Annals of Epidemiology xxx (2018) 1–7

Contents lists available at ScienceDirect

Annals of Epidemiology

journal homepage: www.annalsofepidemiology.org

Original article

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

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

[a] CREST, Ensai, Campus de KerLan, Bruz, France
[b] Université Paris-Sud and INSERM UMR 1181 B2PHI, Villejuif, France
[c] CESP, Inserm U1018, Université Paris-Sud, UVSQ, Université Paris-Saclay, 948052, Villejuif, France
[d] Gustave Roussy Institute, Villejuif, France
[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

ARTICLE INFO

Article history:
Received 23 November 2016
Accepted 15 May 2018
Available online xxx

Keywords:
Cluster analysis
Classification
Disinfectants
Household cleaning
Latent class

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 with 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.

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 multipollutant 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

* Corresponding author. Inserm UMRS 1168, VIMA- Aging and chronic diseases - Epidemiological and public health approaches, 16 avenue Paul Vaillant Couturier, 94807 Villejuif cedex, France.
E-mail address: orianne.dumas@inserm.fr (O. Dumas).

https://doi.org/10.1016/j.annepidem.2018.05.004
1047-2797/

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, that is, 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 nonobserved 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 article 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 (a French national health insurance plan covering mostly teachers) [28]. A total of 98,997 women aged 40–65 years were included at baseline and have been followed up approximately every 2 years. The current analysis uses data from a nested case-control study on asthma (Asthma-E3N) conducted in 2011–2013 [29]. A total of 7100 women with asthma and 14,200 age-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 (10 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 present 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 Adjusted Rand Index (ARI), 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, whereas 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 sociodemographic 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 carried out 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 present 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 parameter ρ_{kb} reflects the strength of the intraclass 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

Please cite this article in press as: Marbac M, et al., Patterns of cleaning product exposures using a novel clustering approach for data with correlated variables, Annals of Epidemiology (2018), https://doi.org/10.1016/j.annepidem.2018.05.004