Appendix

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

Generalities on finite mixtures

The data $x = (x_1, \ldots, x_n)$ are assumed to consist of n independent realizations of the random variable $X = (X_1, \ldots, X_n)$ which corresponds to J ordinal variables with L levels (for instance, the household cleaning dataset has J=24 variables with L=4 levels). We postulate that the observed heterogeneous population consists of K classes of homogeneous individuals where a class is defined by the subset of the individuals generated from the same distribution. Thus, modelling one observation requires two random variables: the categorical variable Z having K modalities which follows a multinomial distribution $M_k(\pi_1, \ldots, \pi_k)$ where $\pi_k = P(Z = k)$ is the marginal probability that an individual belongs to class K, and the multivariate ordinal variable $\boldsymbol{X}$ whose distribution is modelled conditionally on Z. In clustering, the realizations of $\boldsymbol{X}$ are observed while the realizations of Z are missing and should be estimated. Thus, the distribution of the observed variables is a mixture model with K classes defined by

$$P(X = x) = \sum_{k=1}^{K} \pi_k \, P(X = x \mid Z = k)$$

When the distributions of the mixture components are defined (*i.e* distributions of $\boldsymbol{X} \mid Z = k$), the probability that individual $\boldsymbol{x}$ belongs to class k is defined by

$$P(Z = k \mid X = x) = \frac{\pi_k P(X = x \mid Z = k)}{P(X = x)}$$

Hence, the clustering goal can be easily achieved by affecting an individual to the class that maximizes the posterior probability (*i.e.* the class k maximizing $P(Z = k \mid X = x)$). When data with missing values can be managed, the formula can be used by assuming that variables are missing at random [1]. In this case, distribution of the observed variables replaces $P(X = x \mid Z = k)$ by marginalization over the set of the possible values of the missing variables.

Within-class independence and within-class independence per blocks

The simplest way to cluster ordinal data is to use the latent class model which assumes within-class independence. Therefore,

$$P(X = x \mid Z = k) = \prod_{j=1}^{J} P\left(X_j = x_j \mid Z = k\right)$$

and the univariate variable $X_j \mid Z = k$ follows a multinomial distribution $M_L(\alpha_{kj1}, \ldots, \alpha_{kjL})$. Therefore, each class is interpreted with its univariate probabilities $P(X_j = l \mid Z = k)$ as the probability that an individual belonging to class k takes level l for variable j.

Although the within-class independence assumption is useful for modelling ordinal variables, it leads to severe biases when within-class dependencies occur [2]. Therefore, to cluster the household cleaning data, we propose to use an extension of the model of Marbac et al. [3], which relaxes this assumption. The model splits the variables into B within-class independent blocks:

$$P(X = x \mid Z = k) = \prod_{b=1}^{B} P\left(X_{\{b\}} = x_{\{b\}} \mid Z = k\right)$$

where $X_{\{b\}} = (X_j; j \in \Omega_b)$ corresponds to the subset of variables of block b and where Ω_b contains the indices of the variables of block b. We now detail the distribution of $X_{\{b\}} \mid Z = k$ which models intra-class dependencies between the variables of a block.

Specific block distribution

The block distribution is a mixture of the two extreme distributions according to the Cramer's V: the independence and the maximum dependency. The latter has been introduced for categorical variables but can be extended to ordinal data by imposing constraints for considering the order between the levels of the ordinal variables [3]. So, we introduce the binary random variable Y_b where $Y_b = 1$ indicates that the variables of block b follow the maximum dependency distribution while these variables follow the independence distribution if $Y_b = 0$.

Since $Y_b \mid Z = k$ follows a Bernoulli distribution $B(\rho_{kj})$ where $\rho_{kb} = P(Y_b = 1 \mid Z = k)$, the conditional distribution of $X_{\{b\}}$ is

$$P(X_{\{b\}} = x_{\{b\}} \mid Z = k) = (1 - \rho_{kb})P(X_{\{b\}} = x_{\{b\}} \mid Z = k, Y_b = 0) + \rho_{kb}P(X_{\{b\}} = x_{\{b\}} \mid Z = k, Y_b = 1)$$

Conditionally on $(Z = k, Y_b = 0)$, the block variables are independent, so

$$P(X_{\{b\}} = x_{\{b\}} \mid Z = k, Y_b = 0) = \prod_{j \in \Omega_b} P\left(X_j = x_j \mid Z = k, Y_b = 0\right).$$

Each univariate random variable $X_j \mid Z = k, Y_b = 0$ follows a multinomial distribution $M_L(\beta_{kj1}, \dots, \beta_{kjL})$ where β_{kjl} indicates the probability that the variable *j* takes the level *l* under the independence distribution for class *k*.

The maximum dependency constrains all the block variables to take the same level. In class *k* and block *k*, this level follows a multinomial distribution $M_L(\tau_{kj1}, \dots, \tau_{kjL})$. So,

$$P(X_{\{b\}} = x_{\{b\}} \mid Z = k, Y_b = 1) = \begin{cases} \tau_{kbl} \ if \ \forall j \in \Omega_b : x_j = l \\ 0 \ else \end{cases}$$

The parameter τ_{kbl} corresponds to the probability that all variables of block *b* take level *l* under the maximum dependency distribution of class *k*.

Model interpretation

The importance of each class is defined by its proportion. Moreover, the class *k* can be summarized by the univariate probability of the variables $P(X_j = l \mid Z = k)$, *i.e.* the probability that an individual takes level *l* for the variable X_j, conditionally on belonging to class *k* (often referred to as "posterior probabilities"). The probability is obtained from the model parameters by

$$P(X_j = l \mid Z = k) = (1 - \rho_{kb})\beta_{kjl} + \rho_{kb}\tau_{kbl} \ where \ b := \sigma_j$$

The interpretation of class *k* can be refined with the within-class dependencies which are mainly characterized by parameters $\rho_{k1}, \dots, \rho_{kB}$. Indeed, for block *b* of class *k*, ρ_{kb} is similar to a correlation coefficient between all variables assigned into block *b* since $0 < \rho_{kb} < 1$. Finally parameters $\tau_{kb1}, \ldots, \tau_{kbL}$ bring out the more linked modality association between the variables of block b under class k.

Parameter inference and model selection

For a known model, the maximum likelihood estimates can be obtained by an EM algorithm [4,5]. The model selection is performed via the Integrated Completed Likelihood (ICL) criterion [6] since it focuses on the goal of clustering. This criterion favors a model that provides partition with strong evidence since it makes a trade-off between the model evidence and the partitioning evidence. Since the model space is discrete, the search for the model that maximizes the ICL criterion is a combinatorial problem which can be circumvented by a Metropolis-Hastings algorithm [7] performing a random walk over the model space. The mode of its stationary distribution is located on the model that maximizes the ICL criterion since its unique invariant distribution is proportional to $exp\big(ICL(K, B, \sigma)\big)$.

Case of variables with different number of levels

The number of levels can be different between variables. Indeed, the mixture model of dependency blocks introduced for categorical data does not force the variables to have the same number of levels [3]. Here, the approach is an extension of this model to ordinal data. The main idea is to impose constraints on the maximum dependency distribution (one of the two distributions used to model one block of variables). Indeed, the relations between the levels of the variables of a block are monotone. If two variables have the same number of levels, this distribution implies a one-to-one relation between the levels of two variables. If the number of levels is not equal, this distribution implies a many-to-one relation between the levels (as defined by [3]).

References (Appendix)

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

[2] 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.

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

[4] 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.

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

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

[7] Robert CP, Casella G. Monte Carlo statistical methods. Springer Verlag; 2004.

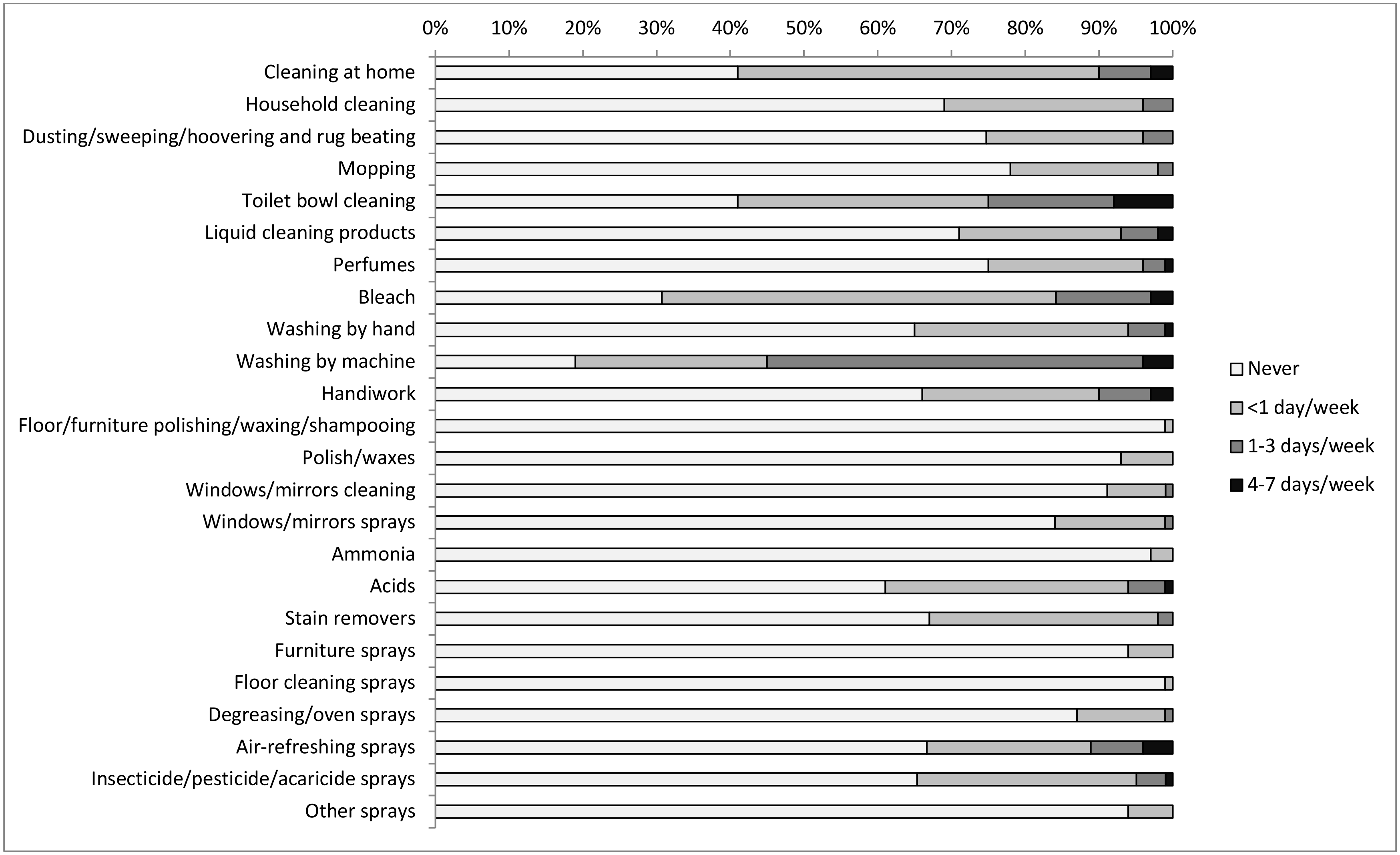

Figure E1. Description of the class “Very sparse cleaning”. Results presented as posterior probabilities of each variable.

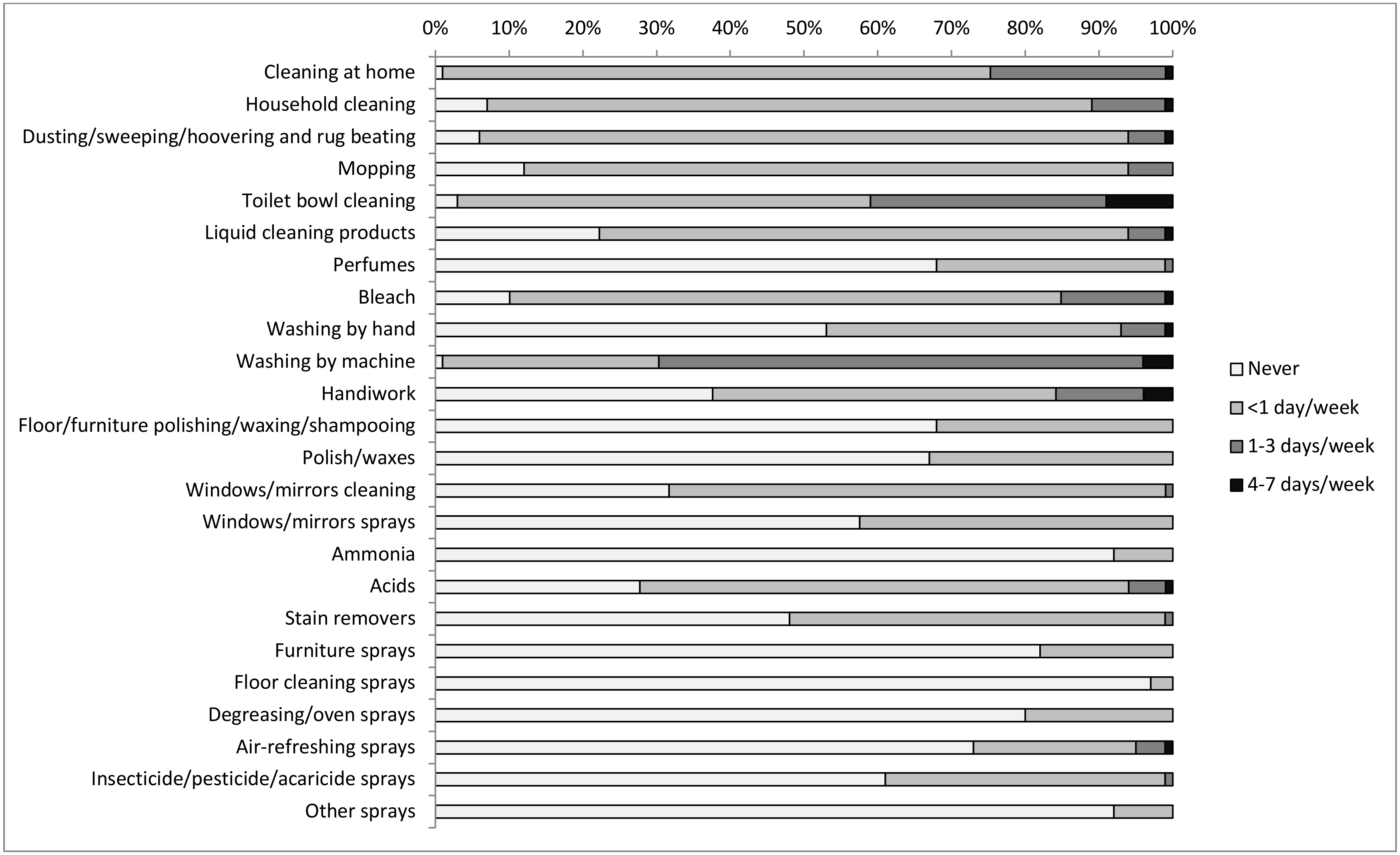

Figure E2. Description of the class “Sparse cleaning”. Results presented as posterior probabilities of each variable.

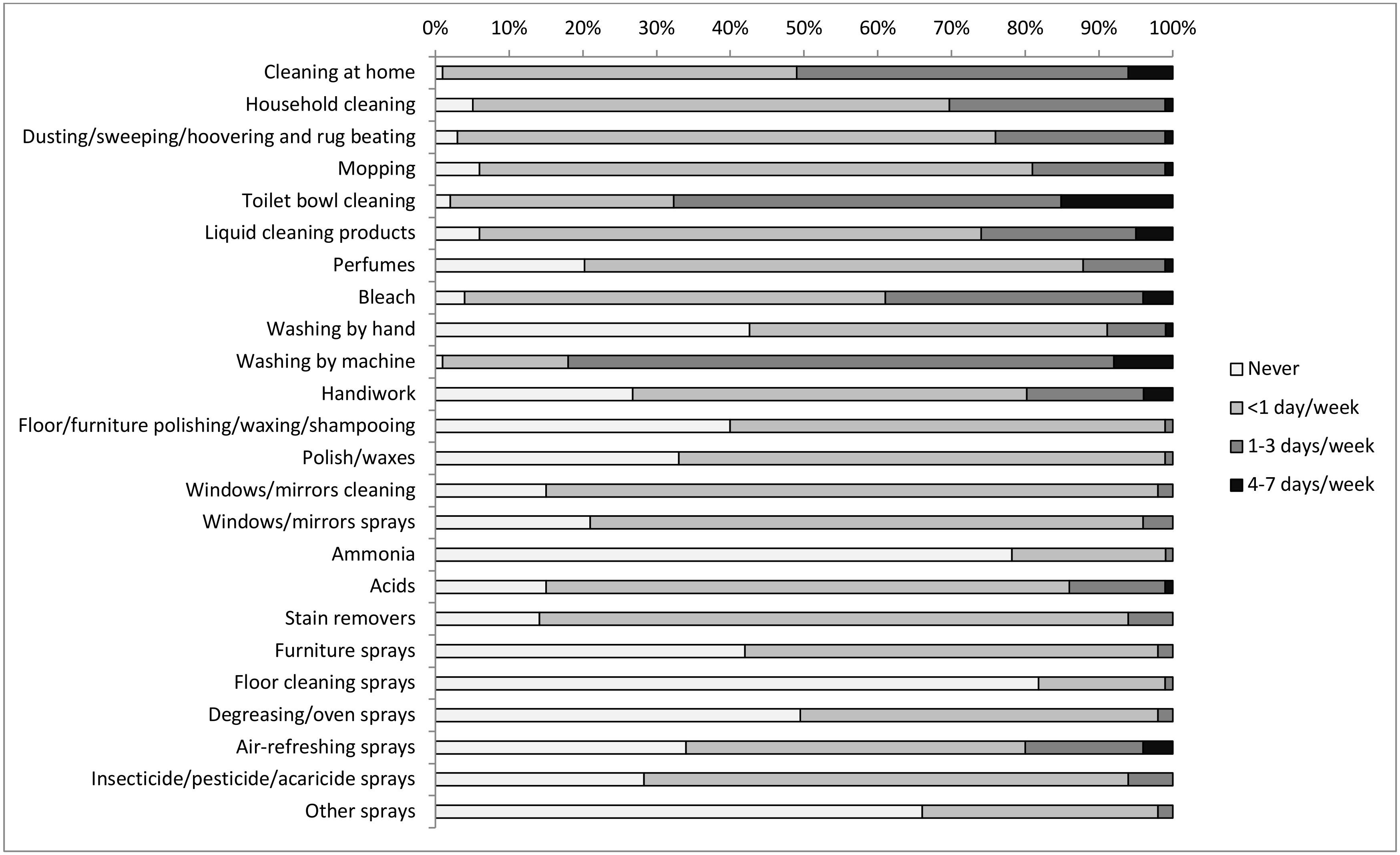

Figure E3. Description of the class “Medium cleaning”. Results presented as posterior probabilities of each variable.

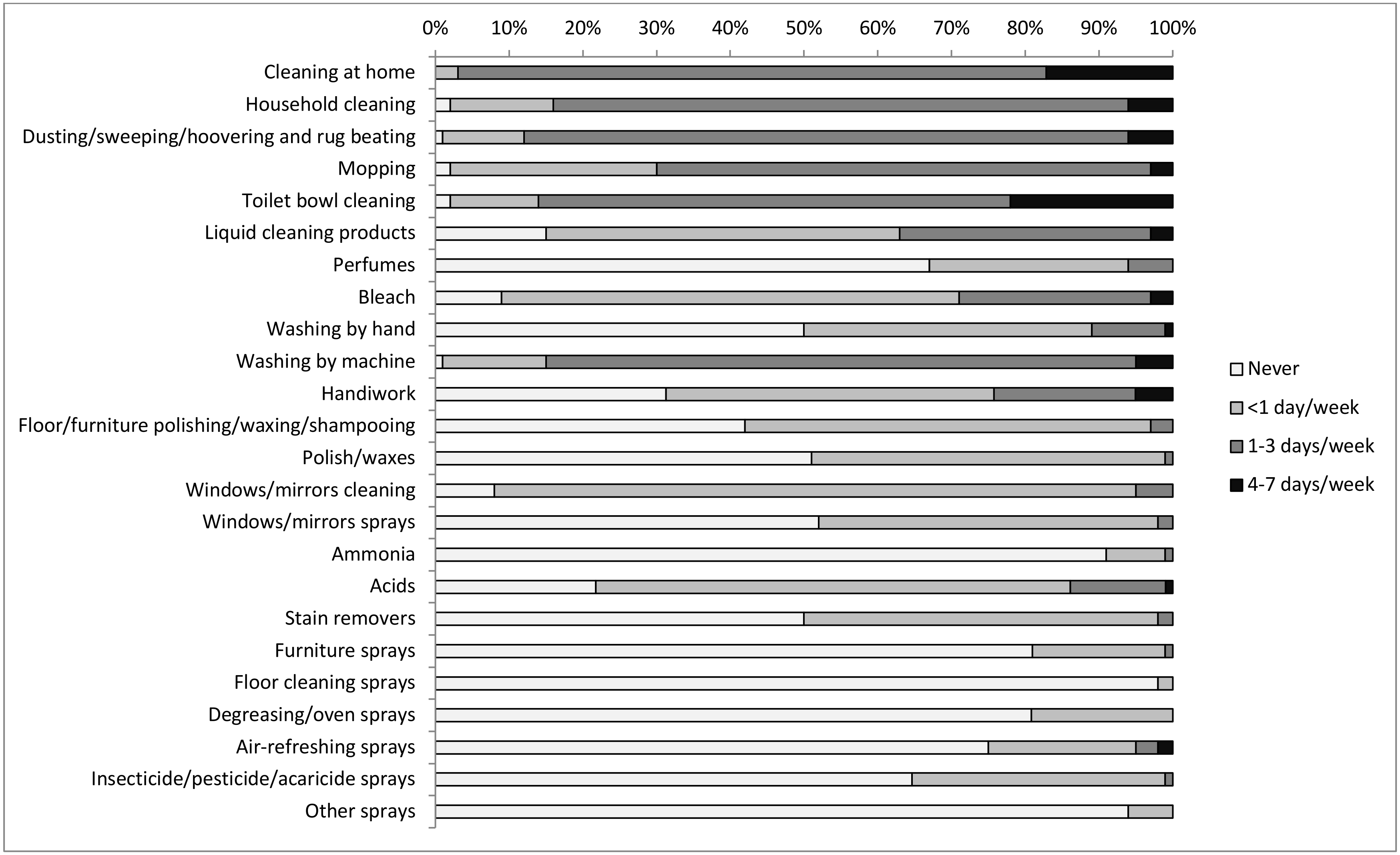

Figure E4. Description of the class “Frequent general cleaning”. Results presented as posterior probabilities of each variable.

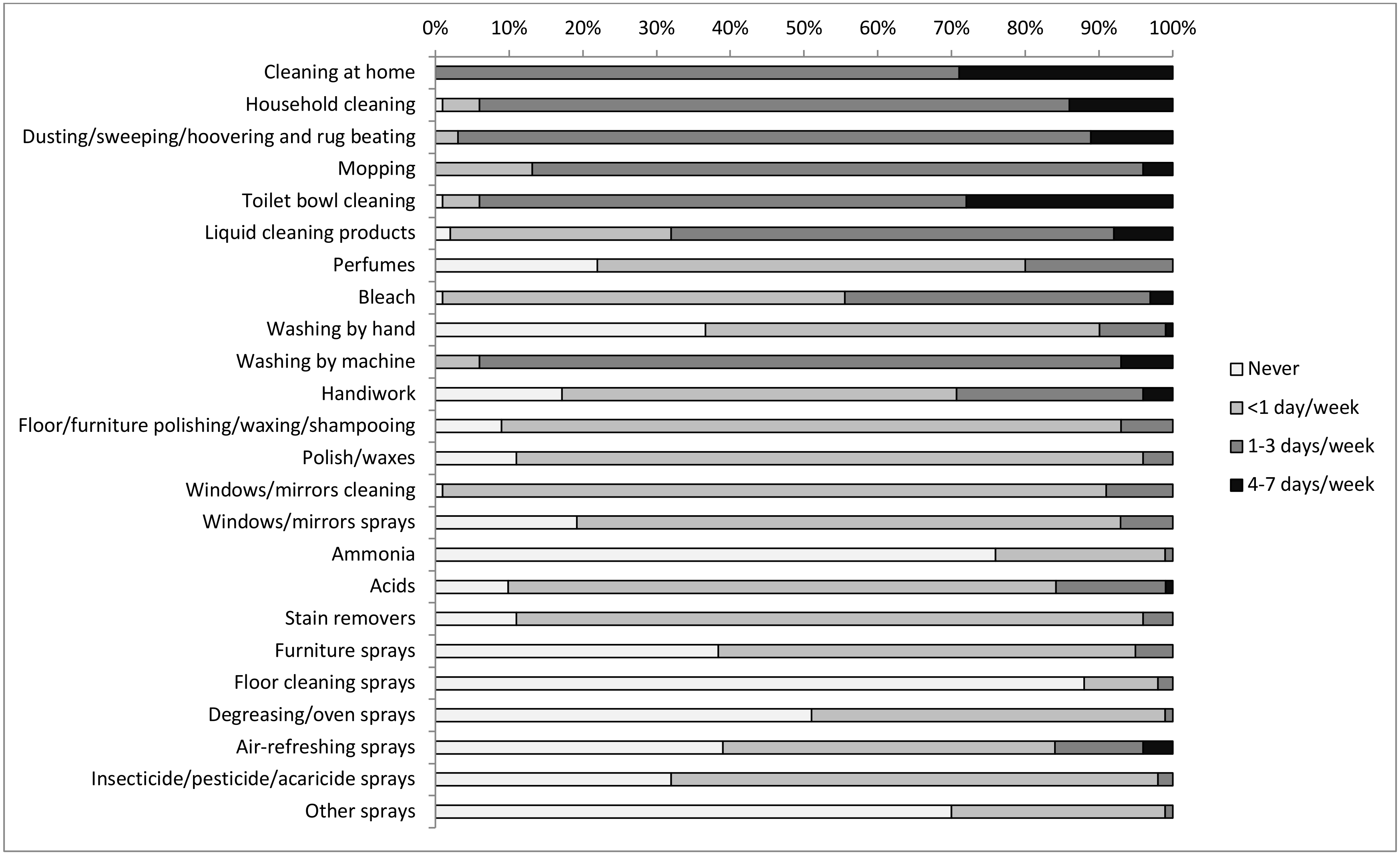

Figure E5. Description of the class “Frequent use of products”. Results presented as posterior probabilities of each variable.

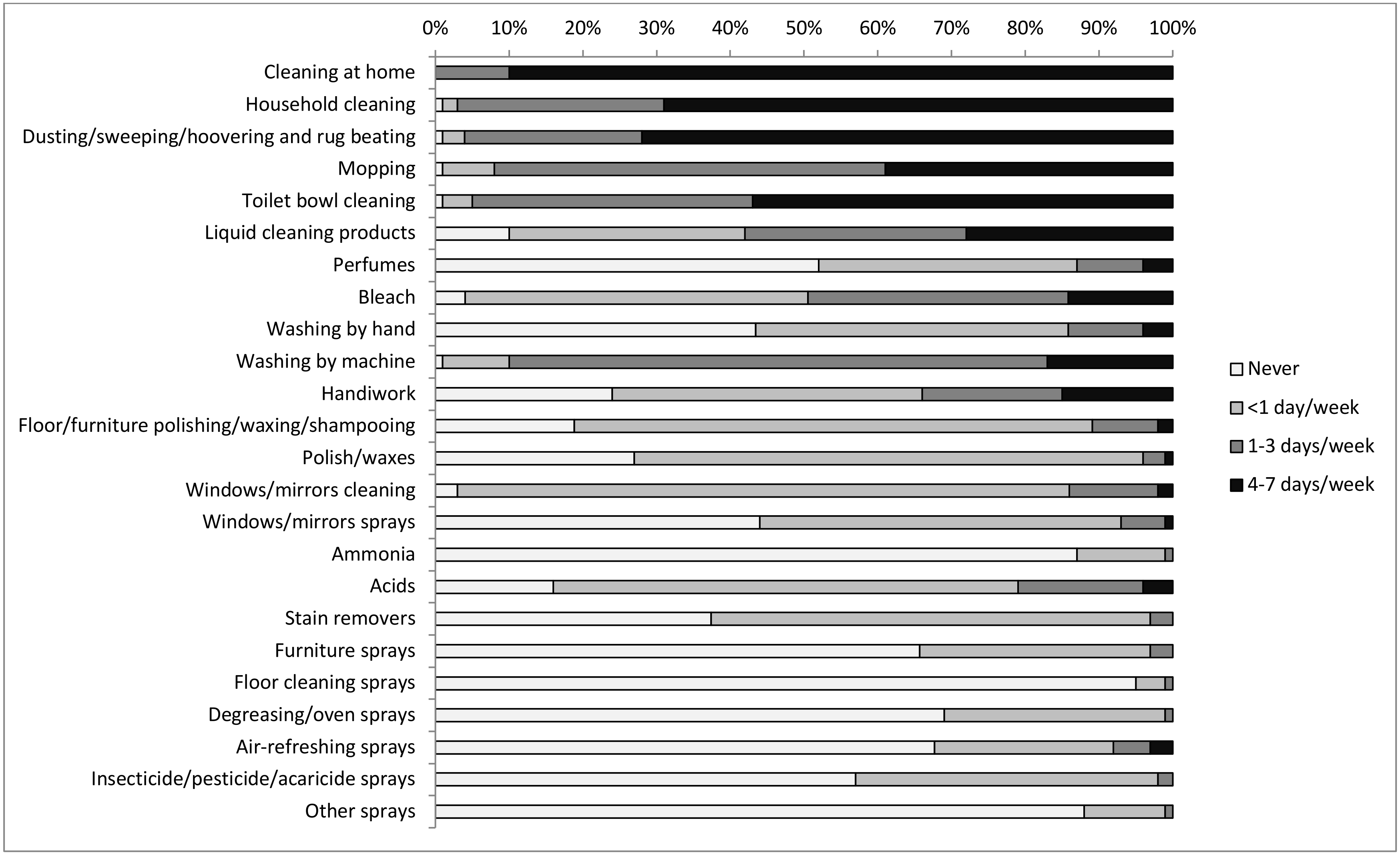

Figure E6. Description of the class “Very frequent general cleaning”. Results presented as posterior probabilities of each variable.

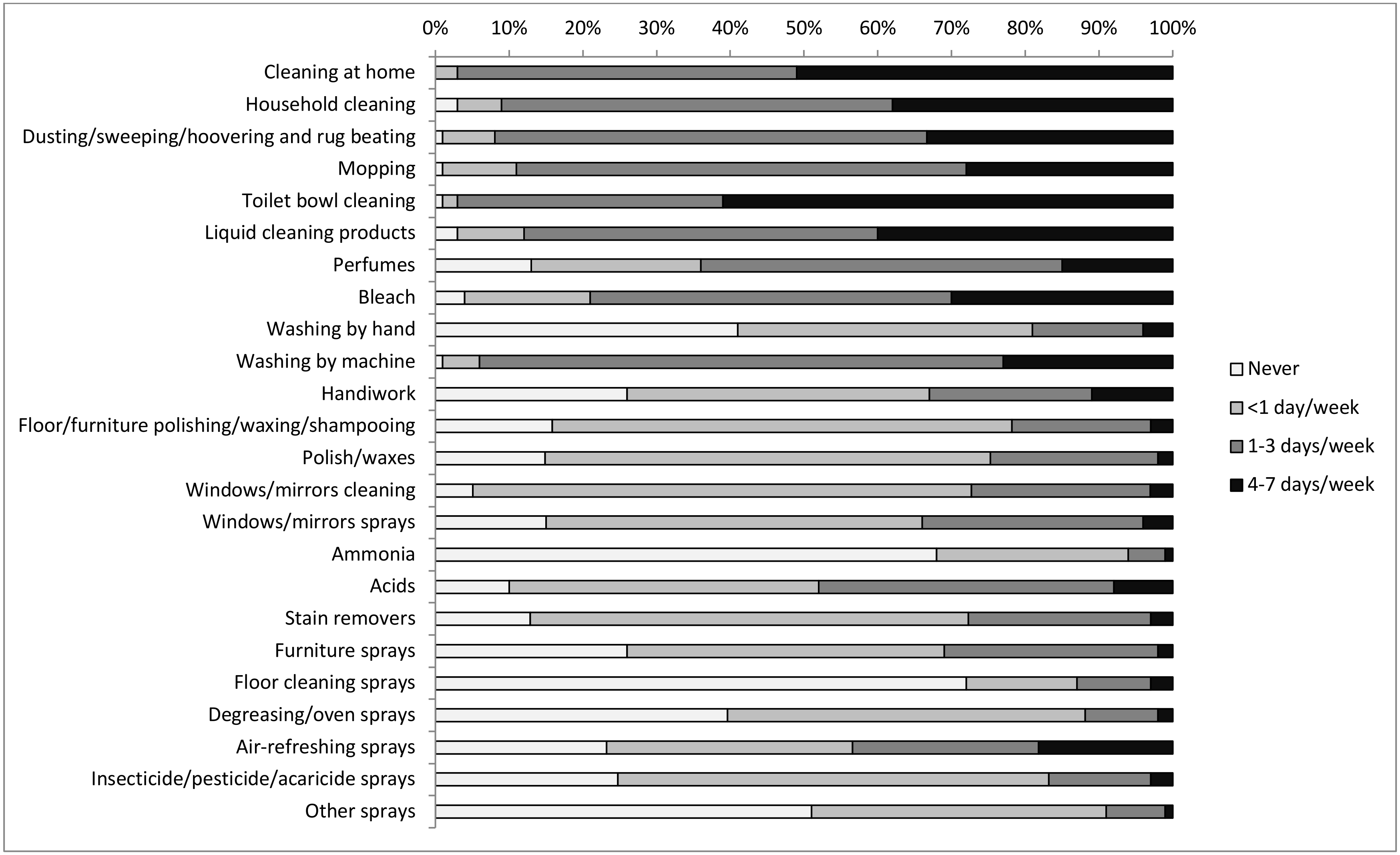

Figure E7. Description of the class “Very frequent use of products”. Results presented as posterior probabilities of each variable.