doi.org/10.26398/IJAS.0029-010

STATISTICAL TREATMENT OF FREE SORTING DATA: A BRIEF REVIEW OF METHODS AND A NEW ASSOCIATION MODEL

El Mostafa Qannari[1], Evelyne Vigneau

StatSC, ONIRIS, INRA, 44322, Nantes, France

***Abstract**. Several statistical procedures, mainly pertaining to multidimensional scaling, multiple correspondence analysis and cluster analysis, have been proposed for the analysis of data from a free sorting procedure. A brief review of these methods is sketched and a statistical model to assess the association between products is discussed. Among other possibilities, these models make it possible to set up a hypothesis testing framework to assess the significance of the effect of external factors on free sorting data. It is also performed in conjunction with a latent class strategy to identify segments of consumers and better highlight the relationships among the products. An illustration on the basis of a case study is outlined.*

***Keywords**: Free sorting, Multiple correspondence analysis, Co-occurrence matrix, Association model, Logistic regression, Latent class model.*

1. INTRODUCTION

In the food industry as well as in other sectors of activity such as the cosmetics and car industry, sensory analysis is nowadays a vital step in quality control and product development. Several sensory evaluation procedures can be implemented according to the objectives and the nature of the products under study (Meilgaard et al., 2006). For three decades or so, the most important procedure of sensory evaluation has been the so-called quantitative descriptive analysis (Chapman et al., 2001) whereby a set of products is described by a panel of trained assessors using a list of sensory attributes pertaining to the appearance, the taste, the flavor etc. Nowadays, the trend in sensory analysis is towards less costly and less time consuming procedures which, moreover, directly involve the final consumers of the products instead of trained assessors. The cross-fertilization of ideas in different scientific fields such as experimental psychology and psycho-acoustic has promoted the use of free

[1] Corresponding author: El Mostafa Qannari, email: elmostafa.qannari@oniris-nantes.fr

sorting procedure as a new method of investigation in sensory analysis. Although introduced in the sensory field only recently (Lawless et al., 1995; Faye et al., 2004), the free sorting task has been since the late sixties a central topic in cognitive and social psychology (see for instance Rosenberg and Olshan, 1970). Based on a holistic approach of characterization, this procedure of evaluation is simple and very intuitive since it relies on the natural human tendency to grouping objects into categories as a cognition process for learning and apprehending the complexity of the world.

In a free sorting task, each subject (judge, consumer...) is instructed to partition the products in as many groups as they deem it necessary with the understanding that the products in the same group are perceived as being similar. Optionally, in addition to partitioning the products, each subject may be instructed to describe the various groups of the partition of the products that he or she has set up, using a predefined list of attributes or lists of attributes freely suggested by the subjects themselves. Most often, these attributes are used as supplementary elements in the statistical analyses to enhance the interpretation of the outcomes from the sorting task and investigate the perceptive differences among the products.

In Section 2, we sketch a brief review of methods for the analysis of the free sorting data. In Section 3, we advocate using (simple) correspondence analysis (CA) to analyze the free sorting data. In Section 4, we introduce a new association model and discuss its properties and possibilities of extension, using a latent class model. An illustration on the basis of a case study on luxury perfumes is outlined in Section 5 and we end the paper by some concluding remarks.

2. OVERVIEW OF METHODS FOR THE STATISTICAL TREATMENT OF THE FREE SORTING DATA

As mentioned above, a free sorting test yields a family of partitions of the same set of products; each partition being associated with a particular subject. For the statistical treatment of these data, several strategies of analysis have been proposed. We refer to the book by Coxon (1999) for a comprehensive review of these strategies. We also refer to a French paper by Faye, Courcoux and Qannari (2011) for a comparison of several methods of analysis on the basis of a case study.

Broadly speaking, the statistical strategies of analysis can be classified in two families of methods, namely (i) factor analytical methods including, in particular, multidimensional scaling (MDS) and multiple correspondence analysis (Takane, 1981, 1982, Qannari et al., 2009; Cadoret et al., 2009) (ii) methods pertaining to cluster analysis (Faye et al., 2011; Courcoux et al., 2014)) and additive trees (Dubois, 1991).

Let us suppose that K subjects take part in a free sorting task in the course of which they were instructed to sort n products (or more generally, stimuli) into groups. Let us assume that subject k (k=1,…, K) has sorted the stimuli into Q_k clusters. As mentioned above, the data can be seen as a set of K partitions of the products, each partition being associated with a particular subject. Alternatively, the data for each subject can be expressed in a (stimuli by stimuli) symmetric matrix containing ones and zeroes. More precisely, an entry of this matrix contains one if the stimuli associated with the row and the column of this entry are sorted into the same group, otherwise the entry contains a zero. This matrix is usually interpreted as a similarity matrix between stimuli and can be easily transformed into a dissimilarity matrix by changing the zeroes into ones and vice versa. At the panel level, the individual similarity or dissimilarity matrices can be added up across the subjects yielding a global similarity or a global dissimilarity matrix between the products. As a matter of fact, the global similarity matrix thus obtained is also known as the co-occurrence matrix since it gives the number of times each pair of products have been set in the same group by the subjects. This matrix that we shall denote by $\mathbf{N}=(n_{ij})$ will be the focus of the present paper.

Several authors have proposed performing MDS on the global dissimilarity matrix to investigate the relationships among the stimuli (MacAdams et al., 1995; Lawless et al., 1995; King et al., 1998; Faye et al., 2004). Besides applying MDS to sorting data, Van der Kloot and Van Herk (1991) also suggested applying multiple correspondence analysis (MCA; Greenacre, 2007). In this procedure, the authors consider that there are as many categorical variables as subjects and the categories for each variable (i.e. subject) are the groups of stimuli formed by the subject under consideration. On the basis of a comparison study of several procedures including MCA, these authors draw the conclusion that the differences among the outcomes of the various methods are on average very small. Takane (1981) proposed a procedure of analysis called MDSORT for analyzing sorting data. From the derivation of the solution to this problem, it can be seen that, as a matter of fact, we are led to the same solution as MCA. This remark was also stressed by Van der Kloot and Van Herk (1991) who stated that their program for running MCA gave outcomes which were identical to those of MDSORT. Takane (1982) proposed yet another procedure called IDSORT for analyzing sorting data that takes account of individual differences among the subjects. This is done by combining ideas from MDSORT (i.e. MCA adapted to the free sorting data) and INDSCAL algorithm (Carroll and Chang, 1970). This yields a space of representation for the stimuli (stimulus space) and a set of weights associated with the subjects which reflect the importance they attach to the various dimensions of the stimuli space. Although it follows a different pattern of thinking pertaining to multi-block data analysis, another strategy of analysis was developed by Qannari et al. (2009) under the acronym Sort-CC. Sort-CC aims at exactly the same purpose as IDSORT and yields similar kinds of outcomes, namely a stimulus space and saliences (or weights) that the subjects attach to the underlying dimensions.

Another strategy of analysis of the free sorting data which stands at the crossroads of MDS and multiblock data analysis was proposed by Abdi et al. (2007) under the acronym DISTATIS. It is designed to be "a generalization of classical multidimensional scaling which allows one to analyze three-ways distance tables". Basically, DISTATIS consists in computing the so-called Torgerson's forms (i.e. matrices of cross-products between stimuli) associated with the individual (i.e. subjects') dissimilarity tables and running the STATIS method (Lavit et al., 1994) on these matrices. This makes it possible to obtain a map for the subjects and a map for the stimuli which depict the relationships between subjects on the one hand and stimuli on the other hand. However, unlike IDSORT and CC-Sort, the representation space of the subjects is fundamentally one-dimensional, reflecting the overall agreement of the subjects on how the stimuli relate to each others.

3. CORRESPONDENCE ANALYSIS

Up to our knowledge, nobody has proposed so far performing CA on the co-occurrence matrix, **N**, as a strategy of analysis for the statistical treatment of the free sorting data. Clearly, this is a very simple and straightforward statistical method which, among others, leads to graphical displays that reflect the relationships among the stimuli. As a matter of fact, both the rows and the columns of matrix **N** refer to the stimuli but since **N** is symmetric, the graphical representations of the rows and the columns are identical. It is worth noting that we run this strategy of analysis together with MCA on several free sorting data and we systematically obtained the same configurations of the stimuli except, for the percentages of variation recovered by the factorial axes from both the analyses. An illustration of this finding is discussed below. Although a thorough investigation of the formal connection between both analyses (i.e. CA on the co-occurrence matrix and MCA on the categorical variables) is still missing, we believe that this connection rests on the fact that, if we denote by **X** the matrix formed of the dummy variables (or indicators) associated to the various categories (i.e. groups formed by the various subjects) then the Burt table associated with **X**, which forms the cornerstone of

MCA applied to $\mathbf{X}$ is given by $\mathbf{X}^T\mathbf{X}$. By comparison, the co-occurrence matrix can be computed as $\mathbf{X}\mathbf{X}^T$ and can be seen as the Burt table associated to matrix $\mathbf{X}^T$.

4. ASSOCIATION MODEL

4.1 PARAMETER ESTIMATION

We propose hereinafter an association model which is, as it is the case for any statistical model, a way of simplifying the complexity of the data. Moreover, it offers a statistical hypothesis testing framework which is so far lacking in the context of the free sorting data. For instance it could be used to assess the effect of experimental design factors (subjects, session…). The association model enjoys several features, among which we single out: (i) it is intuitively appealing, (ii) it bears some similarities to Bradley-Terry model for paired comparison data (Bradley and Terry, 1952) and (iii) the estimation of the parameters is straightforward.

For each subject *k*, we consider the Bernoulli variable $y_{ij,k}$ which is equal to 1 if the products *i* and *j* are set in the same group by subject *k* and, 0 otherwise. The association model assumes that the expectation of $y_{ij,k}$, which is the probability that products *i* and *j* are set in the same group by subject *k*, is equal to:

$$p_{ij} = \frac{\pi_i \pi_j}{1+\pi_i \pi_j} \qquad \text{for } i \neq j \qquad (1)$$

The π_i ($i=1,\ldots,n$), assumed to be equal or larger than 0, are scores associated with the products. It is clear that if π_i (or π_j) increases then p_{ij} increases. Thus, π_i reflects the overall propensity of stimulus *i* to gather with the other stimuli (i.e. closeness). The model assumes that p_{ij} which, in some way, reflects the similarity between products *i* and *j* can be recovered by the parameters π_i ($i=1,\ldots,n$).

Let us consider $p^*_{ij}=1-p_{ij}$ (the probability that stimuli *i* and *j* are not in the same group), we can easily show that:

$$p^*_{ij} = 1 - p_{ij} = \frac{\pi^*_i \pi^*_j}{1+\pi^*_i \pi^*_j} \qquad (2)$$

where $\pi^*_i = \frac{1}{\pi_i}$. This parameter reflects the propensity of product *i* to stay apart from the other products (i.e. aloofness).

Finally, by dividing in Equation (1) the numerator and the denominator by π_i

or π_j we are led to the following expression of the association model:

$$p_{ij} = \frac{\pi_i}{\pi_i + \pi_j^*} = \frac{\pi_j}{\pi_j + \pi_j^*} \tag{3}$$

The interpretation of this model is straightforward: for the comparison of stimuli *i* and *j*, π_i is weighed against π_j^*. If π_i is larger than π_j^* then p_{ij} will be larger than 0.5, indicating an association between stimuli *i* and *j* above the average; otherwise p_{ij} will be smaller than 0.5 indicating a poor association. The same conclusions stand if π_j is weighed against π_i^*.

It is worth noting that there is a clear resemblance between the association model as stated in Equation (2) and the well known Bradley-Terry model for paired comparison data (Bradley and Terry, 1952; Courcoux and Semenou, 1997) which states that the probability (q_{ij}) that a stimulus *i* is preferred to a stimulus *j* is given by:

$$q_{ij} = \frac{\theta_i}{\theta_i + \theta_j}$$

where the parameters θ_i reflect preference scores associated to the products.

For the estimation of the parameters of the association model, it is possible to set up a statistical framework which would lead us to an estimation of the parameters by means of maximum likelihood estimation. For this purpose, we consider the variable y_{ij} which is equal to the number of times that products *i* and *j* are set in the same group. If we assume that the Bernoulli variables $y_{ij,k}$ defined above are independent, then y_{ij} follows a binomial distribution and the likelihood of the total experiment is the product of binomial density functions.

Alternatively, the model can be viewed as a special case of generalized linear models. The model stated in Equation (3) can be written as:

$$\text{logit}\,(p_{ij}) = \alpha_i + \alpha_j \tag{4}$$

where $\text{logit}(x) = \ln\left(\frac{x}{1-x}\right)$ and $\alpha_i = \ln(\pi_i)$.

It follows that the parameters α_i could be estimated using standard software of generalized linear models. More precisely, the data can be presented in two forms as illustrated in Figure 1 for the case of three products and four subjects. In Figure 1A (left), the raw data are presented for each subject indicating whether this subject has set the two stimuli in the same group, in which case, variable *y* takes the value

1; otherwise *y* takes the value 0. In Figure 1B (right), the data are summed up across the subjects and give the number of times that each pair of stimuli have been set in the same group. Note that these frequencies are precisely those found in the co-occurrence matrix, **N**, with the diagonal elements being excluded since they are deemed uninformative. From one or the other of these tables, the parameters α_i can be estimated using a logistic regression (assuming a binomial distribution) to explain the variable *y* from the predictors labeled in Figure 1 as "ProductA" and "ProductB". The logistic regression should be constrained to have no intercept, which is a common option in most statistical software packages.

ProductA	ProductB	Subject	y
Product1	Product2	S1	1 or 0
Product1	Product3	S1	1 or 0
Product2	Product3	S1	1 or 0
Product1	Product2	S2	1 or 0
Product1	Product3	S2	1 or 0
Product2	Product3	S2	1 or 0
Product1	Product2	S3	**1 or 0**
Product1	Product3	S3	**1 or 0**
Product2	Product3	S3	**1 or 0**
Product1	Product2	S4	**1 or 0**
Product1	Product3	S4	**1 or 0**
Product2	Product3	S4	**1 or 0**

A

ProductA	ProductB	***(frequency)***
Product1	Product2	n_{11}
Product1	Product3	n_{12}
Product2	Product3	n_{13}

B

Figure 1: Two ways of presenting the free sorting data to be subjected to a logistic regression

Another way to see the same problem is the following. If we denote by **z** the vector whose entries are logit(p_{ij}), then the model in Equation 4 could be written as follows: $\mathbf{z}=\mathbf{X}\alpha$, where $\alpha=(\alpha_1,\ldots\alpha_N)^T$ and **X** is a matrix whose row corresponding to the entry logit(p_{ij}) contains 1 in the i^{th} and j^{th} columns and 0 otherwise. This configuration is illustrated in Figure 2 with the case of three products. From this new setting, it follows that the parameters in vector α could be estimated by performing a linear regression of **z** on **X**. Again, in this model, one should constrain the model to have no intercept.

Z	Product1	Product2	Product3
logit(p_{12})	1	1	0
logit(p_{13})	1	0	1
logit(p_{23})	0	1	1

Figure 2: Representation of the free sorting data to be subjected to a multiple linear regression

4.2 SUBJECTS' EFFECT

It is worth noting that the presentation of the data in Figure 1A where the free sorting data are detailed for each subject entails that the subjects could be introduced as a factor. This may be of interest since it is likely to highlight some specific characteristics of the subjects. In this case, the logistic model can be written as:

$$\text{logit}(p_{ij,k}) = \alpha_i + \alpha_j + \beta_k \tag{5}$$

where β_k is the effect associated with subject k. A positive effect, β_k, will result in an increase of the propensity of the products to be associated. In particular, this is the case of those subjects who form small numbers of groups of products. Contrariwise, a negative effect, β_k, will result in a propensity of the products to be dispersed. In particular, this is the case of those subjects who form large numbers of groups of products. This aspect will be illustrated below through the case study.

Similarly, other factors could also be included. For instance, these factors could relate to the subjects (gender, region of residence…), the products (manufacturing process, constituents…) or the experimental design (sessions, periods of time…).

4.3 LATENT CLASS MODELS

A single parameters vector $\pi = (\pi_1, \ldots, \pi_n)^T$ may not adequately fit the data if the panel of subjects is not homogeneous. By introducing the subjects' effect as in Equation (5), we assumed that the same products scores hold for the whole panel of subjects but the probability that the subject k sets the products i and j in the same group is corrected by a scalar (positive or negative) that reflects the effect of this subject. The assumption of presence of several segments of subjects entails that from one segment to another, the subjects associate the products differently. In other words, the scores associated with the products vary from one segment to another. Obviously, it is of paramount interest to identify these segments. To do this, we suggest setting up a finite mixture model, also known as latent class (LC) analysis

(McLachlan and Peel, 2000).

In the present situation, the LC model assumes that there are unobserved classes (also called components) of subjects and the association between any two products depends on which class (or component) we consider. More precisely, by considering a mixture of models with *A* classes, we assume that the likelihood function associated with the vector of observations $\mathbf{y} = (y_{ij,k})^T$ is of the form:

$$L\left(\mathbf{y}\middle|\omega_1,\ldots,\omega_A,\boldsymbol{\pi}_1,\ldots,\boldsymbol{\pi}_A,\right) = \sum_{a=1}^{A}\omega_a f\left(\mathbf{y}\middle|\mathbf{y},\boldsymbol{\pi}_a\right) \quad (6)$$

where ω_a and $(\pi_a=\pi_1^{(a)},\ldots,\pi_n^{(a)})^T$ are the model parameters defined, respectively, as the a priori probability and the vector of scores for the component *a* (*a* = 1, …, *A*) and , *f* is the Bernoulli distribution function.

With a fixed number of classes, these parameters are usually estimated by means of the so-called expectation-maximization (EM) algorithm (Dempster et al., 1977). The general strategy is to fit a sequence of models with one class, two classes, and so on. Several fit statistics are provided to help with model selection and comparison. These fit statistics include, in particular, the likelihood-ratio statistic, Akaike's Information Criterion (AIC; Akaike, 1974) and Bayesian Information Criterion (BIC; Schwarz, 1978). In this paper, we shall use BIC to compare competing models (e.g. models with different numbers of latent classes). A smaller BIC for a particular model in comparison to another suggests that the trade-off between fit and parsimony is better achieved.

The R package FlexMix was used to fit the LC model (Gruen and Leisch, 2007).

5 ILLUSTRATION

5.1 CA PERFORMED ON THE CO-OCCURRENCE MATRIX

The data are presented in (Cadoret et al., 2009) to illustrate a strategy of analysis of the free sorting data called FAST (Factorial Approach for Sorting Task). FAST is articulated around MCA and provides, in addition to the usual outputs of this method of analysis, graphical displays and statistics which are of particular interest in the free sorting data. The data and the program FAST are implemented in the *R* package SensoMineR (Lé and Husson, 2008).

A panel of 98 consumers were instructed to carry out a free sorting task of twelve luxury perfumes whose names can be found in Table 1 which gives the co-occurrence matrix issued from the free sorting task.

Table1: Co-occurrence matrix from the sorting task of 12 luxury perfumes

	Jadore(ET)	Jadore(EP)	Pleasures	Coco Mademoiselle	Cinema	Pure Poison	Linstant	Lolita Lempicka	Angel	Chanel5	Aromatics Elixir	Shalimar
Jadore(ET)	98	56	48	38	24	26	22	18	12	14	7	7
Jadore(EP)	56	98	38	28	23	28	28	18	12	12	12	6
Pleasures	48	38	98	28	22	29	23	18	11	14	11	6
Coco Mademoiselle	38	28	28	98	30	32	20	21	11	11	12	9
Cinema	24	23	22	30	98	27	26	42	19	9	8	10
Pure Poison	26	28	29	32	27	98	24	17	10	21	12	11
Linstant	22	28	23	20	26	24	98	22	14	10	13	13
Lolita Lempicka	18	18	18	21	42	17	22	98	35	8	6	9
Angel	12	12	11	11	19	10	14	35	98	16	27	21
Chanel5	14	12	14	11	9	21	10	8	16	98	51	30
Aromatics Elixir	7	12	11	12	8	12	13	6	27	51	98	42
Shalimar	7	6	6	9	10	11	13	9	21	30	42	98
Total	370	359	346	338	338	335	313	312	286	294	299	262

We run CA on the co-occurrence matrix. Figure 3 shows the configuration of the 12 products on the basis of the first two factorial axes. These axes explain around 56% of the inertia.

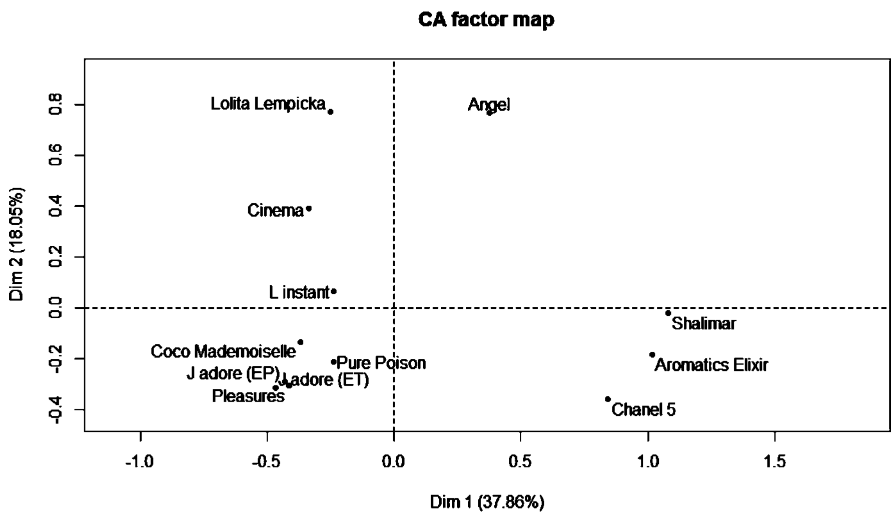

Figure 3: Representation of the 12 luxury perfumes on the basis of the first two CA axes

By way of comparing methods, we also run MCA on the sorting data. The representation of the products on the basis of the first two axes is not shown herein because it is identical to that given in Figure 3. The only difference is the percentage of inertia recovered by the first two factorial axes of MCA which are 17.7% and 13.5%, respectively.

Along the first factorial axis, "Shalimar", "Aromatic Elixir" and "Chanel 5" are singled out and seem to form a group by themselves. The other products are situated on the opposite side of this axis except "Angel" which has a central position. In particular, the products "J'adore ET", "J'adore EP", "Pleasures", "Coco Mademoiselle" and "Pure Poison" seem to be very close to each others. The second axis (18% of the inertia) highlights the products "Angel" and "Lolita Lempicka" which are both on the positive side of this axis.

5.2 ASSOCIATION MODEL

The scores associated with the various products obtained by means of the association model are depicted in Figure 4 together with bars indicating the 95% confidence intervals for these estimates.

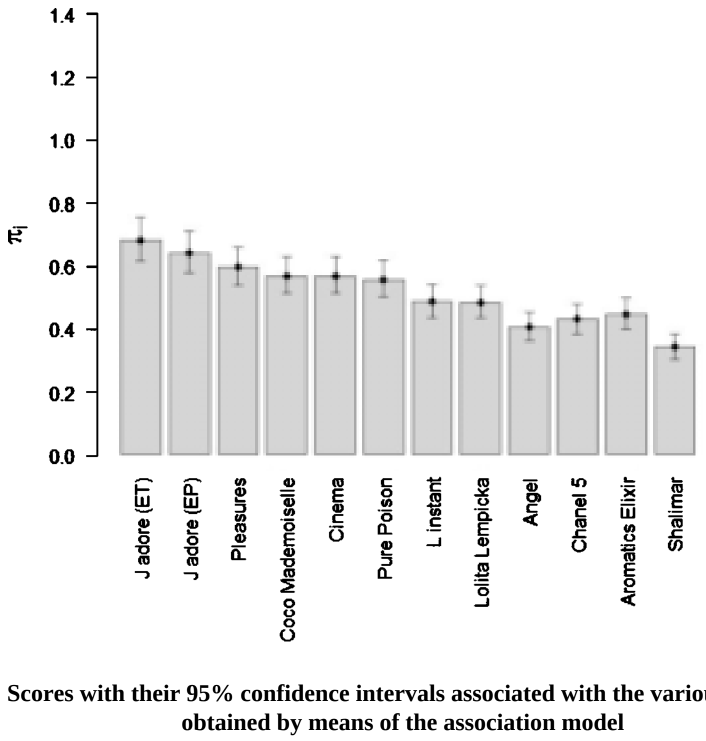

Figure 4: Scores with their 95% confidence intervals associated with the various products obtained by means of the association model

From Figure 4, we can see that the highest scores are those corresponding to "J'adore ET", "J'adore EP", "Pleasures", "Coco Mademoiselle", "Cinema" and "Pure Poison". These products are associated to each others as it can be seen on the graphical display from CA (Figure 3). The scores associated with products "Chanel 5" and "Aromatic Elixir" are also relatively large although smaller than the scores of the products in the previous group of perfumes. These two products are associated to each other. The scores associated with products "Angel" and "Shalimar" are the smallest indicating a less important association to the other products.

5.3 SUBJECTS' EFFECT

In order to assess the differences among the subjects, we introduced this factor in the logistic regression. When model (5), which includes the subjects' effects, is compared with the simplified model without the subjects' effect terms, the log-likelihood test is highly significant (p-value < 10^{-5}, for a deviance of 494 and 97 degrees of freedom). Therefore, we conclude that there are significant differences among the subjects; some showing a positive effect and others showing a negative effect (results not shown herein for the sake of saving space).

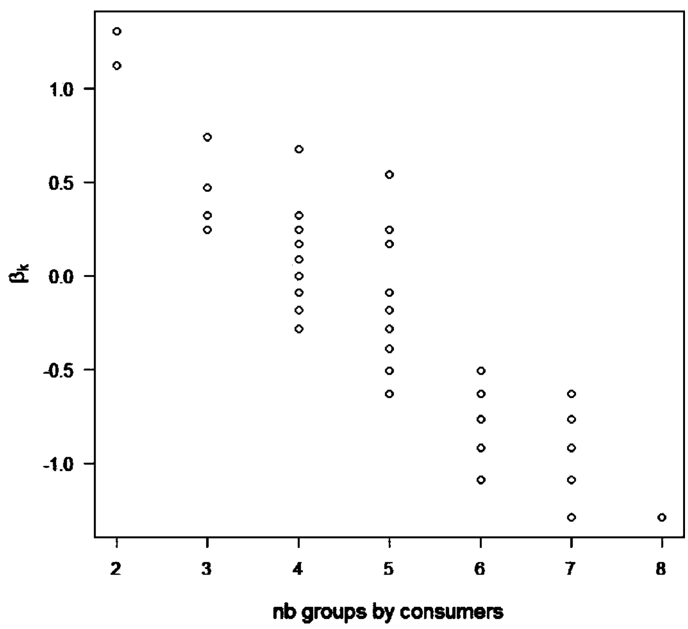

Figure 5: Relationship between the number of groups formed by the subjects and the subjects' effects, β_k

Figure 5 highlights the relationships between the number of groups formed by the subjects in the course of the free sorting procedure and the subjects' effects, β_k, associated with these subjects. It is clear that there is a very tight connection between these quantities indicating that as the number of groups increases the subject's effect becomes negative. This is only normal, since a relatively large number of groups does not favor the association of the products. The implication of this finding is that the subjects should be instructed beforehand not to use too many or too few groups when they perform the free sorting task. More precise instructions depend on the context of the study (number of products, assumed differences among the products…).

5.4 LATENT CLASS MODEL

Figure 6 shows the evolution of BIC as a function of the number of components (i.e. classes) from the latent class model (6). It can be seen that when we move from one component to three components, the BIC decreases then, it starts to increase. This indicates that the choice of three classes is appropriate.

The prior class probabilities of the three latent classes are 61%, 6% and 33%, respectively. The posteriori class probabilities (not shown) indicate that a large proportion of the 98 consumers are clearly assigned to a single class (i.e. only very few consumers have more or less equal posteriori probabilities for two clusters).

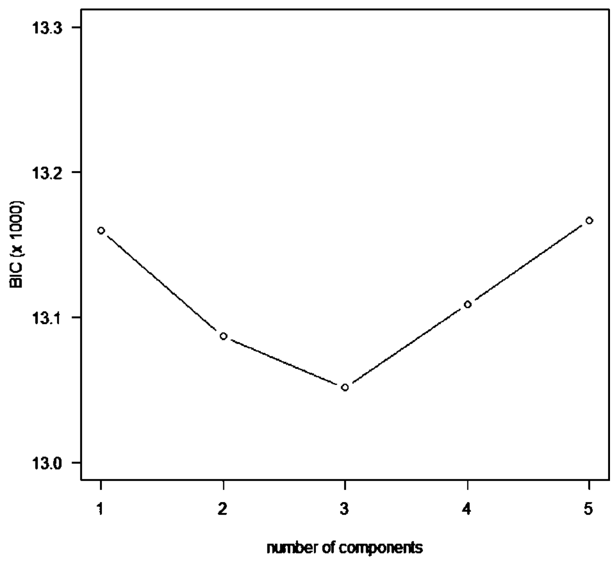

Figure 6: Evolution of BIC as a function of the number of latent class components

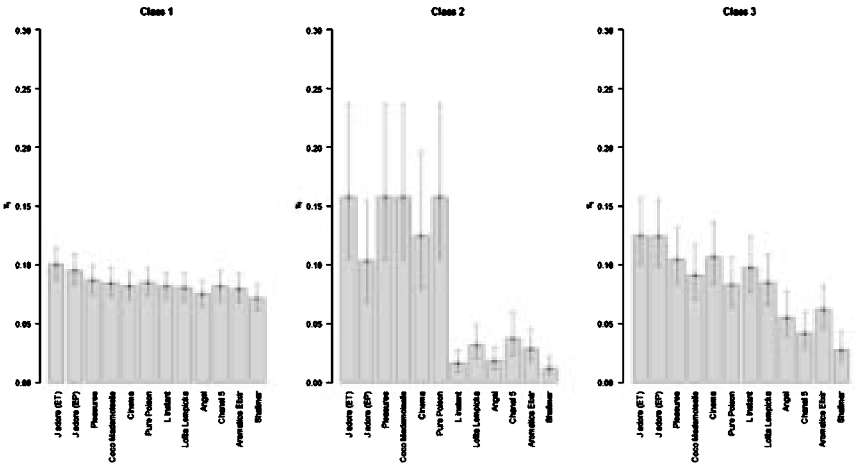

Figure 7: Association parameters π_i in the three latent lasses, together with their 95% confidence intervals

The Figure 7 shows the estimated parameters π_i in the three classes (after transformation so that, in each class, their sum is equal to 1, for a better readability of the graph). It is clear that, in the first class (prior probability: 61%), the scores are more or less similar to those estimated when only one class was considered (Figure 4). The third class (prior probability: 33%) highlights those consumers who have a tendency to set apart “Chanel 5” and “Shalimar” and to associate “J’adore ET” and “J’adore EP” in a more marked manner than in class1. The second class (prior probability: 6%) corresponds to a marginal class of consumers who have a tendency to lump together the products “J’adore”, “Pleasure”, “Coco Mademoiselle” and “Pure Poison”, on the one hand and to set apart the products “L’instant”, “Angel” and “Shalimar” on the other hand. However, it also appears that the estimated scores in this latent class have much larger uncertainties than in the other classes as shown by the width of the 95% confidence intervals in Figure 7.

6. CONCLUDING REMARKS

Notwithstanding its simplicity, the free sorting task is an effective way to investigate the relationships among a set of products. Judging from the wealth of methods that have been proposed to analyze the data issued from this task, it seems that it has very much inspired generations of practitioners in psychology, food science, data analysis…

Our contribution to analyzing such data revolves around the co-occurrence matrix. On the one hand, we suggest performing correspondence analysis which seems to lead to outcomes very similar to those of MCA. On the other hand, we propose a new model to investigate the association among the products. Among other features, this model presents the advantage of being intuitively appealing and the estimation of the parameters is straightforward. Moreover, when this model is performed in conjunction with a latent class strategy, it offers an efficient approach to analyze the free sorting data.

More research work is needed to strengthen the efficiency of this new model which we hope will stand for the free sorting data in the same position as Bradley-Terry model stands for paired comparison data.

More investigations are also needed to extend the scope of application of the association model beyond the free sorting data. More precisely, it is clear that the model could be applied to any symmetric contingency table. This is, in particular, the case of Burt's matrix derived from MCA. This means that it is likely to enhance the interpretation of the outcomes of this latter method of analysis by using it in conjunction with the association model discussed herein.

REFERENCES

Abdi, H., Valentin, D., Chollet, S. and Chrea, C. (2007). Analyzing assessors and products in sorting tasks: *DISTATIS*, theory and applications. In *Food Quality and Preference*. 18: 627-640.

Bradley, R.A. and Terry, M. (1952). The rank analysis of incomplete block designs: I. the method of paired comparisons. In *Biometrika*. 39: 324–345.

Cadoret, M., Lê, S. and Pagès, J. (2009). A factorial approach for sorting task data (FAST). In *Food Quality and Preference*. 20: 410-417.

Carroll, J.D. and Chang, J.J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalisation of 'Eckart-Young' decomposition. In *Psychometrika*. 35: 283-319.

Chapman, K.W., Lawless, H.T. and Boor, K.J. (2001). Quantitative descriptive analysis and principal component analysis for sensory characterization of ultrapasteurized milk. In *Journal of Dairy Science*. 84: 12–20.

Coxon, A.C. (1999). *Sorting Data: Collection and Analysis*. Thousand Oaks, CA: Sage Publications (Quantitative Applications in the Social Sciences).

Courcoux, Ph. and Semenou, M. (1997). Preference data analysis using a paired comparison model. In *Food Quality and Preference*. 8(5/6): 353-358.

Courcoux, P., Faye, P. and Qannari, E.M. (2014). Determination of the consensus partition and cluster analysis of subjects in a free sorting task experiment. In *Food Quality and Preference*. 32: 107-112.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. In *Journal of the Royal Statistical Society, Series B (Methodological).* 39(1): 1-38.

Dubois, D. (1991). Sémantique et cognition. Catégories, prototypes, typicalité. Editions du CNRS, Collection Sciences du langage, Paris.

Faye, P., Brémaud, D., Durand-Daubin, M., Courcoux, Ph., Giboreau, A. and Nicod, H. (2004). Perceptive free sorting and verbalization tasks with naive subjects: an alternative to descriptive mappings. In *Food Quality and Preference*. 15(7-8): 781-791.

Faye, P., Courcoux, P. and Qannari, E.M. (2011). Méthodes de traitement statistique des données issues d'une épreuve de tri libre. In *Revue Modulad*. 43 : 1-24.

Greenacre, M.J. (2007). *Correspondence Analysis in Practice*. Second Edition. Chapman & Hall / CRC, Boca Raton FL.

Gruen, B. and Leisch, F. (2007). Fitting finite mixtures of generalized linear regressions in R. In Computational Statistics & Data Analysis. 51(11): 5247-5252.

King, M.C., Cliff, M.A. and Wall, J.W. (1998). Comparison of projective mapping and sorting data collection and multivariate methodologies for identification of similarity-of-use of snack bars. In *Journal of Sensory Studies*. 13: 347-358

Lavit, Ch., Escoufier, Y., Sabatier, R. and Traissac, P. (1994). The ACT (Statis method). In *Computational Statistics and Data Analysis.* 18: 97-119.

Lawless, H.T., Sheng, T. and Knoops, S. (1995). Multidimensional scaling of sorting data applied to cheese perception. In *Food Quality and Preference*. 6: 91-98.

Lé, S. and Husson, F. (2008). Sensominer: a package for sensory data analysis. In *Journal of Sensory Studies*. 23: 14–25.

MacAdams, S., Winsberg, S., Donnadieu, S., De Soete, G. and Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres : Common dimensions, specificities and latent subject classes. In *Psychological Research*. 58: 177-192.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York.

Meilgaard, M.C., Carr, B.T. and Civille, G.V. (2006). Sensory Evaluation Techniques, Fourth Edition. CRC Press, Boca Raton.

Qannari, E.M., Cariou, V., Teillet, E. and Schlich, P. (2009). SORT-CC A procedure for the statistical treatment of free sorting. In *Food Quality and Preference*. 21: 302-308.

Rosenberg, S. and Olshan, K. (1970). Evaluative and descriptive aspects in personality perception. In *Journal of Personality and Social Psychology*. 16: 619-626.

Schwarz, G. (1978). Estimating the dimension of a model. In *Annals of Statistics*. 6: 461–464.

Takane, Y. (1981). MDSORT: A special purpose multidimensional scaling program for sorting data. In *Journal of Marketing Research*. 18: 480-481.

Takane, Y. (1982). IDSORT: An individual differences multidimensional scaling for sorting data. In *Behavior Research Methods and Instrumentation.* 14: 546.

Van der Kloot, W. A. and van Herk, H. (1991). Multidimensional scaling of sorting data: a comparison of three procedures. In *Multivariate Behavioral Research*. 26(4): 563-581.