Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v13n2p436

CLUSTATIS: Cluster analysis of blocks of variables

By Llobell, Qannari

Published: 14 October 2020

Electronic Journal of Applied Statistical Analysis
Vol. 13, Issue 02, October 2020, 436-453
DOI: 10.1285/i20705948v13n2p436

CLUSTATIS: Cluster analysis of blocks of variables

Fabien Llobell*[a,b] and El Mostafa Qannari[a]

[a] *StatSC, ONIRIS, INRA, Nantes, France*
[b] *Addinsoft, XLSTAT, Paris, France*

Published: 14 October 2020

The STATIS method is one of many strategies of analysis devoted to the unsupervised analysis of multiblock data. A new optimization criterion to define this method of analysis is introduced and an extension to the cluster analysis of several blocks of variables is discussed. This consists in a hierarchical cluster analysis and a partitioning algorithm akin to the K-means algorithm. Moreover, in order to improve the cluster analysis outcomes, an additional cluster called noise cluster which contains atypical blocks of variables is introduced. The general strategy of analysis is illustrated by means of two cases studies.

keywords: Cluster analysis, Multiblock data, STATIS, CLUSTATIS, Noise cluster.

1. Introduction

The instances where practitioners in different fields of application are required to collect multiblock data are getting more and more frequent. For instance, in consumer studies, a panel of consumers may be asked to express their appreciation of a set of products for various attributes. In such a situation, we end up by having as many blocks of variables as consumers, each block having the products as rows and the attributes as columns. In other situations, the various blocks at hand may not refer to the same variables. For instance, in sensory analysis, and more particularly in the evaluation procedure called free choice profiling (Jack and Piggott, 1991), a panel of subjects are instructed to assess the intensity of several sensory variables for a set of products. Moreover, each

*Corresponding author: fllobell@xlstat.com

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

subject is free to choose his or her own list of variables. In these situations, it is clear that the issue regarding the analysis and the clustering of the blocks of variables are of paramount interest. For the analysis of multiblock data, there are a plethora of methods since this topic has been a burning issue for the last thirty years or so (De Roover et al., 2012). For the cluster analysis of multiblock data, the list of methods is much more limited. Dahl and Næs (2004) used cluster analysis in sensory evaluation with the aim of identifying homogeneous sub-groups of panellists and outlying panellists. For this purpose, they computed the Procrustes distances between pairs of blocks of variables and subjected the distance matrix thus obtained to hierarchical cluster analysis. Cariou and Wilderjans (2018) proposed a strategy of clustering multiblock data, called CLV3W, which is particularly designed for the clustering of three way data where the blocks of variables refer to the same individuals and the same variables. We advocate using the STATIS method (Lavit et al., 1994) for the analysis of multiblock data and we propose a method of analysis, called CLUSTATIS, which is tightly linked to STATIS for the clustering of blocks of variables. CLUSTATIS consists in a hierarchical cluster analysis and a partitioning algorithm. Both these two strategies aim at optimizing the same criterion and can be run independently or in combination in an attempt to achieve an even better solution than that obtained by running one or the other of the two strategies alone. CLUSTATIS can be seen as an extension of the cluster analysis of variables called CLV (Vigneau and Qannari, 2003) to the case of blocks of variables. Indeed, these two methods have the same rationale and follow the same pattern of analysis. CLV aims at clustering the variables at hand around latent components, whereas CLUSTATIS aims at clustering the blocks of variables around latent configurations. Within each cluster, this latent configuration is obtained by means of the STATIS method and can be used to depict the relationships among the individuals. In section 2 devoted to the material and methods, we start by sketching a reminder of the STATIS method and we introduce a new criterion to define this strategy of analysis (subsection 2.1). Based on this criterion, we discuss a general strategy of cluster analysis of several blocks of variables (subsection 2.2). By way of improving the outcomes of the cluster analysis, we outline in subsection 2.3 how the atypical blocks of variables could be set aside following the concept of noise cluster introduced by Dave (1991). In section 3, we illustrate the approach on the basis of case studies. Finally, we end the paper by some concluding remarks.

2. Methods

2.1. A new criterion for the STATIS method

Let us consider m blocks of variables denoted by $X_1, \ldots, X_m$, which are assumed to be column centred. These blocks of variables are measured on the same n individuals but the variables may not be the same from one block to another. The Figure 1 represents the data structure of the blocks of variables.

The STATIS method is based on the scalar product matrices associated with the blocks of variables at hand. These matrices are computed as follows: $W_1 = X_1 X_1^\top, \ldots, W_m =$

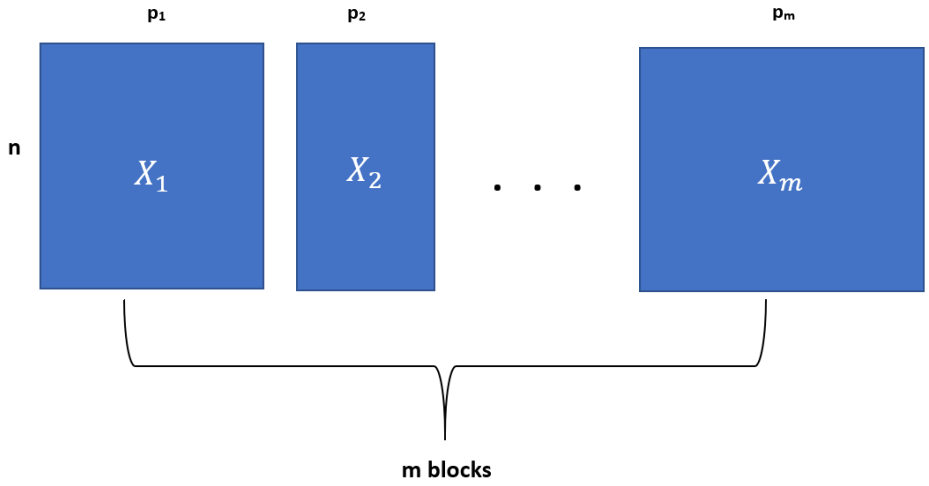

Figure 1.: m blocks measured on the same individuals.

$X_m X_m^\top$. We assume that these matrices are pre-scale so as to have their norm equal to 1. This is achieved by dividing each W_i by its Frobenius norm, namely $||W_i|| = \sqrt{trace(W_i W_i)}$.

The scalar product between two matrices W_i and W_s (for $i, s = 1, \ldots, m$) is given by $trace(W_i W_s)$. Since W_i and W_s are assumed to be of norm equal to 1, this quantity is equal to the so-called RV coefficient (Robert and Escoufier, 1976), which is central to the STATIS method and whose popularity goes far beyond this method of analysis (Schlich, 1996; El Ghaziri and Qannari, 2015). In the following, we shall refer to this coefficient as $RV(W_i, W_s)$. It reflects the similarity between the blocks of variables X_i and X_s. It ranges between 0 and 1; it is equal to 0 if all the variables in X_i are orthogonal to those in X_s and it is equal to 1 if X_i and X_s can be matched by a rotation and multiplication by a scalar factor (Glaçon, 1981).

The STATIS method can be defined by means of the following original criterion. We seek a group average matrix, W, and weighting scalars, α_i, assumed to be such that $\sum_{i=1}^m \alpha_i^2 = 1$ so as to minimize the following quantity:

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

We can show (see appendix A) that the weights, α_i, are obtained by computing the eigenvector of the matrix which contains the pairwise RV coefficients between the various blocks of variables associated with the largest eigenvalue that we shall denote by λ_1. The matrix W is given by $W = \sum_{i=1}^m \alpha_i W_i$. The rationale behind this solution is that the weight α_i is relatively large if X_i tends to agree with the other blocks of variables. Contrariwise, the weighting coefficients tend to be relatively small for differing X_i. By

considering the spectral decomposition of W, we can write $W = CC^\top$. The matrix C is the group average configuration of the blocks of variables $X_1, \ldots, X_m$, and can be used to depict the relationships among the individuals.

Additional properties also shown in appendix A are the following. Most of these properties are already known (see for instance Robert and Escoufier 1976; Glaçon 1981).

(i) $||W||^2 = \sum_{i=1}^m RV^2(W_i, W) = \lambda_1$

(ii) $\alpha_i = \frac{trace(W_i,W)}{||W||^2} = \frac{RV(W_i,W)}{\sqrt{\lambda_1}}$

(iii) $\sum_{i=1}^m ||W_i - \alpha_i W||^2 = m - \sum_{i=1}^m RV^2(W_i, W) = m - \lambda_1$

(iv) $\sum_{i=1}^m ||W_i||^2 = m = ||W||^2 + \sum_{i=1}^m ||W_i - \alpha_i W||^2 = \lambda_1 + \sum_{i=1}^m ||W_i - \alpha_i W||^2$

From the property (i), it emerges that λ_1 can be seen as an overall agreement or homogeneity index between the various blocks of variables since it reflects the extent to which the blocks at hand are related to the group average configuration W. The property (iv) can be interpreted by stating that the total variation measured by $\sum_{i=1}^m ||W_i||^2 = m$ can be decomposed into a variation explained by the group average configuration (*i.e.*, $||W||^2 = \lambda_1$) and a residual variation (*i.e.*, $\sum_{i=1}^m ||W_i - \alpha_i W||^2$). It follows that the index $I = \lambda_1/m$ reflects the part of variation in the various matrices W_i explained by W. This index ranges between $1/m$ and 1. The larger this index, the higher is the agreement among the blocks of variables $X_1, \ldots, X_m$.

2.2. The CLUSTATIS approach

We propose a cluster analysis approach of multiblock data. It aims at minimizing a criterion which reflects the fact that we are seeking homogeneous clusters of blocks of variables. More precisely, the blocks of variables 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 $X_1, \ldots, X_m$ the blocks of variables at hand, which are assumed to be centered. We compute the scalar product matrices: $W_1 = X_1 X_1^\top, \ldots, W_m = X_m X_m^\top$. These matrices are pre-scaled so as to have their Frobenius norm equal to 1. We seek to determine K clusters of blocks of variables $G_1, \ldots, G_K$ so as to minimize the following criterion:

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

where α_i ($i \in G_k$) are scalars to be determined and assumed to be such that $\sum_{i \in G_k} \alpha_i^2 = 1$, and, for $k = 1, \ldots, K$, $W^{(k)}$ is the group average configuration of cluster G_k. Obviously, when there is only one cluster of blocks of variables (*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 above 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 blocks of variables that can be improved by means of the partitioning algorithm. This latter step is called "consolidation" in the French literature (Saporta, 2006).

The hierarchical algorithm follows an ascending (or merging) strategy. We start with the situation where each block of variables forms a group by itself. Obviously, in this case, $D = 0$. At each step, we merge two blocks of variables or, more generally as the algorithm proceeds, two groups of blocks until all the blocks are merged in a single cluster.

From the properties of the STATIS method stated above and more precisely the property (iii), it follows at the current stage, the criterion D is equal to $D = m - \sum_{k=1}^{K} \lambda_1^{(k)}$, where $\lambda_1^{(k)}$ is the largest eigenvalue of the matrix which contains the pairwise RV coefficients of the blocks of variables in group G_k. In the course of the hierarchical algorithm, the variation of criterion D when two groups of blocks of variables A and B (say) are merged is equal to $D_{A\cup B} - D_{A+B} = \lambda_1^{(A)} + \lambda_1^{(B)} - \lambda_1^{(A\cup B)}$, where $\lambda_1^{(A)}$, $\lambda_1^{(B)}$ and $\lambda_1^{(A\cup B)}$ are respectively the largest eigenvalue of the matrix of the RV coefficient between pairs of configurations in clusters A, B and $A\cup B$. We can show that this variation is positive (see appendix B). This means that the merging of two clusters A and B ineluctably results in a deterioration of the within clusters homogeneity (*i.e.*, increase of criterion D). The rationale of the aggregation strategy is to merge those two clusters A and B which result in the smallest increase of criterion D (*i.e.* the smallest variation $\lambda_1^{(A)} + \lambda_1^{(B)} - \lambda_1^{(A\cup B)}$).

One should trace the increase of the criterion D at each stage of the hierarchical cluster analysis because it reflects the increase of the within-cluster heterogeneity as the merging strategy proceeds. A jump of this quantity indicates that we are trying to merge two clusters which are heterogeneous and this should be considered as a signal that the merging strategy should be stopped. In practice, the increase of criterion D is reflected in the hierarchical tree (or dendrogram) as the height of the branches that connect two embedded nodes. Alternatively, the jump between successive steps could be depicted using bar plots showing their evolution as the number of clusters decreases.

The clustering problem based on criterion D given above can also be solved by means of a partitioning algorithm akin to the K-means algorithm (Everitt et al., 2011). In the course of this algorithm, the blocks of variables 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 blocks of variables: K groups of blocks are given by the practitioner. These clusters could be chosen by a random assignment of the

blocks to the K groups. A better initialization can be made from the outcomes of the hierarchical clustering described above. This point will be further discussed below.

- Step 2 (Determination of the group average scalar product matrix in the various clusters): In cluster G_k, $W^{(k)}$ and the associated weights α_i are determined by means of the STATIS method as described above.

- Step 3 (changing clusters): New clusters of blocks of variables are formed by moving each block, X_i, to the cluster G_k for which $||W_i - \alpha_i W^{(k)}||$ is the smallest or, equivalently, $RV(W_i, W^{(k)})$ is the largest. This equivalence stems from the fact that $||W_i - \alpha_i W^{(k)}||^2 = 1 - RV^2(W_i, W^{(k)})$ (property (iii)). The steps 2 and 3 are iterated until convergence, that is the criterion D stops to decrease.

It is worth noting that the partitioning algorithms also seeks to maximize the quantity $\sum_{k=1}^{K} \sum_{i \in G_k} RV^2(W_i, W^{(k)}) = \sum_{k=1}^{K} \lambda_1^{(k)}$. This property can be derived from the property (iii): $D = \sum_{k=1}^{K} \sum_{i \in G_k} ||W_i - \alpha_i W^{(k)}||^2 = m - \sum_{k=1}^{K} \sum_{i \in G_k} RV^2(W_i, W^{(k)})$.

In practice, both the hierarchical and the partitioning algorithms should be performed to reach a better solution than when each algorithm is performed alone. Firstly, the hierarchical strategy can be used to hint to an appropriate number of clusters by examining the evolution of the aggregation criterion, $\Delta = \lambda_1^{(A)} + \lambda_1^{(B)} - \lambda_1^{(A \cup B)}$ in the course of the clustering process. Secondly, the blocks of variables 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 changing of cluster memberships, when running the partitioning algorithm, the solution obtained by the hierarchical clustering is likely to be improved since we attempt to further minimize the criterion D.

Several indices associated with the final solution are of paramount interest. In the first place, we consider for each cluster G_k, $(k = 1, \ldots, K)$, 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 blocks of variables in group G_k and m_k is the number of blocks in this group. This index ranges between $1/m_k$ and 1 and reflects the homogeneity in G_k. An overall index to assess the quality of the partition of the blocks of variables obtained by the clustering approach is given by the weighted average of the indices I_k : $I = \sum_{k=1}^{K} \frac{m_k I_k}{m} = \sum_{k=1}^{K} \frac{\lambda_1^{(k)}}{m}$. This index can be interpreted as the percentage of variation in the original blocks of variables explained by the group average configurations in the various groups. Within the group G_k, we can compute for each block of variables X_i $(i \in G_k)$, the RV coefficient between W_i and $W^{(k)}$. This index reflects how each block of variables is close to its associated group configuration. Alternatively, we could consider the coefficient α_i $(i = 1, \ldots, m)$ which reflects the same idea. Finally, in order to assess how the various groups of blocks of variables are close to each others, we can compute the RV coefficients between their

associated group average configurations. These indices will be illustrated through the two case studies in section 3.

2.3. Cluster analysis while setting aside atypical blocks of variables

2.3.1. The "K+1" or noise cluster

When clustering objects, it often occurs that some atypical objects do not fit any pattern of the clusters that have been determined. By identifying these atypical objects and setting them aside, the clustering solution is likely to be of better quality. Dave (1991) introduced the concept of "noise cluster" or "K+1 cluster", which contains the objects (blocks of variables in our case) which are deemed to be atypical. Our strategy of analysis draws from this concept.

In a first stage, we choose a threshold value, ρ, between 0 and 1, below which the link between two blocks of variables as measured by the RV coefficient will be considered as insignificant. In a second stage, we perform an iterative algorithm akin to the "K-means" with the aim to maximizing the following quantity:

$$H_\rho = \sum_{k=1}^{K} \sum_{i=1}^{m} (\delta_i^k RV^2(W_i, W^{(k)}) + \delta_i^{K+1} \rho^2) \tag{3}$$

where δ_i^k is the Kronecker symbol which is equal to 1 if the block of variables X_i belongs to cluster k and 0 otherwise, with the constraint $\sum_{k=1}^{K+1} \delta_i^k = 1$. This constraint entails that a block of variables belongs to one and only one cluster, including the "K+1" cluster.

The rationale behind this criterion is clear: each block of variables, X_i, is assigned to a cluster G_k for which $RV(W_i, W^{(k)})$ is the largest. However, there is a requirement that this quantity should be larger than ρ. If this requirement is not fulfilled, then X_i is assigned to the "K+1" cluster because its similarity (*i.e.*, $RV(W_i, W^{(k)})$) with all the clusters is deemed to be very weak (*i.e.*, smaller than the threshold value, ρ).

To solve this optimization problem, we propose a three steps partitioning algorithm:

- Step 1 (Initial partition of the blocks): K groups of blocks and a threshold ρ are given by the practitioner
- Step 2 (Determination of the group average scalar product matrix in the various clusters): In cluster G_k, $W^{(k)}$ and the associated weights α_i are determined by means of the STATIS method.
- Step 3 (changing clusters): New clusters of blocks are formed by moving each block of variables, X_i, to the cluster G_k for which the quantity $RV(W_i, W^{(k)})$ is the largest providing that this quantity is larger than ρ, otherwise X_i is assigned to the cluster "K+1".

2.3.2. Choice of the threshold value, ρ

We propose a procedure to select an appropriate threshold value, ρ. We consider as an atypical block of variables, a block which is an outlier in that sense that it is far removed from all the clusters, or a block which is straddling two or more clusters. These occurrences are depicted in Figure 2, where each block of variables is represented by a point. The three circles delineate three clear clusters and the points outside these circles represent atypical blocks of variables.

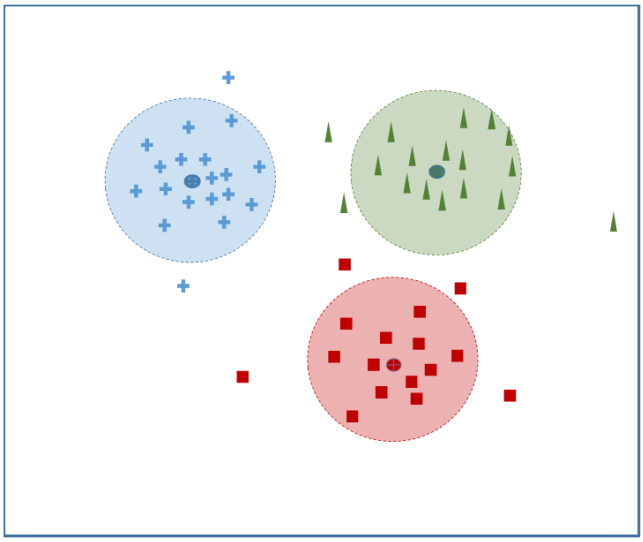

Figure 2.: Fictitious representation of the blocks of variables in three clusters. The blocks of variables out of the three circles are considered as atypical.

Since the parameter ρ represents a similarity index as measured by the RV coefficient between a block of variables and a cluster group average, it should be comprised between 0 and 1. For $\rho = 0$, the "K+1" cluster is empty, and for $\rho = 1$, all the blocks of variables are set in the "K+1" cluster. Dave (1991), who first introduced the concept of noise cluster for the cluster analysis of observations based on their distances, suggested a strategy for the choice of ρ. This consists, in a first step, in computing the average of the squared distances between each observation and each center of the various clusters. Then, an average of all these distances is computed and multiplied by yet another parameter (hyper-parameter). However, few indications are given regarding this hyper-parameter. Vigneau et al. (2016), who were concerned by the cluster analysis of variables and used the concept of "K+1" cluster, proposed for the selection of the threshold parameter to explore the range between 0 and 1 by considering several values of ρ, and assess the evolution of the within-cluster homogeneity as ρ varies. In parallel, the evolution of the number of variables assigned to the "K+1" cluster is investigated. From these perspectives, an appropriate ρ should correspond to a satisfactory within-

cluster homogeneity without, however, setting aside a large proportion of the variables in the "K+1" cluster.

We propose to automatically compute an appropriate value of the ρ parameter in CLUSTATIS. This value is defined as the average of the RV coefficients between each block of variables and the two nearest cluster group average configurations, which correspond to the two largest values of RV coefficients. Formally, the selected value ρ that we propose is the following:

$$\rho = \frac{\sum_{i=1}^{m} RV(W_i, W^{(k_i)}) + RV(W_i, W^{(k_{i'})})}{2m} \tag{4}$$

where $W^{(k_i)}$ is the group average scalar products matrix of the cluster to which the i^{th} block of variables belongs, and $W^{(k_{i'})}$ is the group average scalar products matrix of the nearest cluster to the i^{th} block of variables.

This index takes into account the proximity of each block of variables with its own cluster and the nearest cluster (the second nearest group average configuration). The larger the similarity of the blocks of variables to their group average, the larger is ρ. Furthermore, the closer the blocks of variables are to the neighbouring clusters the larger is ρ. The first property entails that for compact clusters with a high within-homogeneity, we should choose a relatively large threshold value, ρ. The second property entails that if the various clusters are relatively close to each others, then the threshold value should also be large so as to delineate clear boundaries between the clusters.

3. Case studies

3.1. Perception of luxury perfumes

The data considered herein can be found in the *R* package SensoMineR Lê and Husson (2008). A panel of 103 Dutch consumers were asked to smell 12 luxury perfumes and score their perception for each perfume with respect to 21 attributes. Examples of such attributes are "Intensity", "Freshness", "Jasmin", "Rose", *etc.* The perfumes "Shalimar" and "PurePoison" were replicated twice. All in all, we have 103 blocks of variables associated to the various subjects and the aim is to segment these subjects on the basis of how they perceive the perfumes. For this purpose, we run the CLUSTATIS method by using the *R* package ClustBlock (Llobell et al., 2020).

If we consider the whole panel as a unique cluster, the homogeneity index is equal to 40.1% which indicates a rather poor agreement among the consumers. The dendrogram or hierarchical tree and the graph which shows the evolution of the merging criterion in the course of the hierarchical clustering (Figure 3) indicate to consider four clusters of consumers since the merging criterion has jumped when passing from four to three clusters. The overall homogeneity index obtained as the weighted average of

the homogeneity indices in the four clusters is equal to 46.7%. By way of consolidating this partition, we run the partitioning algorithm considering the partition in four clusters derived from the hierarchical algorithm as a starting point. Only six consumers changed clusters and the homogeneity index is now equal to 47.1% indicating but a slight improvement over the solution obtained by means of the hierarchical algorithm. In a subsequent stage, we run the partitioning algorithm by activating the "K+1" option. This means that we added a noise cluster that should contain all the consumers who do not fit to the pattern of any cluster. This resulted in up to 36 consumers being discarded and the homogeneity index significantly increased (55.3%). The fact that 36 out of 103 consumers were set in the noise cluster may seem as relatively large. However, this is a common occurrence in this kind of studies which involve non-trained consumers who, in addition, do not get any pecuniary retribution. Some of these consumers perform the task casually or without a real motivation, others may not perceive a significant difference among the products or may confuse or misinterpret the attributes. Other studies dealing with consumers also reported that around one third of the panel may be deemed as atypical (Vigneau et al., 2016). The homogeneity results of our analysis are summed up in Table 1, where we also give the sample sizes of the various clusters. It can be seen that the cluster 2 was originally the cluster with the poorest homogeneity (39.0%). This explains why up to 23 consumers of this cluster moved to the noise cluster.

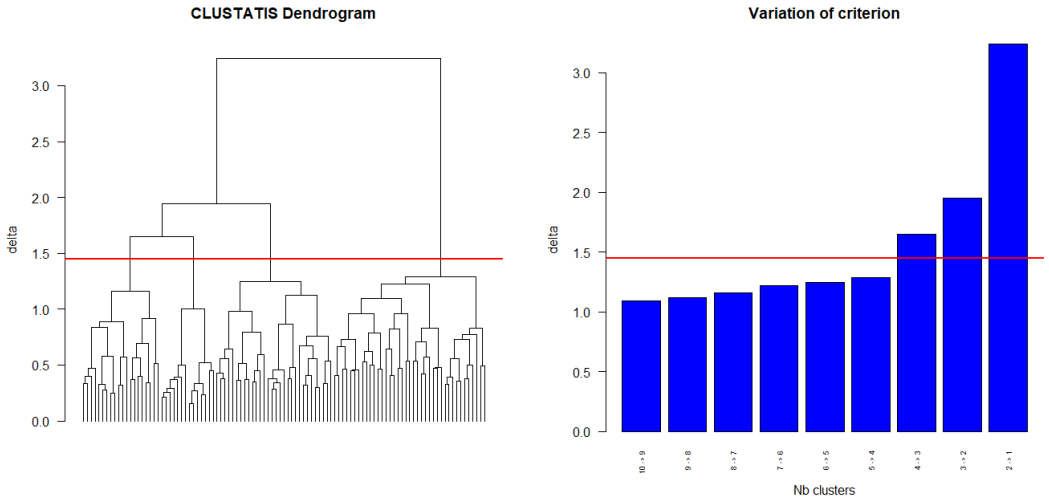

Figure 3.: The perfume data: Dendrogram and variation of criterion D given by the CLUSTATIS hierarchical algorithm.

CLUSTATIS yields, in addition to the partition of the blocks of variables into clusters, a group average scalar product matrix $W^{(k)}$ $(k = 1, \ldots, 4)$ within each cluster. By performing the spectral decomposition of $W^{(k)}$: $W^{(k)} = C^{(k)}C^{(k)^\top}$, we can depict the relationships among the perfumes on the basis of the columns of $C^{(k)}$. In Figure 4,

Table 1.: The perfume data: Size, homogeneity and global homogeneity indices using CLUSTATIS without and with the "K+1" strategy.

	CLUSTATIS		CLUSTATIS with "K+1"	
Cluster	Size	Homogeneity index (%)	Size	Homogeneity index (%)
1	21	49.3	16	55.1
2	38	39.0	15	50.7
3	18	59.3	14	64.4
4	26	48.7	22	52.6
Overall (weighted average)	-	47.1	-	55.3
The whole panel	84	40.1		

we show the representation of the perfumes on the basis of the first two components (*i.e.*, eigenvectors of $W^{(k)}$ associated with the largest eigenvalues) associated with the four clusters. We can see that these configurations are on the whole different from one another indicating a difference in the perception of the products in the four clusters. A further indication of the relatively poor performance in cluster 2 is that, unlike the other clusters, the two replicates corresponding to "Shalimar" are far removed from each other along the second component. In order to achieve a better characterization of the perfumes, we can investigate the correlations of the principal components in the various clusters with the original variables (data not given to save space).

3.2. Projective mapping data

In sensory evaluation, we have witnessed over the last decade or so a strong tendency to recourse to simple and quick procedures to evaluate a set of products. Very often, these procedures involve the final consumers of the products instead of trained assessors (Varela and Ares, 2014). Projective mapping (Risvik et al., 1994), also called Napping (Pagès, 2005), is one of these procedures that have gained ground. In this procedure, a set of products are presented to a panel of subjects and each subject is instructed to arrange the products on a sheet of paper, usually 60cm x 60cm, in such a way that similar products are arranged near one another, whereas different products are placed far apart. The data from each subject can be stored in a two dimensional block of variables where the variables refer to the coordinates on the sheet of paper. Note that although the blocks of variables associated to the various subjects are two-dimensional, the variables do not necessarily refer to the same sensory perceptions from one subject to another.

The data used in this case study relate to an experiment performed by 97 students

from Agrocampus-France. Each student was instructed to place the same 12 luxury perfumes as in case study 1 on a sheet of paper.

By way of segmenting the students who took part in the experiment, we run CLUSTATIS on the 97 two-dimensional blocks of variables. The dendrogram derived from the hierarchical algorithm is shown in Figure 5. It indicates that a partition into five clusters is appropriate.

We can see in Table 2 that the homogeneity of the whole panel is very poor (26.0%). It significantly increases after the partitioning of the panel into four clusters (39.9%). It very significantly increased to 53.5% when we added the noise cluster. Up to 34 blocks of variables moved to the noise cluster, with 15 blocks providing from cluster 3, which originally had a very poor homogeneity (23.7%). As in the previous case study, we can note that around one third of the panel of consumers were set in the noise cluster. This might seem at first sight too high a proportion but the reader should keep in mind that the panellists were non-trained students who were asked to perform this evaluation task as an assignment to practice the projective mapping procedure, that is without major issues at stake.

Table 2.: The projective mapping data: Size, homogeneity and global homogeneity indices using CLUSTATIS without and with the "K+1" strategy.

	CLUSTATIS		CLUSTATIS with "K+1"	
Cluster	Size	Homogeneity index (%)	Size	Homogeneity index (%)
1	27	49.1	23	54.4
2	13	47.5	11	51.9
3	23	23.7	8	48.7
4	19	36.7	11	51.1
5	15	45.9	10	59.5
Overall (weighted average)	-	39.9	-	53.5
The whole panel	97	26.0		

Table 3 gives the RV coefficients between the $W^{(k)}$ $(k = 1, \ldots, 5)$ that is the group average scalar product matrices associated with the five clusters. It highlights how far the clusters are distant from one another. It turns out that the closest clusters are clusters 1 and 2 (RV=0.64) and the two most distant clusters are clusters 4 and 5 (RV=0.22).

Table 3.: The projective mapping data: RV coefficients between the cluster group averages.

Cluster	1	2	3	4	5
1	1	0.64	0.56	0.46	0.61
2		1	0.53	0.42	0.58
3			1	0.51	0.46
4				1	0.22
5					1

4. Conclusion

The collection of several blocks of variables is becoming more and more frequent. This explains why the investigation of methods of analysis of this kind of data has been central in multivariate data analysis in the last thirty years or so (De Roover et al., 2012). However, the problem of clustering several blocks of variables has not been sufficiently addressed notwithstanding its interest in situations such as consumer segmentation, for example. CLUSTATIS fills this gap. It is a method of clustering blocks of variables that can be applied to blocks of variables that pertain to the same individuals but not necessarily the same variables. It encompasses a hierarchical cluster analysis and a partitioning algorithm. It is based on optimization criteria that, on the one hand, suggest clustering procedures and, on the other hand, provide performance indicators that make it possible to evaluate the quality of the obtained solutions. The introduction of a "K+1" class or "noise cluster" improves the quality of the solution obtained by means of CLUSTATIS by discarding those blocks of variables that do not fit to the pattern of any cluster. To do this, we have drawn from ideas developed by Dave (1991) and Vigneau et al. (2016). The proposed procedure requires the determination of a threshold parameter that governs the decision whether a block of variables should be assigned to a main cluster or to the "noise cluster". We proposed a procedure for automatically selecting a threshold taking into account the proximity between the clusters.

The *R* package ClustBlock (Llobell et al., 2020) allows to perform the CLUSTATIS method including its different options as well as functions for the cluster analysis of blocks of variables from specific sensory evaluation procedures.

Performing the hierarchical algorithm in a first step has two purposes (i) it helps choosing the number of clusters from the structure of the dendrogram or by examining the diagram showing the evolution of the aggregation criterion in the course of hierarchical clustering; (ii) it makes it possible to automatically compute the ρ parameter associated with the noise cluster. The partitioning algorithm can improve the solution obtained by

means of the hierarchical and, if the "K+1" option is activated, it makes it possible to set aside atypical blocks, resulting in yet a further improvement in terms of within clusters homogeneity. The counterpart of performing the two clustering algorithms in succession is that the computation time may be relatively large if the number of blocks of variables is large and if the number of rows (*i.e.*, individuals) is also large. In particular, the hierarchical algorithm is costly in terms of computation time. That is the reason why we advocate, in a setting with very many blocks of variables and individuals, performing only the partitioning algorithm. However, we recommend in this case using a multi-start strategy consisting in running the partitioning algorithm several times (say, 30 times) by choosing different staring points at random. As an indication, for the first case study which involved 103 blocks of variables, 21 variables and 14 products, the running time for both the hierarchical and partitioning algorithms was around 35 seconds.

The problem of determining the number of clusters is an important and tricky issue that should be more precisely investigated. For the time being, we have recommended determining this parameter by examining the evolution of the aggregation criterion associated with the hierarchical cluster analysis. Analytical procedures, possibly with hypothesis testing strategies, would certainly be more appropriate. Ongoing research on this indicates promising perspectives.

A. Proof of the STATIS properties

In this appendix, several properties are proved. One should bear in mind that the blocks of variables at hand were centred and normalized so as to have $||W_i|| = 1$. As a consequence, we have $RV(W_i, W_j) = trace(W_i W_j)$.

Let us consider the problem of minimization of the Q quantity with the constraint $\sum_{i=1}^{m} \alpha_i^2 = 1$. We have:
$Q = \sum_{i=1}^{m} ||W_i - \alpha_i W||^2 = \sum_{i=1}^{m} ||W_i||^2 - 2\sum_{i=1}^{m} \alpha_i < W_i, W > + \sum_{i=1}^{m} \alpha_i^2 ||W||^2 = m - 2\sum_{i=1}^{m} \alpha_i < W_i, W > + ||W||^2$.

The Lagrangian expression associated with the minimization problem is given by:
$L(\alpha_i, W, \mu) = m - 2\sum_{i=1}^{m} \alpha_i < W_i, W > + ||W||^2 + \mu(\sum_{i=1}^{m} \alpha_i^2 - 1)$ where μ is the Lagrange multiplier. Setting the partial derivative of L with respect to W to zero, it follows:
$\frac{\partial L}{\partial W} = 0 \Leftrightarrow -2\sum_{i=1}^{m} \alpha_i W_i + 2W = 0 \Leftrightarrow W = \sum_{i=1}^{m} \alpha_i W_i$.

Replacing in the expression of Q, $\sum_{i=1}^{m} \alpha_i W_i$ by W, we have:
$Q = m - 2\sum_{i=1}^{m} \alpha_i < W_i, W > + ||W||^2 = m - 2||W||^2 + ||W||^2 = m - ||W||^2$.

If instead, we replace W by $\sum_{i=1}^{m} \alpha_i W_i$, it follows:
$Q = m - \sum_{l=1}^{m} \sum_{i=1}^{m} \alpha_i \alpha_l < W_i, W_l >= m - \sum_{i=1}^{m} \alpha_i \alpha_l RV(W_i, W_l) = m - \alpha^\top R \alpha$ where $\alpha = (\alpha_1, \ldots, \alpha_m)^\top$ and R is the matrix which contains the pairwise RV coefficients. It

follows that the minimization of Q with respect to α leads to the maximization of the quantity $\alpha^\top R\alpha$. The solution to this problem is given by the eigenvector of R associated with the largest eigenvalue, that we denote by λ_1. This shows that we are led to the same solution as the STATIS method.

We have $||W||^2 = \alpha^\top R\alpha = \lambda_1\alpha^\top\alpha = \lambda_1$ (property (i)).
From the expression $R\alpha = \lambda_1\alpha$, it follows that for each i,
$\sum_{l=1}^m RV(W_i, W_l)\alpha_l = \lambda_1\alpha_i \Leftrightarrow \sum_{l=1}^m < W_i, W_l > \alpha_l = \lambda_1\alpha_i \Leftrightarrow < W_i, W >= \lambda_1\alpha_i \Leftrightarrow$
$\alpha_i = \frac{<W_i,W>}{\lambda_1} = \frac{RV(W_i,W)}{||W||} = \frac{RV(W_i,W)}{\sqrt{\lambda_1}}$ (property (ii)).

Since $||W||^2 = \lambda_1$, $Q = m - ||W||^2 = m - \lambda_1$ (property (iii)) and $m = Q + \lambda_1$ (property (iv)).

B. Proof of the CLUSTATIS properties

In this appendix, we show that for two clusters of blocks of variables A and B (say), $\lambda_1^{(A\cup B)} \leq \lambda_1^{(A)} + \lambda_1^{(B)}$:

As a preliminary remark, let us note that $RV(W_i, W_j) = v_i^\top v_j$ where $v_i = Vec(W_i)$ and $v_j = Vec(W_j)$. v_i (respectively, v_j) is obtained by stacking the columns of W_i (respectively, W_j) so as to form a unique vector.

Let us denote by $R_{A\cup B}$ the pairwise RV coefficients between the blocks of variables in $A \cup B$. We have $R_{A\cup B} = V_{A\cup B}^\top V_{A\cup B}$ where $V_{A\cup B}$ contains as columns the vectorised W_i associated with the blocks of variables in $A \cup B$. In a similar way, we denote by V_A (respectively, V_B) the vectorised W_i associated with blocks of variables in A (respectively, B). Since the largest eigenvalue of $V_{A\cup B}^\top V_{A\cup B}$ is equal to that of $V_{A\cup B}V_{A\cup B}^\top$, it follows:
$\lambda_1^{(A\cup B)} = \max_{\alpha,||\alpha||=1}\{\alpha^\top V_{A\cup B}V_{A\cup B}^\top\alpha\} = \max_{\alpha,||\alpha||=1}\{\alpha^\top V_A V_A^\top\alpha + \alpha^\top V_B V_B^\top\alpha\}$
$\leq \max_{\alpha,||\alpha||=1}\{\alpha^\top V_A V_A^\top\alpha\} + \max_{\alpha,||\alpha||=1}\{\alpha^\top V_B V_B^\top\alpha\} = \lambda_1^{(A)} + \lambda_1^{(B)}$.

References

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.

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.

Dave, R. N. (1991). Characterization and detection of noise in clustering. *Pattern Recognition Letters*, 12(11):657–664.

De Roover, K., Ceulemans, E., and Timmerman, M. E. (2012). How to perform multiblock component analysis in practice. *Behavior Research Methods*, 44(1):41–56.

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. –john wiley & sons. *Ltd., New York*, page 330.

Glaçon, F. (1981). *Analyse conjointe de plusieurs matrices de données: Comparaison de différentes méthodes*. PhD thesis, Université scientifique et médicale de Grenoble.

Jack, F. R. and Piggott, J. (1991). Free choice profiling in consumer research. *Food quality and preference*, 3(3):129–134.

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

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

Llobell, F., Vigneau, E., Cariou, V., and Qannari, E. M. (2020). *ClustBlock: Clustering of Datasets*. R package version 2.2.0.

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.

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.

Robert, P. and Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods: the RV-coefficient. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 25(3):257–265.

Saporta, G. (2006). *Probabilités, analyse des données et statistique*. Editions Technip.

Schlich, P. (1996). Defining and validating assessor compromises about product distances and attribute correlations. In *Data handling in science and technology*, volume 16, pages 259–306. Elsevier.

Varela, P. and Ares, G. (2014). *Novel techniques in sensory characterization and consumer profiling*. CRC Press.

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

Vigneau, E., Qannari, E., Navez, B., and Cottet, V. (2016). Segmentation of consumers in preference studies while setting aside atypical or irrelevant consumers. *Food quality and preference*, 47:54–63.

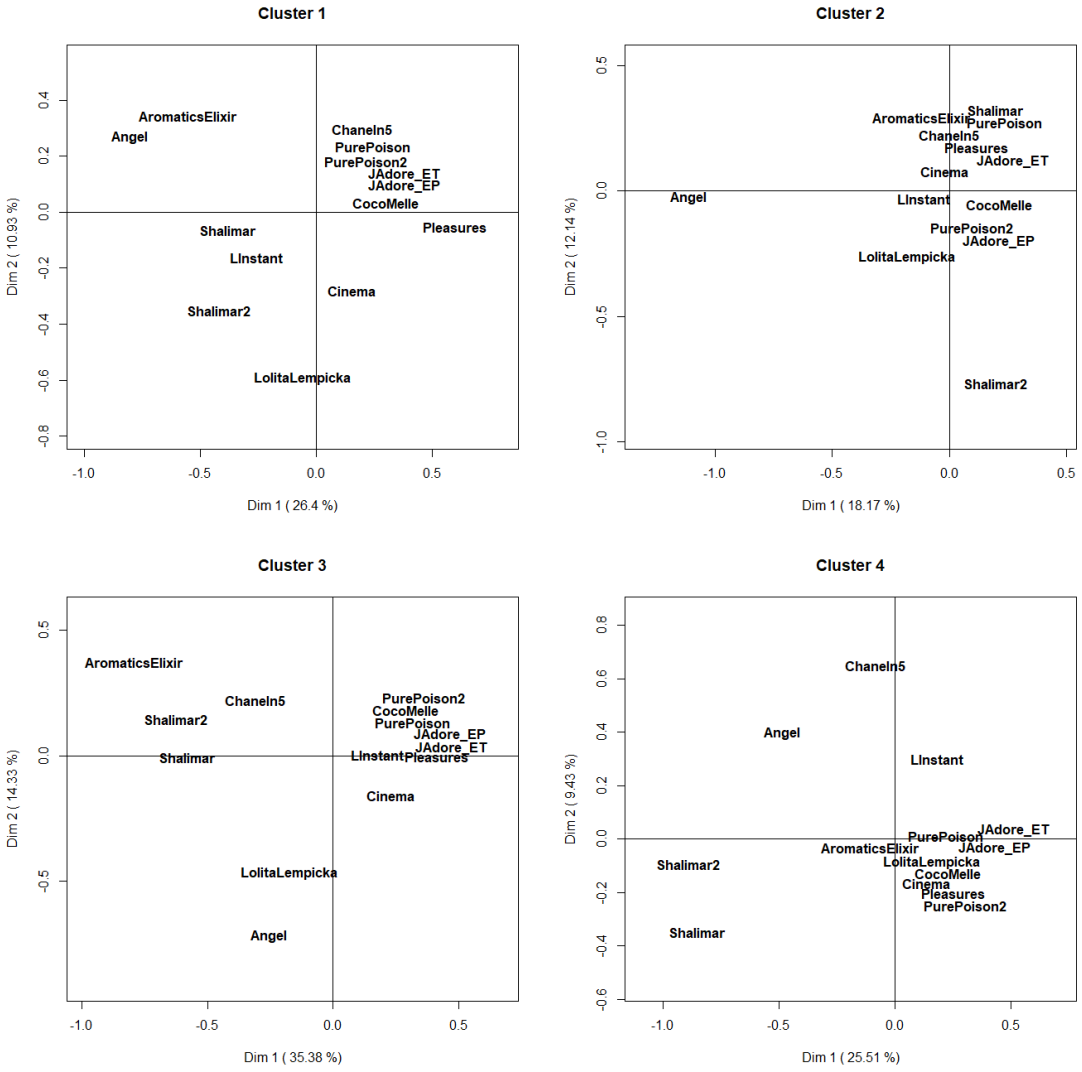

Figure 4.: The perfume data: Graphical representation of the perfumes on the basis of the first two principal components in each cluster.

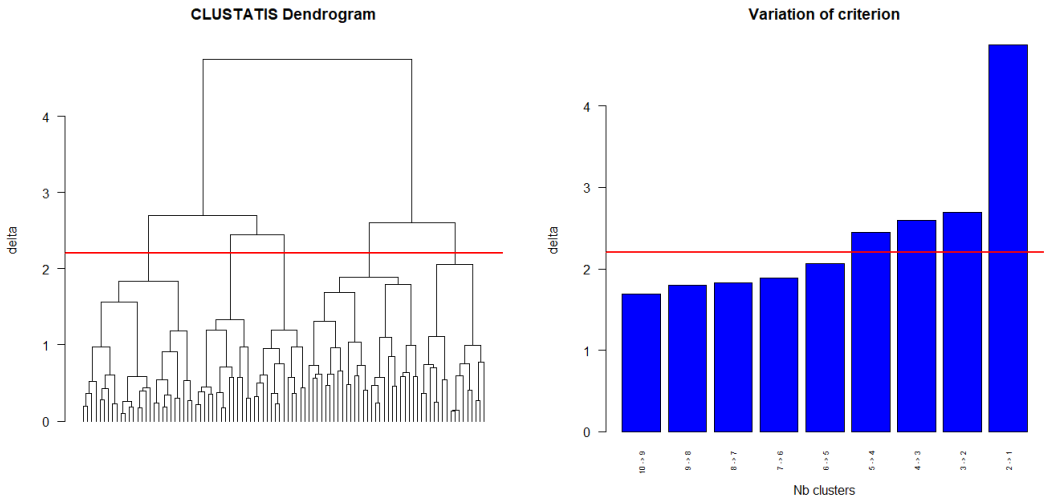

Figure 5.: The projective mapping data: Dendrogram and variation of criterion D given by the CLUSTATIS hierarchical algorithm.