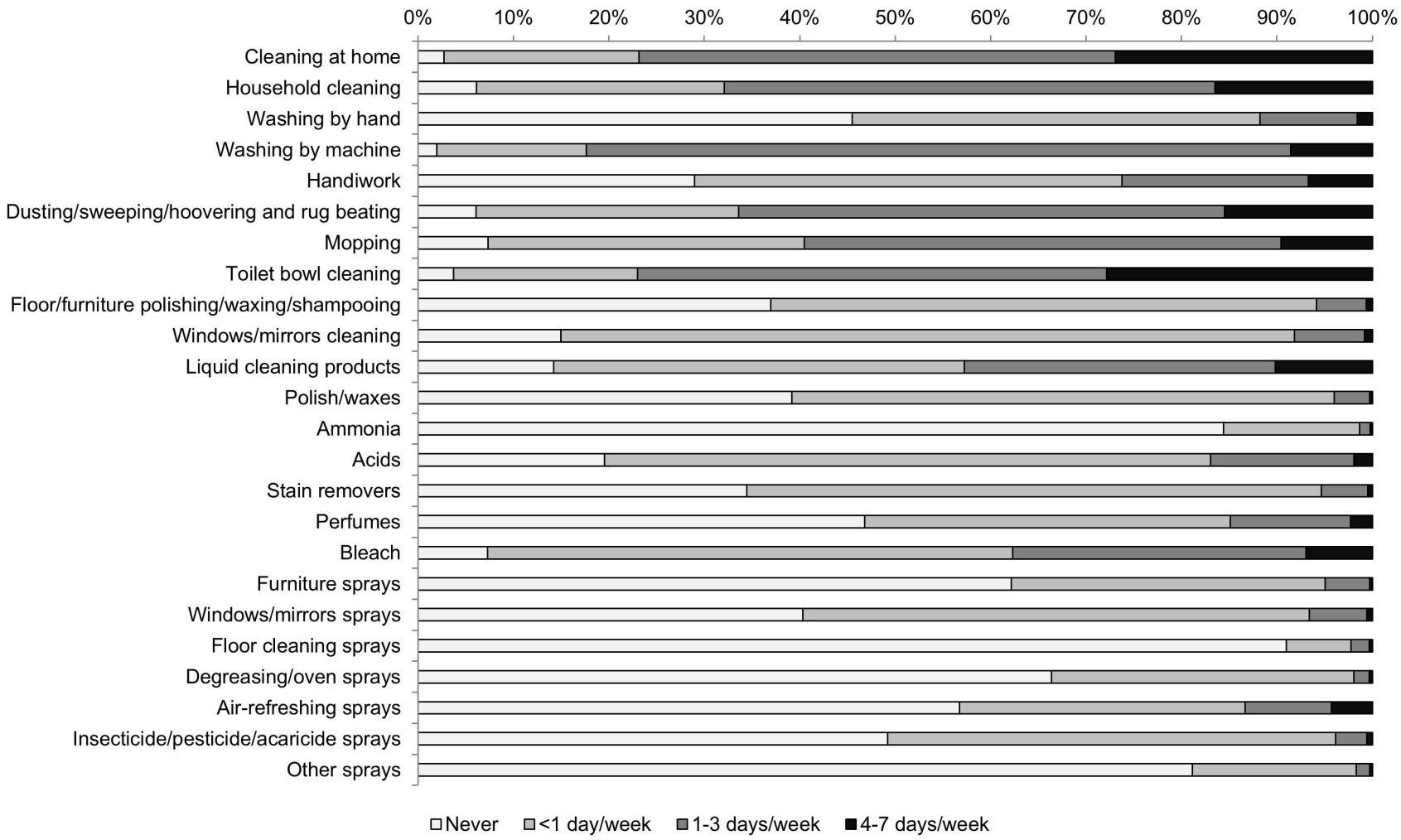
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Cleaning at home
Household cleaning
Washing by hand
Washing by machine
Handiwork
Dusting/sweeping/hoovering and rug beating
Mopping
Toilet bowl cleaning
Floor/furniture polishing/waxing/shampooing
Windows/mirrors cleaning
Liquid cleaning products
Polish/waxes
Ammonia
Acids
Stain removers
Perfumes
Bleach
Furniture sprays
Windows/mirrors sprays
Floor cleaning sprays
Degreasing/oven sprays
Air-refreshing sprays
Insecticide/pesticide/acaricide sprays
Other sprays
Never
<1 day/week
1-3 days/week
4-7 days/week

Figure 1. Description of the frequency of household cleaning tasks and use of cleaning products among the study participants (women of the Asthma-E3N study, France, 2011-2013).

Variables had 3% to 11% missing values.