Analysis and clustering of multiblock datasets by means of the STATIS and CLUSTATIS methods. Application to sensometrics

Fabien Llobell, Véronique Cariou, Amaury Labenne, El Mostafa Qannari

ANALYSIS AND CLUSTERING OF MULTIBLOCK DATASETS BY MEANS OF THE STATIS AND CLUSTATIS METHODS. APPLICATION TO SENSOMETRICS

Fabien Llobell[1,2], Véronique Cariou[1], Evelyne Vigneau[1], Amaury Labenne[2], El Mostafa Qannari[1]
[1]StatSC, ONIRIS, INRA, Nantes, France.
[2]Addinsoft, XLSTAT, Paris, France.
fllobell@xlstat.com

Abstract

The STATIS method has been successfully applied to the analysis of sensory profiling data and other kinds data in sensometrics. We discuss its use and benefits and compare its outcomes to alternative methods for the analysis of multiblock data arising in situations such as projective mapping and free sorting experiments. More importantly, a method of clustering a collection of datasets measured on the same individuals, called CLUSTATIS, is introduced. It is based on the optimization of a criterion and consists in a hierarchical cluster analysis and a partitioning algorithm akin to the K-means algorithm. The procedure of analysis can be seen as an extension of the cluster analysis of variables around latent components (CLV, Vigneau and Qannari, 2003) to the case of blocks of variables. Alongside the determination of the clusters, a latent configuration is determined by the STATIS method. The interest of CLUSTATIS in sensometrics is discussed and illustrated on the basis of two case studies pertaining to the projective mapping also called Napping and the free sorting tasks, respectively.

Keywords - Cluster analysis, Multiblock datasets, STATIS, Projective mapping, Napping, Free sorting.

I. INTRODUCTION

Several sensory procedures directly lead to multiblock datasets. For instance, in sensory profiling evaluation, we are presented with a collection of data organized in blocks of variables. Each data block is associated with an assessor and gives the intensity scores of the products (rows) for several sensory attributes, which may be the same from one assessor to another (fixed vocabulary profiling) or different from one assessor to another (free choice profiling).

The STATIS method is a method of analysis of multiblock datasets (Lavit, Escoufier, Sabatier and Traissac, 1994). It was introduced to the sensometrics domain by Schlich (1996). It is nowadays popular among the practitioners of sensory analysis *(*Pizarro, Esteban-Díez, Rodríguez-Tecedor, González-Sáiz, 2013*).* We show how this method of analysis can be applied to projective mapping and free sorting data. We also compare the outcomes from this strategy of analysis to those obtained by means of standard procedures of analysis, namely Generalized Procrustes Analysis (GPA) and Multiple Factor Analysis (MFA) for projective mapping/Napping data, and DISTATIS for free sorting data. Another and more important aim of the paper is to introduce a cluster analysis approach of multiblock datasets, called CLUSTATIS. In the applications discussed hereinafter, CLUSTATIS will be used to segment the subjects involved in a free sorting task or a projective mapping/napping experiments. This clustering method is tightly linked to the STATIS method and consists in a hierarchical cluster analysis and a partitioning algorithm. Both these two strategies aim at optimizing the same criterion; either locally for the hierarchical clustering or globally for the partitioning algorithm They can be run independently or in combination in an attempt to achieve an even better solution than that obtained by one of the two strategies alone. Again, the efficiency of this clustering approach is demonstrated on the basis of data pertaining to projective mapping/Napping and free sorting tasks.

CLUSTATIS can be seen as an extension of the cluster analysis of variables around latent components (CLV, Vigneau and Qannari, 2003) to the case of blocks of variables. It follows the same pattern of analysis and enjoys the same properties.

Cluster analysis of datasets is not a new topic in sensometrics since this approach has been proposed with the aim of identifying sub-groups of assessors or outlying assessors (Dahl and Næs, 2004). These authors computed the Procrustes distances between pairs of datasets and subjected the distance matrix thus obtained to algorithms of hierarchical cluster analysis by considering various aggregation criteria. In particular, the clustering strategies are compared for the purpose of detecting outlying assessors. One of the findings was that single and centroid linkages are better suited for this aim than the other strategies of aggregation.

In the context of three-way data where the datasets refer to the same individuals (*e. g.,* products) and the same variables, Wilderjans and Cariou (2016) proposed a strategy of clustering the sensory descriptors in conventional profiling, called CLV3W. By adding a non-negativity constraint, these authors used this method to segment the consumers (Cariou, Wilderjans, 2018). Interestingly enough, this strategy of analysis is also an extension of the CLV method, which is a feature shared by CLUSTATIS. However, CLV3W is particularly designed for three-way data. Moreover, the configuration of the individuals (*e. g.,* products) within each cluster is restricted to be one dimensional. It is worth noting that CLV3W is a special case of the clusterwise Parafac model, which allows us to have higher-dimensional configurations within each cluster (Wilderjans, Ceulemans, 2013). By comparison, CLUSTATIS operates on multiblock data, which is a larger setting than three-way data since the variables may not be the same for all the datasets at hand.

In section 2 devoted to the material and methods, we start by recapitulating the STATIS method (subsection 2.1). Thereafter, we discuss the general strategy of cluster analysis (subsection 2.2) and show how it can be applied to projective mapping/Napping (subsection 2.3) and free sorting data (subsection 2.4). In section 3, we illustrate the approach on the basis of case studies pertaining to each of these tasks. Finally, we close the paper by some concluding remarks.

II. Material and methods

2.1. The STATIS method

We consider the setting where we dispose of m blocks of variables denoted by $\boldsymbol{X_1}$,… , $\boldsymbol{X_m}$, which are assumed to be column centred. These datasets are measured on the same n individuals (e. g., products) but the variables may differ in nature as well as in number from one dataset to another.

The STATIS method was introduced to the sensometrics domain by Schlich (1996) and is nowadays popular among the practitioners of sensory analysis *(*Pizarro, Esteban-Díez, Rodríguez-Tecedor, González-Sáiz, 2013*)*. The cornerstone of this strategy of analysis is the scalar product matrices associated with the datasets at hand. These matrices are computed as follows: $\boldsymbol{W_1} = \boldsymbol{X_1 X_1^T}$,…, $\boldsymbol{W_m} = \boldsymbol{X_m X_m^T}$. These are $n \times n$ symmetric matrices which reflect the spatial configuration of the individuals (*e.g.,* products) since the entry corresponding to the *l*th row and *j*th column gives the scalar product between these two individuals. It is recommended to scale the matrices $\boldsymbol{W_i}$ so as to have their norm equal to 1. This is achieved by dividing each $\boldsymbol{W_i}$ by its Frobenius norm. We recall that the Frobenius norm of a matrix $\boldsymbol{A}$ (say) is given by $\|\boldsymbol{A}\| = \sqrt{\sum_l \sum_j a_{lj}^2}$, where a_{lj} is the $(l, j)^{\text{th}}$ entry of matrix $\boldsymbol{A}$. Central to the STATIS method is the RV coefficient which is also very popular in sensometrics (El Ghaziri,

Qannari, 2015; Næs, Berget, Liland, Ares, Varela, 2017). This coefficient reflects the similarity between two configurations. Roughly speaking, the RV coefficient between $\boldsymbol{X_i}$ and $\boldsymbol{X_s}$ can be seen as the correlation coefficient between $\boldsymbol{W_i}$ and $\boldsymbol{W_s}$. It ranges between 0 and 1 and increases as the similarity, in terms of spatial configuration of the individuals, between the two datasets at hand increases. In the following we shall indifferently refer to this coefficient as $\boldsymbol{RV}(\boldsymbol{X_i}, \boldsymbol{X_s})$ or $\boldsymbol{RV}(\boldsymbol{W_i}, \boldsymbol{W_s})$.

The various steps to perform the STATIS method are depicted in Figure 1. It can be seen that the STATIS method seeks a weighted average configuration of the $\boldsymbol{W_i}$ matrices. As indicated in Figure 1, the weights, α_i, are obtained by computing the first eigenvector of the matrix of the RV coefficients between the various datasets. The weight α_i is relatively large if $\boldsymbol{X_i}$ tends to agree with the other datasets. Contrariwise, the weighting coefficients tend to be relatively small for differing $\boldsymbol{X_i}$. It is worth noting that the first eigenvalue of the matrix of RV coefficients, λ_1, stands for an overall agreement or homogeneity index between the various datasets. The standardized index $I = \frac{\lambda_1}{m}$ reflects the part of variation in the various matrices $\boldsymbol{W_i}$ explained by the group average matrix $\boldsymbol{W}$. It ranges between $\frac{1}{m}$ and 1. The larger this index, the higher is the agreement among the datasets $\boldsymbol{X_1}, \ldots, \boldsymbol{X_m}$.

Formally, we can show that the STATIS method seeks to find weighting (positive) scalars, α_i, and a compromise or group average scalar products matrix, $\boldsymbol{W}$ so as to minimize the following criterion:

$$\sum_{i=1}^{m} ||\boldsymbol{W_i} - \alpha_i \boldsymbol{W}||^2$$

As a determination constraint, we assume that $\sum_{i=1}^{m} \alpha_i^2 = 1$.

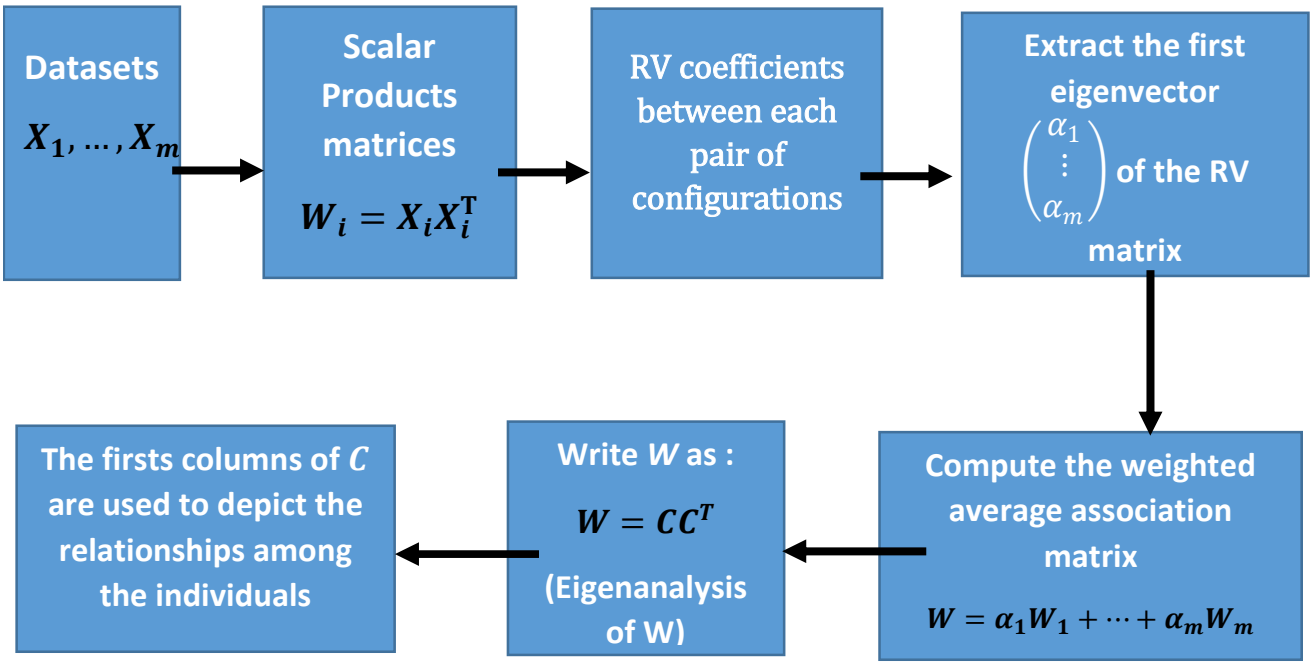

Figure 1. Finding a group average configuration by means of The STATIS method

2.2. The CLUSTATIS approach

We propose a cluster analysis approach for multiblock datasets. This approach consists in an extension of the CLV method (Vigneau and Qannari, 2003; Vigneau, Chen and Qannari, 2015). As CLV, CLUSTATIS is based on a hierarchical and a partitioning algorithms whose aim is to cluster the objects at hand (*i.e.,* variables in the case of CLV and datasets in the case of CLUSTATIS) in such a way that the objects in each cluster are as close as possible to a central (or latent) entity (*i.e.,* latent variable for CLV and a compromise cross product matrix in the case of CLUSTATIS). Moreover, the aggregation criteria used in the hierarchical algorithms for both methods are very similar.

CLUSTATIS aims at minimizing a criterion which reflects the fact that we are seeking homogeneous clusters of datasets. More precisely, the datasets in each cluster are assumed to be highly related to a latent configuration which is determined by means of the STATIS

method. Formally, let us denote by $\boldsymbol{X_1}, \ldots, \boldsymbol{X_m}$ the datasets at hand, which are assumed to be columnwise centered. We compute the scalar products matrices: $\boldsymbol{W_1} = \boldsymbol{X_1}\boldsymbol{X_1^T}$, …, $\boldsymbol{W_m} = \boldsymbol{X_m}\boldsymbol{X_m^T}$, and we seek to minimize the following criterion:

$$D = \sum_{k=1}^{K} \sum_{i \in G_k} ||\boldsymbol{W_i} - \alpha_i \boldsymbol{W^{(k)}}||^2$$

where α_i $(i \in G_k)$ are scalars to be determined and assumed to be such that $\sum_{i \in G_k} \alpha_i^2 = 1$, K is the number of groups of datasets, G_k is the k^{th} group and, for k $(k = 1, \ldots, K)$, $\boldsymbol{W^{(k)}}$ is the compromise of the group G_k. Obviously, when there is only one group of datasets (*i.e.*, $K = 1$), we retrieve the same criterion that underlies the STATIS method.

The procedure of cluster analysis to solve this problem is called CLUSTATIS and entails two complementary clustering strategies. The first strategy consists in a hierarchical cluster analysis. The second strategy consists in a partitioning algorithm akin to the K-means algorithm. Both strategies aim at optimizing the same criterion either locally or globally and, in practice, complement each other. More precisely, the hierarchical cluster analysis can help selecting the appropriate number, K, of clusters and provides a starting partition of the datasets that can be improved by means of the partitioning algorithm.

The hierarchical algorithm follows an ascending (or merging) strategy. We start with the situation where each dataset forms a group by itself. Obviously, in this case, $D = 0$. At each step, we merge two datasets or, more generally as the algorithm proceeds, two groups of datasets until all the datasets are merged in a single cluster. At each step, we can show that the criterion D increases. More precisely, we can show that when the two clusters A and B are merged, the criterion D increases by $delta = \lambda_1^{(A)} + \lambda_1^{(B)} - \lambda_1^{(A \cup B)}$, where $\lambda_1^{(A)}$, $\lambda_1^{(B)}$and

$\lambda_1^{(A\cup B)}$ are respectively the largest eigenvalue of the RV coefficients matrix between pairs of configurations in clusters A, B and $A \cup B$. The rationale of the aggregation strategy is to merge those two clusters A and B (say) which result in the smallest increase of criterion D. One should trace the increase of the criterion D as the hierarchical clustering proceeds because it reflects the loss of heterogeneity when we merge the clusters A and B. A jump of this quantity indicates that we are trying to merge two clusters which are heterogeneous. In practice, these quantities are reflected in the hierarchical tree (or dendrogram) as the height of the branches that connect two embedded nodes. Alternatively, these quantities could be represented as a bar plot showing their evolution as the number of clusters decreases.

The clustering problem based on the criterion D given above can also be solved by means of a partitioning algorithm akin to the K-means algorithm (Lloyd, 1982; Everitt, Landau, Leese, Stahl, 2011). In the course of this algorithm, the datasets are allowed to move in and out of the groups achieving at each step a decrease of the criterion D. This algorithm assumes that the number of clusters, K, is given beforehand and runs as follows:

- Step 1 (Initial partition of the datasets): K groups of datasets are given by the practitioner. These groups could be chosen by a random assignment of the datasets. A better initialization can be performed from the outcomes of the hierarchical clustering described above. This point will be further discussed below.
- Step 2 (Determination of the cluster compromise scalar product matrices) : In cluster G_k, the compromise scalar products matrix, $\boldsymbol{W}^{(k)}$, and the associated weights α_i are determined by means of the STATIS method as sketched in Figure 1.
- Step 3 (changing clusters): New clusters of datasets are formed by moving each dataset, $\boldsymbol{X_i}$, to the cluster G_k for which $RV(\boldsymbol{W_i}, \boldsymbol{W}^{(k)})$ is the largest.

The steps 2 and 3 are iterated until there is no change in cluster memberships. This means that the criterion D stops to decrease.

In practice, both the hierarchical and the partitioning algorithms should be performed to reach a better solution. Firstly, the hierarchical strategy can be used to hint at an appropriate number of clusters by examining the evolution of the aggregation criterion in the course of the aggregation process, as discussed above. Secondly, the datasets are submitted to the partitioning algorithm using as an initial solution, the partition obtained by cutting the hierarchical tree at the indicated level (*i.e.*, with the selected number of clusters). By allowing the switching of cluster memberships, the solution obtained by the hierarchical clustering may be improved since the criterion D is likely to be further minimized.

It is worth noting that since the hierarchical algorithm can be time consuming in situations where the number of datasets at hand is large, one can use only the partitioning algorithm. However, it is advised to perform a multi-start random partition by running this algorithm using several (random) partitions as starting points. Eventually, the final solution which corresponds to the smallest value of criterion D is retained.

Associated with each cluster of subjects G_k, CLUSTATIS yields a weighted average matrix $\boldsymbol{W}^{(k)}$, which is the compromise scalar products as obtained by the STATIS applied to the datasets in G_k. An eigenanalysis of the matrix $\boldsymbol{W}^{(k)}$ makes it possible to write $\boldsymbol{W}^{(k)} = \boldsymbol{C}^{(k)}\boldsymbol{C}^{(k)T}$. The matrix $\boldsymbol{C}^{(k)}$ stands for the group average configuration in cluster G_k.

Several indices associated with the final solution are of paramount interest. In the first place, we consider for each cluster G_k, $(k = 1, \dots, K)$, the index $I_k = \frac{\lambda_1^{(k)}}{m_k}$, where $\lambda_1^{(k)}$ is the largest eigenvalue of the matrix of the RV coefficients between the datasets in group G_k and m_k is

the number of datasets in this group. We know that this index ranges between $\frac{1}{m_k}$ and 1 and reflects the homogeneity in G_k. It can be interpreted as the percentage of variation in the datasets of group G_k which is explained by the group average configuration associated with $\boldsymbol{W}^{(k)}$. Within the group G_k, we can compute for each dataset $\boldsymbol{X_i}$ $(i \in G_k)$, the RV coefficient between $\boldsymbol{W_i}$ and $\boldsymbol{W}^{(k)}$. This index reflects how each dataset is close to its associated group configuration. Alternatively, we could consider the coefficient α_i $(i = 1, \dots, m)$ which reflects the same idea. An overall index to assess the quality of the partition of the datasets obtained by the clustering approach is given by the weighted average of the indices I_k : $I = \frac{\sum_{k=1}^{K} m_k I_k}{m} = \frac{\sum_{k=1}^{K} \lambda_1^{(k)}}{m}$. This index can be interpreted as the percentage of variation in the original datasets explained by the group average configurations in the various groups. Finally, in order to assess how the various groups of datasets are close to each other, we can compute the RV coefficients between their associated group average configurations. All these indices will be illustrated through the two case studies below.

2.3. The case of projective mapping or Napping

From a technical point of view, the procedures called Projective mapping (Risvik, McEwan, Colwill, Rogers and David, 1994) and Napping (Pages, 2005) are similar. They consist in instructing the subjects who participate in the experiment to position a set of products on a sheet of paper, considering that similar products should be located near one another and differing products should be placed far apart**.** Thus, the data can be presented as a collection of datasets; each dataset being associated with a subject and consists of the x and y coordinates of the products on the sheet of paper. Two statistical methods pertaining to the wide range of multi-block data analysis are concurrently proposed, namely Generalized Procrustes Analysis (GPA; Gower, 1975; Risvik *et al.*, 1994) and Multiple Factor

Analysis (MFA, Pagès, 2005). The STATIS method is also appropriate for the analysis of these data. A thorough comparison of the respective merits of these three methods of analysis is beyond the scope of this paper. We will content ourselves by comparing the outcomes of this method of analysis to those of GPA and MFA on the basis of a case study.

Methods of cluster analysis of projective mapping/Napping data were proposed (Vidal *et al.*, 2016) and are backed up by the idea that the subjects are likely to use different criteria to assess the products. These authors advocated using the so-called *Lg* measures derived from Multiple Factor Analysis (MFA, Lê and Worch, 2014). Roughly speaking, the *Lg* measure between a dataset associated with a given subject and a component derived from MFA or any other method of analysis can be seen as a quantity that measures the variation in this dataset explained by the component under consideration (Lê and Worch, 2014). By selecting four MFA components, Vidal *et al.* (2016) computed for each dataset associated with a subject four *Lg* measures. Thus, each subject was considered as a four dimensional data point. Thereafter, these data points were subjected to a hierarchical cluster analysis using Ward's criterion. This approach of clustering the subjects is questionable for several reasons. In the first place, it is based on the configuration of the products obtained by means of an analysis performed on the whole panel. In case of a high disagreement among the panellists, which is precisely the reason why we wish to cluster them, this overall configuration is likely to be not very reliable. The second reason is that it appears somehow odd to set up a similarity measure among the datasets on the basis of the variations explained by a set of components when there are direct and easy to interpret measures of similarity among the datasets such as the RV coefficients (Robert and Escoufier, 1976). INDSCAL, which can be used for the analysis of multiblock datasets as an alternative method to MFA (Næs et al, 2017), yields, for each subject (*i.e.*, dataset), a set of saliences which reflect the weights that

this subject attaches to the various INDSCAL components. Technically speaking, these saliences reflect the variation in the datasets explained by the INDSCAL components and, therefore, can be interpreted in a similar way as the *Lg* measures. Jackson (2005) stated that "although one might be tempted to cluster the saliences or carry out other formal analysis on them, these quantities do not lend themselves to that purpose". He gives additional references to support his claim (MacCallum, 1977; Coxon and Davis, 1982). CLUSTATIS can be applied in a straightforward way to the projective mapping or Napping data. By setting the norm of each matrix $\boldsymbol{W_i}$ to 1, we take into account the variations among the subjects in terms of dispersion of the products on the sheets of paper. Some subjects may use the whole surface of the sheet of paper whereas, others may restrict themselves to a smaller part. The fact that we are using the scalar product matrices, which are invariant by rotation, instead of the coordinates of the products on the sheets of paper means that we account for the orientation of the axes which may differ from one subject to another.

2.4. The case of free sorting

As stated above, in a free sorting task, the subjects who participate in the experiment are instructed to partition the products, considering that the products in each group are perceived as similar. The number of groups may differ from one subject to another. Several ways of coding the data are proposed and lead to different methods of analysis of the data from a free sorting task. These methods include strategies pertaining to multidimensional scaling (Lawless, Sheng, Knoops, 1995; Faye *et al.*, 2004; Abdi, Valentin, Chollet and Chrea, 2007), multiple and simple correspondence analysis (Takane, 1981; Qannari*,* Cariou, Teillet, and Schlich, 2010; Cadoret, Lê and Pagès, 2009; Cariou and Qannari, 2018). Of particular interest to us is the DISTATIS method (Abdi *et al.*, 2007), which, similarly to our approach,

revolves around the STATIS method. For this reason, we will give more details regarding this method and compare its outcomes to those of our approach before running the cluster analysis of the subjects.

We adopt herein a coding of the data which is a common practice in correspondence analysis. Let us assume that the subject i $(i = 1, \dots, m)$ has sorted the *n* products at hand into p_i clusters. We denote by $\boldsymbol{Y_i}$ (nxp_i) the matrix of dummy variables indicating for each product the group to which it belongs. Let us denote by $\boldsymbol{y}$ the current column of matrix $\boldsymbol{Y_i}$. This column is a dummy (or 0-1) variable which is associated to a specific group, G (say), of the products defined by the subject under consideration. It indicates for each product whether it belongs to this group (in which case it takes the value 1) or not (in which case it takes the values 0). Let us denote by f the proportion of products in the group G. It is easy to check that f is also the average of $\boldsymbol{y}$**.** The centred and standardized column $\boldsymbol{x}$ associated with $\boldsymbol{y}$ is given by $\boldsymbol{x} = {(\boldsymbol{y} - f)}/{\sqrt{f}}$. This standardization is usual with categorical data, particularly within the framework of correspondence analysis and can be backed by several considerations. More details can be found in the paper by Qannari *et al.* (2010). In the following, we shall denote by $\boldsymbol{X_i}$ the matrix obtained from $\boldsymbol{Y_i}$ by standardization of the columns according to this strategy. It is worth noting that the agreement between two subjects i and i' as assessed by the *RV* coefficients between $\boldsymbol{X_i}$ and $\boldsymbol{X_{i'}}$ is proportional to the chi-square statistic of the contingency table which cross-tabulates the groups of products defined by subject i and those by subject i' (Qannari *et al.*, 2010).

For an overall analysis of the free sorting data, the data tables $\boldsymbol{X_i}$, $(i = 1, \dots, m)$ can be submitted to the STATIS method. By way of clustering the subjects, we propose to perform CLUSTATIS on the matrices $\boldsymbol{X_i}$ $(i = 1, \dots, m)$.

It is worth noting that since we have advocated dividing each matrix $\boldsymbol{W_i} = \boldsymbol{X_i X_i^T}$ by its norm, this amounts to setting all the subjects on the same footing before STATIS or CLUSTATIS analyses are run. Indeed, we can show that the norm of $\boldsymbol{W_i}$ is proportional to $\sqrt{p_i - 1}$, where, as stated above, p_i is the number of groups in the partition of the products defined by the i^{th} subject (Qannari *et al.*, 2010). Indeed, this is an important parameter that obviously has a significant impact on the analysis of free sorting data and which is generally overlooked when analyzing this kind of data.

DISTATIS and the approach regarding the analysis of free sorting data discussed herein share the fact they use the same coding of the sorting data by means of dummy variables and they both use the STATIS method. Basically, the main difference is that, in DISTATIS, the dummy variables are not standardized as advocated herein. By using this kind of standardization, we are, *de facto,* considering methods of analysis that are related to correspondence analysis. It is well known that this method of analysis is better suited to frequency and 0-1 data (Greenacre, 2007). To the credit of DISTATIS, we should put the fact that it allowed new developments that yielded interesting tools to better investigate the sorting data (Lahne, Abdi and Heyman, 2018).

The specific issue of clustering the subjects involved in a free sorting task was addressed by Courcoux, Faye and Qannari (2014). These authors computed a global criterion based on the adjusted rand index (Hubert and Arabie, 1985), which is the corrected-for-chance version of the Rand index (Rand, 1971). This criterion assesses the agreement of the subjects with a consensus partition of the products, to be determined by the algorithm. Thereafter, a strategy of clustering the subjects is set up. It combines an ascending hierarchical strategy of analysis and a partitioning procedure to improve the solution obtained by means of the

hierarchical cluster analysis. The originality of this approach is that it leads to the determination of partitions of the products associated to the various clusters of subjects whereas, in CLUSTATIS, we find latent configurations that makes it possible to depict, for each cluster of subjects, the relationships among the products on the basis of principal components.

In a more recent paper, Cariou and Qannari (2018), investigated the agreement among the subjects involved in a free sorting task by computing a co-occurrence matrix that gives for each pair of subjects the numbers of pairs of products that these subjects set in the same group. This co-occurrence matrix is subjected to correspondence analysis and hierarchical cluster analysis. The aggregation criterion of this latter strategy of analysis is devised so as to preserve as much as possible the χ^2 index associated with the co-occurrence matrix. Therefore, the cluster analysis is aligned with correspondence analysis since both these methods aim at preserving as much as possible the χ^2 index. The advantage of CLUSTATIS over this strategy of analysis is that it has a wider scope than the free sorting data and, as discussed in the conclusion, can be extended in various directions.

III. Illustration

3.1 Projective mapping/Napping data

The data which are used to illustrate the application of STATIS and CLUSTATIS to Projective mapping/Napping can be found in the R package SensoMineR (Lê and Husson, 2008). They concern 8 smoothies which were evaluated by 24 consumers. Thus, the data consist of 24 datasets; each dataset is associated with a consumer and consists of the x-y coordinates of the eight smoothies on the sheet of paper.

It is usually advocated performing either GPA or MFA (Tomic, Berget, Næs, 2015). The STATIS method can also be used to analyze this kind of data. In Figure 2, we show the representation of the products on the basis of the first two components computed by means of STATIS, GPA and MFA. It is clear that there is a high similarity between these three configurations. This is confirmed by the high values of the RV coefficients between the pairs of configurations: RV(STATIS, GPA)=0.99, RV(STATIS,MFA)=0.97, RV(GPA,MFA)=0.96.

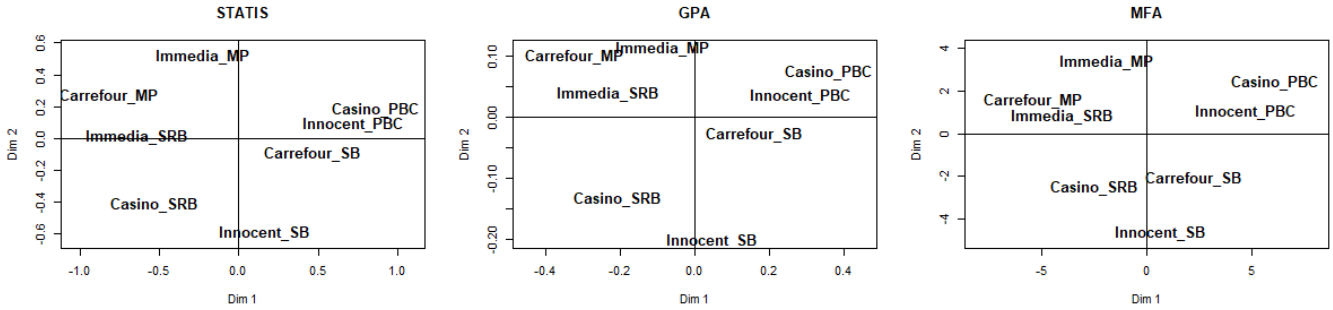

Figure 2. Smoothies data: representation of the products on the basis of the first two components derived from STATIS, GPA and MFA, respectively.

Figure 3 sheds even more light on the differences and similarities between the three methods of analysis considered herein. In this figure, we plot for each of the three methods the cumulative variation in the original datasets explained by the components derived from STATIS, GPA and MFA, respectively. Since all the datasets at hand are two dimensional, the group average configuration derived from GPA is also two dimensional and, therefore, only two components can be computed. These two components explain less than 19% of the variation in the original datasets. The first two components derived from STATIS and MFA explain as much variation as those from GPA. However, with these two latter methods of analysis, further components can be computed to account for additional variation in the datasets. From Figure 3, it is clear that STATIS and MFA show the same pattern in terms of total variance in the original datasets recovered by the successive components.

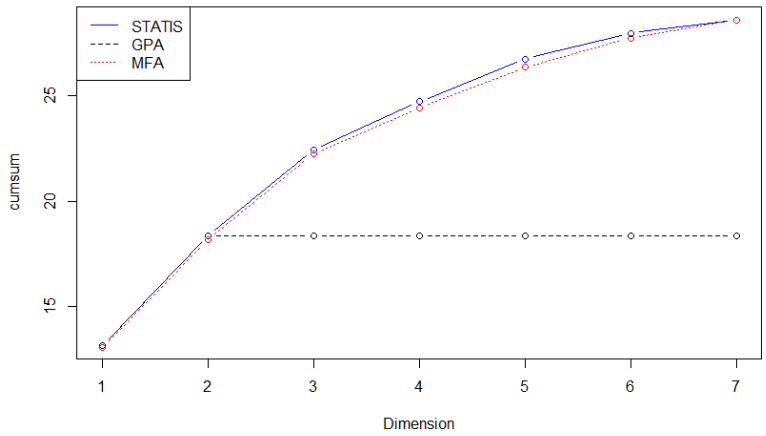

Figure 3. Smoothies data: percentages of total variance in the original datasets explained by the successive components derived from STATIS, GPA and MFA.

Associated with the STATIS method, we computed the overall homogeneity index which was described in section 2.2. It is equal to 42.5 %, indicating a poor agreement among the consumers. A segmentation of these consumers by means of CLUSTATIS should improve this homogeneity within each cluster.

Figure 4 (left) shows the hierarchical tree obtained by running CLUSTATIS on the projective mapping/Napping data. Figure 4 (middle) shows the evolution of the aggregation criterion in the course of the clustering process and, in Figure 4 (right), we depict the evolution of the overall homogeneity index as the number of clusters decreases. These figures suggest to choose three clusters since there is a significant jump of the criterion D when passing from a partition in three clusters to a partition with two clusters. The solution thus obtained is further improved by running the partitioning algorithm associated with CLUSTATIS. As a matter of fact, only one subject changed cluster membership so much so that the overall homogeneity index improved but only slightly (0.5%). This particular point will be further discussed in the conclusion. Table 1 gives the homogeneity index associated

within each cluster of subjects together with the overall homogeneity index. The third cluster, which is the largest cluster, has the smallest homogeneity index (54.5%). Figure 5 shows the RV coefficients of each subject's configuration with the group average configuration of the cluster to which this subject belongs. It can be seen that overall these coefficients are large. However, some few RV coefficients are relatively small. This evidences marginal subjects who do not properly fit in any cluster. This point will be further discussed in the conclusion.

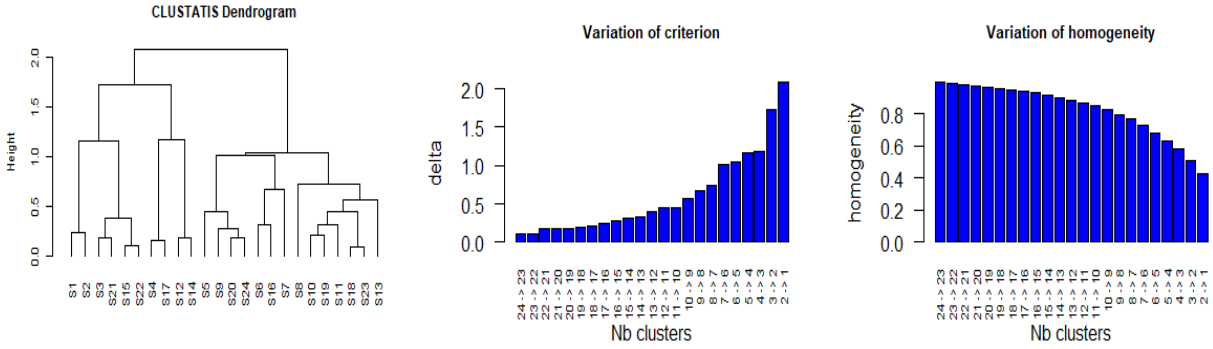

Figure 4. Smoothies data: dendrogram (left), variation of the aggregation criterion (middle) and overall homogeneity index (right) as functions of the number of clusters.

By way of assessing the significance of the overall homogeneity index associated with the partition obtained by means of CLUSTATIS, we performed, as suggested by a reviewer of this paper, a simulation study whereby 10000 random partitions of the subjects were generated, keeping the sizes of the clusters equal to those obtained by means of CLUSTATIS. For each partition, we computed the overall homogeneity index. It turned out that the average of all these simulated indices was equal to 47.1% and the largest value is equal to 54.0%, which is smaller than the overall homogeneity index associated with the CLUSTATIS partition (58.9%).

Figure 6 shows the representation of the products on the basis of the first two columns of $\boldsymbol{C}^{(k)}$ $(k = 1,\ 2,\ 3)$, where, as stated above, $\boldsymbol{C}^{(k)}$ is the group average configuration associated with the cluster G_k. There are obvious differences between these three configurations, particularly regarding the smoothie "Carrefour_SB". Table 2 shows the RV coefficients between the group average configurations associated with the three clusters. This table indicates that the first cluster is closer to the third cluster than it is to the second cluster. Table 2 also shows the RV coefficients of each cluster configuration and the configuration associated with the whole panel. It appears that the largest cluster (cluster 3) is the closest to the global configuration, whereas the second cluster, which is the smallest, has a relatively small RV coefficient with the global configuration.

Cluster	Number of consumers	Homogeneity (%)
1	7	65.4
2	4	61.5
3	13	54.5
Overall index	-	58.9
The whole panel	24	42.5

Table 1: Smoothies data: *sizes and homogeneity indices of the three clusters.*

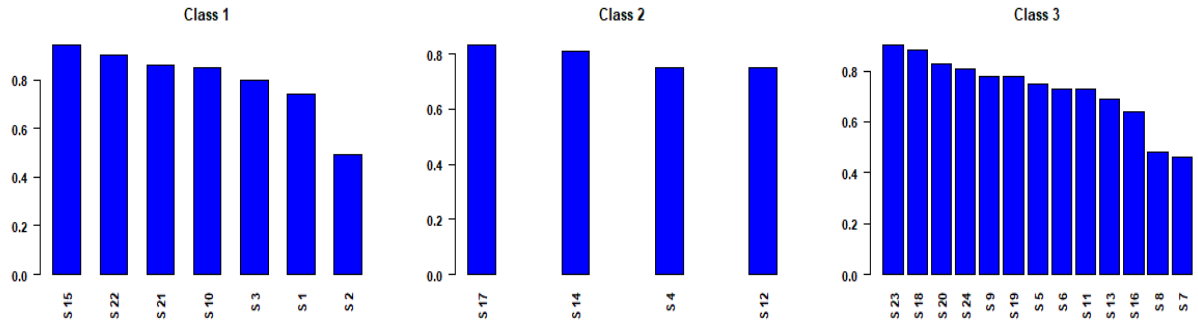

Figure 5. Smoothies data: RV coefficients of the datasets associated with subjects S1 to S24 with the group average configurations of the clusters to which they belong.

	Cluster 1	Cluster 2	Cluster 3	Overall
Cluster 1	1	0.37	0.63	0.84
Cluster 2		1	0.35	0.49
Cluster 3			1	0.94
Overall				1

Table 2: Smoothies data: *RV coefficients between the group average configurations of the three clusters and between the configuration of each cluster and the global configuration associated with the whole panel.*

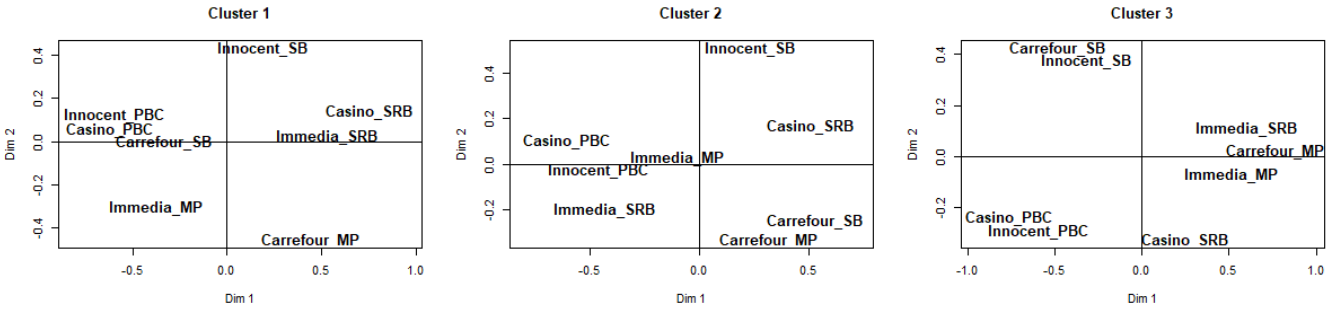

Figure 6. Smoothies data: representation of the smoothies on the basis of the first two STATIS components in each cluster.

3.2 Free sorting data

A panel of 25 subjects from Product Perceptions Ltd.'s pool of sensory panelists participated to the free sorting task. They were instructed to sort 14 brands of chocolate. As a matter of fact, the data considered herein are extracted from a larger experiment the details of which are given in Courcoux *et al.* (2012). The subjects created between 5 and 10 groups of products.

Before undertaking the segmentation of the panellists by means of CLUSTATIS, we compared the outputs of the STATIS method applied on the standardized data as indicated in section 2.4 to those of well established methods. A particular emphasis is put on the DISTATIS method (Abdi *et al.*, 2007) since it is tightly linked to STATIS.

Figure 7 (left) shows the position of the products on the basis of the first two components obtained by means of STATIS applied to the sorting data. The homogeneity index is equal to 63.9%. It indicates a fair agreement among the subjects. Figure 7 (right) shows the configuration of the products on the basis of the first two components obtained by means of DISTATIS. There are clear similarities between these two configurations. For instance, we can see that, the pairs of products CDM and Galaxy, on the one hand, and Tesco Value and JSValue, on the other hand, are close to each other and each of these pairs of products is far removed from the other. However, there are also striking differences among the two configurations. For instance, the chocolate 'Divine' is rather extreme for both the first two components from STATIS, whereas it occupies a central position in the configuration from DISTATIS. These similarities and differences are reflected by the RV coefficient between these two configurations which is equal to 0.71, indicating a fair agreement between the two methods of analysis. As a matter of fact, the differences from these two methods stem from the standardisation that we have adopted in section 2.4. If we use the dummy variables without standardization, then the RV coefficient between the two dimensional configurations obtained by means of STATIS and DISTATIS jumps to 0.96, indicating a high agreement among the two configurations.

It is worth noting that we also compared the configurations obtained by means of STATIS as described herein, DISTATIS, multiple correspondence analysis (Takane, 1981, Qannari *et al.*, 2010; Cadoret, Lê and Pagès, 2009), simple correspondence analysis (Cariou and Qannari, 2018) and MDS (Lawless, Sheng, Knoops, 1995; Faye *et al.*, 2004). Not surprisingly, the configurations from correspondence analysis (simple and multiple) and STATIS were in a very high agreement (RV>0.94). The RV coefficients of these configurations with that obtained by means of MDS were also relatively high (RV>0.82). By contrast, the RV

coefficient between the configuration obtained by means of DISTATIS and that obtained by means of MDS was as small as 0.69. Clearly, DISTATIS seems to stand apart from the other methods.

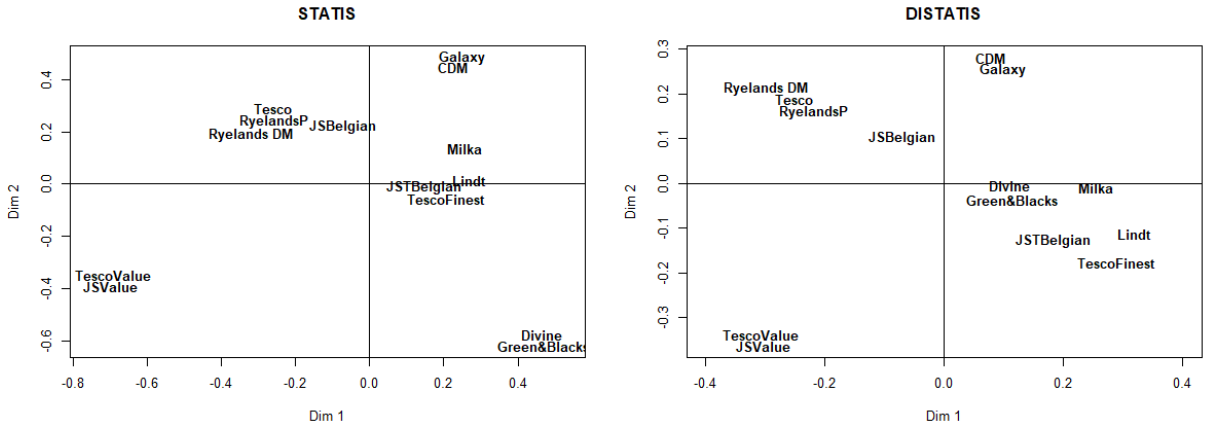

Figure 7. Chocolate data: representation of the products on the basis of the first two components from STATIS (left) and DISTATIS (right).

Following the strategy of analysis described in section 2.4, we performed CLUSTATIS. Figure 8 (left) shows the dendrogram associated with the hierarchical cluster analysis. Figure 8 (middle) shows the evolution of the aggregation criterion in the course of the merging process. Figure 8 (right) shows the evolution of the overall homogeneity index as the number of clusters decreases. These figures indicate to consider two clusters. The partition obtained by cutting the dendrogram at the level corresponding to two clusters was submitted to the partitioning algorithm in an attempt to improve it but no subject changed cluster membership. The first cluster (size=15) has a homogeneity index equal to 68.5% whereas, the second cluster (size=10) has a homogeneity index equal to 71.5%. The overall homogeneity index, which is as stated above a weighted average of these two indices, is equal to I= 69.7%. This represents a rather small improvement over the agreement at the level of the whole panel (63.9%). The fact that this sorting task was performed by panellists with a good experience in sensory evaluation may be an explanation of this finding.

Figure 9 (left and right) shows the representation of the brands of chocolate on the basis of the first two components associated with the first and second clusters of subjects, respectively. We can see that, for the first cluster of subjects (Figure 9-left), the products JSvalue and Tesco value are singled out but not in Figure 9 (right) associated with the second cluster. More importantly, in the second cluster, the products Green&Black and Divine are far removed from the other products and from each other, which is not the case for the first cluster. Another difference concerns the products JS Belgian and JST Belgian which are very close to each other in the configuration associated with the second cluster but not in that associated with the first cluster.

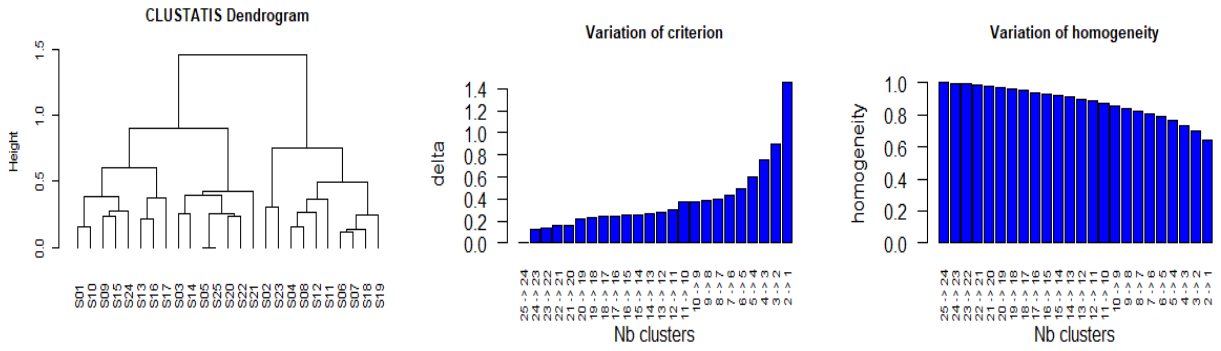

Figure 8. Chocolate data: CLUSTATIS hierarchical analysis

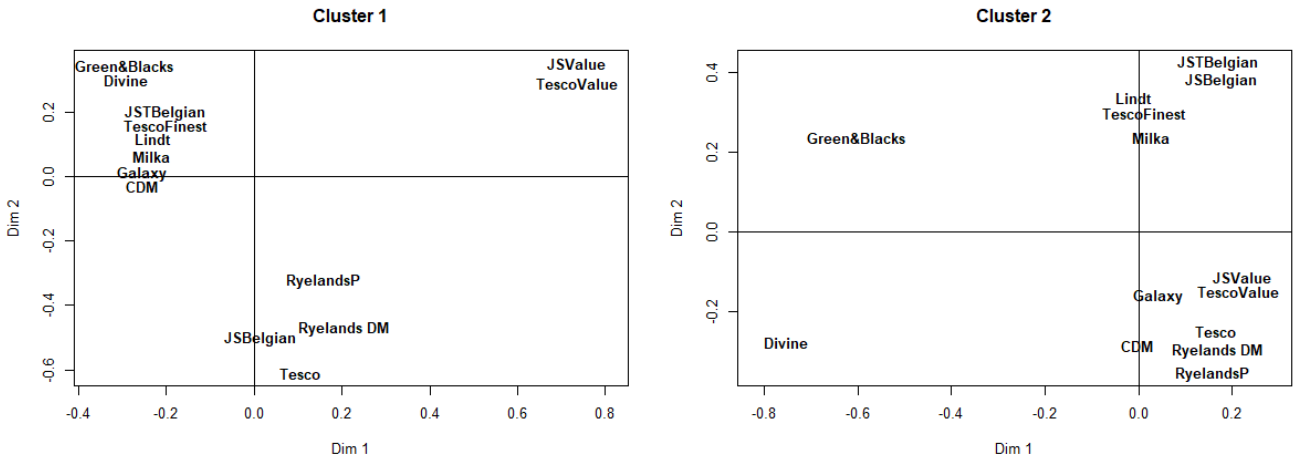

Figure 9. Chocolate data: Representation of the chocolate brands on the basis of the first two STATIS *components in each cluster.*

In table 3, we give the RV coefficients between the configurations associated with Cluster 1 and Cluster 2. This coefficient (0.82) is far agreement between the two clusters. In table 3, we also give the RV coefficients between the configurations associated with the two clusters and the configuration associated with the whole panel. Both the RV coefficients are relatively large. All these results seem to confirm that, in this case study, the panel of subjects is relatively homogeneous.

	Cluster 1	Cluster 2	Overall
Cluster 1	1	0.82	0.97
Cluster 2		1	0.94
Overall			1

Table 3: Chocolate data: *RV between the group average configurations of the two clusters and among the overall compromise (got by the STATIS method).*

As in the previous case study pertaining to the projective mapping/Napping experiment, we performed a simulation study whereby 10000 random partitions of the subjects were generated, keeping the sizes of the clusters equal to those obtained by means of CLUSTATIS. For each partition, we computed the overall homogeneity index. It turned out that the average of all these simulated indices was equal to 65.3% and the largest value is

equal to 68.3%, which is slightly smaller than the overall homogeneity index associated with the CLUSTATIS partition (69.7%).

IV. Conclusion

The CLUSTATIS method is a clustering approach for multiblock datasets. Not only it leads to the clustering of a set of datasets but, within each cluster, it yields a group average configuration which makes it possible to depict the relationships among the individuals (*i.e.*, rows of the datasets). This group average configuration, also called compromise, is computed by means of the STATIS method applied to the datasets of the cluster under consideration. This means that weights associated with the various datasets are determined. These weights reflect the extent to which each dataset agrees with the general "point of view" of the datasets and are used to compute the group average configuration. This entails that those datasets which are in less agreement with the others are down-weighted.

CLUSTATIS provides several indices and tools to select the appropriate number of clusters and to assess the relevance of the obtained solution, the homogeneity of each cluster, the centrality of each dataset within its cluster, the similarity among the clusters, etc.

We have discussed the application of CLUSTATIS to projective mapping/Napping and free sorting data. Obviously, the range of application of this strategy of clustering is very wide and is getting wider as time goes by since the collection of multiblock data is becoming more and more frequent in several domains of applications such as metabolomics, genomics, etc.

From a technical point of view, CLUSTATIS offers two distinct strategies of clustering (*i.e.*, hierarchical and partitioning algorithms) and it is advocated using both of them since

they complement each other. The hierarchical cluster analysis helps the practitioner to choose the appropriate number of clusters and provides an initial solution to the partitioning algorithm. Notwithstanding, each of these two strategies of clustering can stand by itself. We have seen in the first case study that the solution provided by the hierarchical clustering was but very marginally improved since only one subject changed membership. In the second case study, the partitioning algorithm could not further improve the solution obtained by means of the hierarchical algorithm. One should not jump to the conclusion that the partitioning algorithm is redundant. Indeed, there may be situations where it can significantly improve the solution from the hierarchical strategy of clustering. Moreover, we should bear in mind that the hierarchical clustering is time consuming particularly if the number of datasets is very large. In such situations, the partitioning algorithm can be used by itself by considering several initial solutions and, eventually, choosing the best solution that corresponds to the smallest criterion *D*. The difficulty with this strategy of analysis is the choice of the appropriate number of clusters, which is a tricky problem.

CLUSTATIS was designed to be an extension of the CLV approach (Vigneau and Qannari, 2003; Vigneau, Chen and Qannari, 2015). As stated above, this latter approach is concerned with the cluster analysis of variables and have enjoyed some popularity among practitioners in various domains of analysis. As a matter of fact, this method of analysis has several interesting options that can be adapted to the CLUSTATIS approach. For instance, we may be interested in clustering several consumers' datasets (*e.g.*, liking scores for several criteria) taking account of an external dataset (*e.g.*, sensory data). We may also be interested in clustering a collection of datasets (*e.g.*, from a sorting or Napping tasks) while setting aside outlier panellists (Vigneau, Qannari, Navez, Cottet, 2016). This will result in a more stable

and better interpretable clustering of the datasets. Investigations regarding these extensions are currently underway and the findings will be reported elsewhere.

For the illustration of CLUSTATIS, we have used case studies which involve a relatively small number of subjects. The efficiency of the clustering approach would be better demonstrated with case studies involving more subjects. Furthermore, besides the homogeneity indices, more validation tools can be used to better assess the relevance of the results (Vigneau et al, 2016; Brock, Pihur, Datta, Datta, 2008).

Regarding the choice of the number of clusters, we have relied on the structure of the hierarchical tree which reflects the evolution of the aggregation criterion, but this procedure remains more or less subjective. There is a vast literature concerning the tricky problem of selecting the number of clusters in a clustering analysis (Charrad *et al.*, 2014; Sugar, James, 2003). Further investigations should concern the adaptation of some of these strategies of selecting the appropriate number of clusters to CLUSTATIS.

References

Abdi, H., Valentin, D., Chollet, S., and Chrea, C. (2007). Analyzing assessors and products in sorting tasks: DISTATIS, theory and applications. Food quality and preference, 18(4), 627-640.

Brock, G., Pihur, V., Datta, S., & Datta, S. (2011). clValid, an R package for cluster validation. Journal of Statistical Software (Brock et al., March 2008).

Cadoret, M., Lê, S., and Pagès, J. (2009). A factorial approach for sorting task data (FAST). Food Quality and Preference, 20(6), 410-417.

Cariou, V., and Qannari, E. M. (2018). Statistical treatment of free sorting data by means of correspondence and cluster analyses. Food Quality and Preference, 68, 1-11.

Cariou, V., and Wilderjans, T. F. (2018). Consumer segmentation in multi-attribute product evaluation by means of non-negatively constrained CLV3W. Food Quality and Preference 67, 18-26.

Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A., & Charrad, M. M. (2014). Package 'NbClust'. Journal of Statistical Software, 61, 1-36.

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. Food quality and preference, 32, 107-112.

Courcoux, P., Qannari, E. M., Taylor, Y., Buck, D., and Greenhoff, K. (2012). Taxonomic free sorting. Food Quality and Preference, 23(1), 30-35.

Coxon, A. P. M., and Davies, P. M. (1982). The user's guide to multidimensional scaling: With special reference to the MDS (X) library of computer programs. Heinemann.

Dahl, T., and Næs, T. (2004). Outlier and group detection in sensory panels using hierarchical cluster analysis with the Procrustes distance. Food Quality and Preference, 15(3), 195-208.

El Ghaziri, A., and Qannari, E. M. (2015). Measures of association between two datasets; application to sensory data. Food Quality and Preference, 40, 116-124.

Everitt, B. S., Landau, S., Leese, M., and Stahl, D. (2011). Cluster Analysis, 5th Edition, 71-110.

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

Gower, J. C. (1975). Generalized procrustes analysis. Psychometrika, 40(1), 33-51.

Hubert L. J., Arabie P. (1985). Comparing partitions. Journal of Classification, 2, 193-218.

Jackson, J. E. (2005). A user's guide to principal components (Vol. 587). John Wiley and Sons.

Lahne J., Abdi H., Hildegarde H. (2018). Rapid sensory profiles with DISTATIS and Barycentric Text Projection: An example with amari, bitter herbal liqueurs, Food Quality and Preference, 66, 36-43.

Lavit, C., Escoufier, Y., Sabatier, R., & Traissac, P. (1994). The act (statis method). Computational Statistics & Data Analysis, 18(1), 97-119.

Lawless, H. T., Sheng, N., and Knoops, S. S. (1995). Multidimensional scaling of sorting data applied to cheese perception. Food Quality and Preference, 6(2), 91-98.

Lê, S., and Worch, T. (2014). Analyzing sensory data with R. CRC Press.

Lê, S., and Husson, F. (2008). SensoMineR: A package for sensory data analysis. Journal of Sensory Studies, 23(1), 14-25.

Lloyd, S. P. (1982). Least squares quantization in PCM. IEEE transactions on information theory, 28(2), 129-137.

MacCallum, R. C. (1977). Effects of conditionality on INDSCAL and ALSCAL weights. Psychometrika, 42(2), 297-305.

Næs, T., Berget, I., Liland, K. H., Ares, G., and Varela, P. (2017). Estimating and interpreting more than two consensus components in projective mapping: INDSCAL vs. multiple factor analysis (MFA). Food quality and preference, 58, 45-60.

Pagès, J. (2005). Collection and analysis of perceived product inter-distances using multiple factor analysis: Application to the study of 10 white wines from the Loire Valley. Food quality and preference, 16(7), 642-649.

Pizarro, C., Esteban-Díez, I., Rodríguez-Tecedor, S., and González-Sáiz, J. M. (2013). A sensory approach for the monitoring of accelerated red wine aging processes using multi-block methods. Food quality and preference, 28(2), 519-530.

Qannari, E. M., Cariou, V., Teillet, E., and Schlich, P. (2010). SORT-CC: A procedure for the statistical treatment of free sorting data. Food quality and preference, 21(3), 302-308.

Rand W. (1971). Objective criteria for the evaluation of clustering methods. Journal of American Statistical Association, 66, 846-850.

Risvik, E., McEwan, J. A., Colwill, J. S., Rogers, R., and Lyon, D. H. (1994). Projective mapping: A tool for sensory analysis and consumer research. Food quality and preference, 5(4), 263-269.

Risvik, E., McEwan, J. A., and Rødbotten, M. (1997). Evaluation of sensory profiling and projective mapping data. Food quality and preference, 8(1), 63-71.

Robert, P., and Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods: the RV-coefficient. Applied statistics, 257-265.

Schlich (1996). Defining and validating assessor compromises about product distances and attribute Correlations. In: Næs and Risvik (ed.): Multivariate Analysis of Data in Sensory Science, 259-306.

Sugar, C. A., & James, G. M. (2003). Finding the number of clusters in a dataset: An information-theoretic approach. Journal of the American Statistical Association, 98(463), 750-763.

Takane, Y. (1981). MDSORT: A special-purpose multidimensional scaling program for sorting data. Behavior Research Methods, 13(5), 698-698.

Tomic, O., Berget, I., and Næs, T. (2015). A comparison of generalised procrustes analysis and multiple factor analysis for projective mapping data. Food quality and preference, 43, 34-46.

Vidal, L., Antúnez, L., Giménez, A., Varela, P., Deliza, R., and Ares, G. (2016). Can consumer segmentation in projective mapping contribute to a better understanding of consumer perception? Food quality and preference, 47, 64-72.

Vigneau, E., Qannari, E. M. (2003). Clustering of variables around latent components. Communications in Statistics-Simulation and Computation, 32, 4, 1131-1150.

Vigneau E., Chen M., Qannari E. M. (2015). ClustVarLV: An R package for the clustering of variables around latent variables. Rjournal, 7, 2, 134-148.

Vigneau E., Qannari E. M., Navez B., Cottet V. (2016). Segmentation of consumers in preference studies while setting aside atypical or irrelevant consumers. Food Quality and Preference, 47, A, 54-63.

Wilderjans, T. F., Cariou, V. (2016). CLV3W: A clustering around latent variables approach to detect panel disagreement in three-way conventional sensory profiling data. Food Quality and Preference, 47, 45-53.

Wilderjans, T. F., & Ceulemans, E. (2013). Clusterwise Parafac to identify heterogeneity in three-way data. Chemometrics and Intelligent Laboratory Systems, 129, 87-97.