Food Quality and Preference 77 (2019) 184–190

Contents lists available at ScienceDirect

Food Quality and Preference

journal homepage: www.elsevier.com/locate/foodqual

Assessment of the agreement and cluster analysis of the respondents in a CATA experiment

Fabien Llobell[a,b,*], Davide Giacalone[c], Amaury Labenne[b], El Mostafa Qannari[a]

[a] StatSC, ONIRIS, INRA, Nantes, France
[b] Addinsoft, XLSTAT, Paris, France
[c] Universiy of Southern Denmark, Odense, Denmark

ARTICLE INFO

Keywords:
Agreement
Segmentation
Cluster analysis
Atypical subjects
CATA experiment
Permutation test

ABSTRACT

Statistical tools to assess the agreement of respondents in a Check-All-That-Apply (CATA) experiment are discussed. An overall index of agreement is introduced and a hypothesis test to assess the significance of this index is outlined. A similar investigation at the level of each attribute is undertaken. The permutation test proposed in this latter situation can be compared to Cochran's *Q* test. We also propose to cluster the respondents while setting aside those respondents that are deemed to be atypical. This strategy of clustering stands a refinement of the cluster analysis called CLUSCATA that was introduced in a previous paper. The various strategies of analysis are illustrated by means of real case studies.

1. Introduction

Check All That Apply (CATA) experiment is among those so-called rapid methods of sensory evaluation that have gained ground these last two decades or so (Meyners, Castura, & Carr, 2013; Varela & Ares, 2014). In this experiment, respondents are instructed to associate attributes from a predefined list to each one of the products to be evaluated. For more details regarding this experiment, we refer to the paper by Meyners et al. (2013).

Besides setting up graphical displays that depict the similarity and the dissimilarity of the products, practitioners have been interested in assessing the agreement among the respondents. The importance of assessing the agreement of respondents in a CATA experiment was stressed by Worch and Piqueras-Fiszman (2015). For that purpose, these authors proposed to use strategies based on multiple factor analysis on contingency tables and McNemar tests. It is worth noting that Meyners, Castura, and Worch (2016) remarked that although McNemar's test provides insight into products differences, it is not appropriate to assess the repeatability in a CATA experiment. The discrimination of the products by the various attributes (*e.g.* Cochran's *Q* test, Meyners et al., 2013) are also important topics.

We propose a unified framework to address some of these issues. In this regard, the data from each respondent are plunged in a Euclidean space. The cosine between two matrices associated with two respondents stands as a similarity index between these two respondents. An overall agreement index is derived from the first eigenvalue of the matrix that contains the pairwise similarity indices between the respondents. The associated eigenvector yields individual agreement indices that reflect how each respondent agrees with the general point of view of the panel. It is clear that this analysis bears high similarities to the STATIS method where the RV coefficient is used as a similarity index between the datasets (Lavit, Escoufier, Sabatier, & Traissac, 1994). A permutation test is performed in order to assess the significance of the overall agreement index. Following a similar strategy, we investigate the agreement of the respondents for each attribute taken separately. If the agreement index associated with a given attribute turns out to be non-significant, this means that this attribute has been misunderstood by the subjects or interpreted differently by them. Therefore, it is very likely that this inconsistency will result in a non-discrimination of the products. Drawing on selected case studies, we will compare the outcomes of the permutation tests regarding the consistency of the attributes with those of Cochran's *Q* test, which is commonly used to assess the discrimination of the attributes (Meyners et al., 2013).

In a subsequent section, we tackle the issue of clustering the subjects participating to a CATA experiment. We discuss a refinement of a clustering strategy called CLUSCATA (Llobell, Cariou, Vigneau, Labenne, & Qannari, 2019). The idea is to use an additional cluster, called "noise cluster" or "K+1" cluster, to the standard CLUSCATA procedure, with the aim of catching those respondents who are atypical

* Corresponding author.
E-mail address: fllobell@xlstat.com (F. Llobell).

https://doi.org/10.1016/j.foodqual.2019.05.017
Received 21 March 2019; Received in revised form 20 May 2019; Accepted 20 May 2019
Available online 22 May 2019
0950-3293/

in that sense that they do not fit to the pattern of any main cluster. The concept of noise cluster was originally introduced by Dave (1991), and has previously been adapted to various situations in sensometrics (Berget, Varela, & Næs, 2019; Llobell, Vigneau, & Qannari, 2019; Vigneau, Qannari, Navez, & Cottet, 2016).

The paper is organized as follows. In the section "Theory", we start by outlining a brief reminder of the similarity among pairs of respondents. Then, we discuss how the overall agreement of the respondents or, equivalently, the homogeneity of the panel can be assessed and how a permutation test can be set up to evaluate the significance of this overall index. We also tackle the issue of assessing the agreement of the respondents for each attribute. Subsequently, we discuss how the strategy of clustering the respondents called CLUSCATA can be adapted to include a noise cluster. In Section 3, the general strategy of analysis is illustrated of the basis of selected case studies. We end the paper by a conclusion (Section 4).

2. Theory

2.1. Similarity measure between respondents

Consider a CATA experiment involving m subjects who are instructed to evaluate n products on the basis of p attributes. The data from each respondent can be coded as a matrix $(n \times p)$ whose rows refer to the products and columns to the attributes. The entry of this matrix is equal to 1 if the respondent has ticked the attribute in the corresponding column for the product in the corresponding row. Otherwise, the entry is equal to 0 (Meyners et al., 2013). Thus, the data of each respondent are plunged in a Euclidean space, namely the space of matrices of dimensions $n \times p$. We can show that the cosine between two matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ associated with two respondents is given by $s(\boldsymbol{X}, \boldsymbol{Y}) = \frac{<\boldsymbol{X}, \boldsymbol{Y}>}{||\boldsymbol{X}|| \, ||\boldsymbol{Y}||} = \frac{n_{XY}}{\sqrt{n_X}\sqrt{n_Y}}$, where n_{XY} is the total number of times that both respondents checked the same attributes for the same products, and n_X (resp. n_Y) is the total number of checks across attributes and products of respondent $\boldsymbol{X}$ (resp. $\boldsymbol{Y}$). Indeed, since $\boldsymbol{X}$ and $\boldsymbol{Y}$ contains only 0 and 1, it is easy to check that $<\boldsymbol{X}, \boldsymbol{Y}> = \sum_i \sum_j \boldsymbol{X}_{ij} \boldsymbol{Y}_{ij} = n_{XY}$, $||\boldsymbol{X}|| = \sqrt{\sum_i \sum_j \boldsymbol{X}_{ij}^2} = n_X$ and similarly $||\boldsymbol{Y}|| = n_Y$. This index $s(\boldsymbol{X}, \boldsymbol{Y})$ stands as a similarity coefficient between $\boldsymbol{X}$ and $\boldsymbol{Y}$ and is referred to the Ochiai coefficient (Ochiai, 1957). We can show that $s(\boldsymbol{X}, \boldsymbol{Y}) = 0$ if and only if the two respondents at hand have checked completely different attributes for each product. We can also show that $s(\boldsymbol{X}, \boldsymbol{Y}) = 1$ if and only if $\boldsymbol{X} = \boldsymbol{Y}$, which means that the two respondents perfectly agree. More details regarding this similarity index can be found in (Llobell, Cariou, et al., 2019).

It is worth mentioning that, within the context of CATA experiments, Carr, Dzuroska, Taylor, Lanza, and Pansini (2009) and Meyners et al. (2013) proposed an approach of analysis called Multidimensional Alignment (MDA) based on the cosine between pairs of vectors. The aim of MDA is to investigate the relationships between products and attributes. The aim of the approach adopted herein is different since we seek to assess the similarity between the respondents participating to the CATA experiment.

2.2. Overall and individual agreement indices

Let us denote by $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_m$ the 0–1 matrices associated with the m respondents. We also denote by $s_{ik} = s(\boldsymbol{X}_i, \boldsymbol{X}_k)$ the Ochiai coefficient between $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$. Thus, the matrix $\boldsymbol{S} = (s_{ik})_{i,k=1,\ldots,m}$ of dimension $m \times m$ is formed by positive entries. As a consequence, its largest eigenvalue, which we denote by λ, is positive and its corresponding eigenvector $\alpha = (\alpha_1, \cdots, \alpha_m)$ can be chosen in such a way that all its components are positive (Meyer, 2000). As an overall index of agreement among the respondents or homogeneity of the panel, we propose $I = \frac{\lambda}{m}$. We can show that this index is comprised between $\frac{1}{m}$ and 1. The lower bound is achieved if all the respondents are in complete disagreement (*i.e.*, $s(\boldsymbol{X}_i, \boldsymbol{X}_k) = 0$ for $i \neq k$). Contrariwise, $I = 1$ if $s(\boldsymbol{X}_i, \boldsymbol{X}_k) = 1$ for $i, k = 1, \cdots, m$ (*i.e.*, complete agreement). For more details, we refer to the paper by Llobell, Cariou, et al. (2019)

Assessing the significance of this index means that we are putting in balance the two hypotheses H_0: $I = \frac{1}{m}$ against H_1: $I > \frac{1}{m}$. We propose to perform a permutation test consisting in randomly permuting the rows of each dataset $\boldsymbol{X}_i$ leading to the dataset $\boldsymbol{X}_i^*$. Thereafter, the agreement of the datasets $\boldsymbol{X}_i^* (i = 1, \cdots, m)$ thus obtained is evaluated by means of $I^* = \frac{\lambda^*}{m}$, where λ^* is the largest eigenvalue of the matrix $\boldsymbol{S}^* = (s(\boldsymbol{X}_i^*, \boldsymbol{X}_k^*))_{i,k=1,\ldots,m}$. This permutation procedure is repeated a large number of times (say, 1000 times) and the distribution of the values I^* thus obtained will stand for the distribution of the test statistic under H_0. This means that we can compute a p-value as the proportion of simulated values I^* that are larger or equal than the observed value, I. As a consequence, the hypothesis H_0 will be rejected if this p-value is smaller than a significance level chosen by the practitioner (*e.g.*, 5%).

The agreement of respondent $i (i = 1, \cdots, m)$ can be assessed by α_i which is, as stated above, the i^{th} component of the eigenvector of matrix S associated with the largest eigenvalue, λ. The coefficients α_i are comprised between 0 and 1 but unlike the overall agreement index, these coefficients should not be compared to 1 but to each other. This means that they enable us to rank the respondents according to their agreement with the general point of view of the panel. Obviously, their interpretation depends on the value of the overall index of agreement discussed above.

2.3. Attributes consistency

We consider the degree of agreement of the respondents regarding the evaluation of the products with respect to each attribute. More specifically, we are interested in the hypothesis test that evaluates whether there is no agreement at all among the respondents when evaluating the products for the attribute under consideration against the alternative hypothesis that stipulates that there is some degree of consistency among the respondents. In case the null hypothesis is not rejected, the attribute in consideration is very unlikely to discriminate the products. However, the opposite implication is not necessarily true since the respondents may perfectly agree on the assessment of an attribute that is not intrinsically discriminant. To assess the discrimination of the attributes, it is recommended to perform Cochran's Q test (Cochran, 1950; Meyners et al., 2013), which is devoted to investigate the differences among several treatments (*i.e.*, products in the case of CATA) when the data at hand are binary. The consistency test proposed herein may complement the Cochran's Q test in the sense that it checks whether the non-discrimination of the products by the attribute under consideration can be explained by the disagreement among the respondents.

The hypothesis test for the respondents' consistency in evaluating the products by means of a given attribute follows exactly the same process as for the evaluation of the overall agreement (see previous section). Since we are interested in one attribute at a time, the data can be stored in a table where the rows refer to the products and the columns refer to the respondents. The entries of this table are 1 or 0 indicating whether each respondent (column) has checked the attribute under consideration for each product (row) or not. We can note in passing that this is precisely the setting for performing the Cochran's Q test.

We will not give further details regarding the consistency test since it follows point by point the test regarding overall agreement. It proceeds as if we had a CATA experiment with one attribute, that is the attribute for which we investigate the consistency.

2.4. CLUSCATA while setting aside atypical respondents

In a previous paper (Llobell, Cariou, et al., 2019), a cluster analysis

specially tailored for the CATA data – called CLUSCATA – was introduced. The rationale behind this method of analysis is to determine homogeneous clusters of respondents and, for each cluster, a group average dataset which is as close as possible to the datasets associated with the respondents in this cluster.

From now on, the datasets $\boldsymbol{X_i}(i = 1, \cdots, m)$ associated with the various respondents are assumed to be scaled in order to have their norm equal to one. This is achieved by dividing the original binary datasets by their respective norms. This standardization makes it possible to account for the total number of checked attributes which may significantly differ from one respondent to another. For what follows, we also need to extend the similarity index s to measure the similarity between two datasets which are not necessarily binary. This is given for two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ (say) by $s(\boldsymbol{A}, \boldsymbol{B}) = <\boldsymbol{A}, \boldsymbol{B}> = \frac{trace(\boldsymbol{AB}^T)}{\sqrt{trace(\boldsymbol{AA}^T)trace(\boldsymbol{BB}^T)}}$. Obviously, this is the cosine between $\boldsymbol{A}$ and $\boldsymbol{B}$ considered as elements of the space of matrices of dimension $n \times p$. For more details, we refer to the paper by Llobell, Cariou, et al. (2019). In that paper, we proposed an optimisation criterion that clearly highlights the rationale behind CLUSCATA. A new formulation of CLUSCATA which is more conducive to the 'noise cluster' idea is to seek to maximize the quantity:

$$H = \sum_{k=1}^{K} \sum_{i \in G_k} s^2(\boldsymbol{X_i}, \boldsymbol{C^{(k)}}) \tag{1}$$

where $\boldsymbol{C^{(k)}} = \sum_{i \in G_k} \alpha_i \boldsymbol{X_i}$, with $\sum_{i \in G_k} \alpha_i^2 = 1$. $\boldsymbol{C^{(k)}}$ is the group average dataset in cluster G_k, and the weight α_i in cluster G_k reflects the degree of agreement of respondent i with $\boldsymbol{C^{(k)}}$. Clearly, this criterion reflects the idea that we are seeking clusters of respondents so that in each cluster, G_k, the datasets $\boldsymbol{X_i}$ are as much close (*i.e.*, high similarity) as possible to the group average dataset, $\boldsymbol{C^{(k)}}$ associated with G_k.

In order to solve this maximization problem, a hierarchical algorithm followed by a partitioning algorithm was proposed (Llobell, Cariou, et al., 2019). We can also show that $H = \sum_{k=1}^{K} \lambda^{(k)}$ where $\lambda^{(k)}$ is the largest eigenvalue of matrix $\boldsymbol{S}$, which contains the pairwise similarities as measured by s between the respondents in group k. As previously, we propose to assess the homogeneity in cluster k by $I_k = \frac{\lambda^{(k)}}{m^{(k)}}$ where $m^{(k)}$ is the number of respondents in cluster k.

As a matter of fact, we can show the following equality: $\sum_{i=1}^{m} ||\boldsymbol{X_i}||^2 = \sum_{k=1}^{K} ||\boldsymbol{C^{(k)}}||^2 + \sum_{k=1}^{K} \sum_{i \in G_k} ||\boldsymbol{X_i} - \alpha_i \boldsymbol{C^{(k)}}||^2$. Since $||\boldsymbol{X_i}|| = 1$ (standardization), it follows that $\sum_{i=1}^{m} ||\boldsymbol{X_i}||^2 = m$. This quantity stands as the total variation in the original datasets. The quantity $\sum_{k=1}^{K} \sum_{i \in G_k} ||\boldsymbol{X_i} - \alpha_i \boldsymbol{C^{(k)}}||^2$ stands as the within clusters variation and $\sum_{k=1}^{K} ||\boldsymbol{C^{(k)}}||^2$ stands as the between clusters variation. It follows that the ratio $\sum_{k=1}^{K} ||\boldsymbol{C^{(k)}}||^2 / \sum_{i=1}^{m} ||\boldsymbol{X_i}||^2$, which is equal to $\frac{\sum_{k=1}^{K} \lambda^{(k)}}{m}$, (*i.e.*, the overall homogeneity) also reflects how the $\mathbf{C^{(k)}}$ are far removed from each others. The lower bound of this quantity, which is equal to K/m, means there is no grounds for clustering (*i.e.*, between clusters variation negligible) and the upper bound, which is equal to 1, entails that we have perfectly separate clusters and within each cluster the datasets are similar (*i.e.*, within clusters variation is equal to 0).

We will focus hereinafter on the partitioning algorithm, which is akin to the k-means algorithm since the datasets are iteratively moved in and out of the clusters depending on their similarity with the average datasets of these clusters. In parallel, the average datasets are updated after each change (*i.e.*, removal and inclusion of datasets in the clusters). The partitioning algorithm requires to fix the number of cluster and an initial partition of the respondents as a starting point. This can be achieved by examining the outcomes from the hierarchical strategy of clustering (Everitt, Landau, Leese, & Stahl, 2011; Llobell, Cariou, et al., 2019). More precisely, the evolution of the aggregation criterion in the course of the clustering process is reflected by the lengths of branches of the dendrogram or hierarchical tree. Therefore, we can choose the number of clusters that corresponds to a significant jump of the aggregation criterion when passing from a partition with K clusters to a partition with K-1 clusters. The solution thus obtained is further improved by running the partitioning algorithm associated with CLUSCATA (Llobell, Cariou, et al., 2019).

We will discuss how the partitioning algorithm can be extended in order to identify and set aside the respondents who are atypical in that sense that they do not fit to the pattern of any cluster. Following Dave's idea (Dave, 1991), we introduce, besides the main K clusters, an additional cluster called "noise cluster" or "K+1 cluster", which will contain the respondents who are deemed to be atypical. This strategy of analysis requires selecting a threshold value, ρ, between 0 and 1, below which the similarity between a dataset associated with a respondent and the average dataset associated with a given cluster is considered as weak. Naturally, the partitioning algorithm seeks to assign a respondent to the cluster for which this similarity is the largest but, if this similarity is smaller than the threshold value ρ, the respondent is assigned to the "K+1 cluster". Formally, we seek to maximize the following criterion:

$$H_\rho = \sum_{k=1}^{K} \sum_{i=1}^{m} (\delta_{ik} s^2(\boldsymbol{X_i}, \boldsymbol{C^{(k)}}) + \delta_{i(K+1)} \rho^2) \tag{2}$$

where δ_{ik} is the Kronecker symbol which is equal to 1 if the dataset $\boldsymbol{X_i}$ belongs to cluster k and 0 otherwise. We have the constraint $\sum_{k=1}^{K+1} \delta_{ik} = 1$, which stipulates that a subject belongs to one and only one cluster, including the noise cluster. To solve this maximization criterion, we run the following algorithm:

Step 0 (initial partition). Choose a partition into K clusters of datasets (*i.e.,* subjects) as a starting point. This can be done by a random assignment of the datasets at hand to K clusters or by cutting the hierarchical tree at the level corresponding to K clusters.

Step 1 (group average datasets). In each cluster G_k, the group average $\boldsymbol{C^{(k)}}$ is determined as a weighted average of the datasets in this cluster, where the weights are derived from the first eigenvector of the matrix which contains the pairwise similarities between the respondents in cluster G_k.

Step 2 (changing clusters). New clusters of subjects are formed by moving each dataset, $\boldsymbol{X_i}$, to the cluster G_k for which the quantity $s(\boldsymbol{X_i}, \boldsymbol{C^{(k)}})$ is the largest providing that this quantity is larger than ρ otherwise, $\boldsymbol{X_i}$ is assigned to the cluster "K+1".

Steps 1 and 2 are iterated until the datasets stop changing clusters.

As a matter of fact, CLUSCATA is an adaptation of a clustering approach called CLUSTATIS to the particular case of CATA data (Llobell, Cariou, et al., 2019). Within the framework of CLUSTATIS, a strategy to select an appropriate threshold value was proposed and backed up by several considerations (Llobell, Vigneau, & Qannari, 2019). The same considerations could be applied to CLUSCATA. This leads us to select the following quantity:

$$\rho = \frac{\sum_{i=1}^{m} (s(\boldsymbol{X_i}, \boldsymbol{C^{(k_i)}}) + s(\boldsymbol{X_i}, \boldsymbol{C^{(k_i')}})}{2m} \tag{3}$$

where $\boldsymbol{C^{(k_i)}}$ is the weighted average dataset of the cluster to which the i^{th} dataset belongs, and $\boldsymbol{C^{(k_i')}}$ is the weighted average dataset of the second nearest cluster to the i^{th} dataset, that is the cluster for which $s(\boldsymbol{X_i}, \boldsymbol{C^{(k_i')}})$ is the largest after $s(\boldsymbol{X_i}, \boldsymbol{C^{(k_i)}})$ The rationale behind the selection of this parameter is to set up an intermediary value that delineates the boundary between the clusters, so much so that those subjects who are straddling two or more clusters are assigned to the noise cluster.

3. Illustrations

3.1. Beers dataset

The data used to illustrate the general strategy of analysis discussed herein pertain to a CATA experiment where 9 beer images were

Table 1
CATA attributes and product names (beer brands) in the beer dataset.

Attribute	Abbreviation	Attribute	Abbreviation
As a gift for someone	Gift	At a rugby match	Rugby
As a treat for myself	Treat	At work for Friday drinks	Work
At a BBQ with friends	BBQ	On a camping or fishing trip	Camping/fishing
As a thirst-quencher	ThirstQ	To celebrate an achievement	Achievement
At a fine-dining restaurant	Fine dining	To serve to guests	Guests
At a music concert	Concert	Watching TV at home	TV
At a party	Party	With a Snack	Snack
At a public house (bars, *etc.*)	PubHouse		

Beer brands and abbreviations between brackets: Steilager Classic (SC), Gold Medal Ale (GMA), Lion Red (LR), Mac's Hop Rocker (MHR), Monteith's Black Beer (MBB), Stoke Gold (STG), Hopwired Ipa (HPW), Stonecutter Scotch Ale (SSA) and Pot Kettle Black (PKB).

evaluated by 76 consumers using 15 attributes. The consumers were instructed to check the relevant usage contexts to drink each beer. The attributes (or contexts) are shown in Table 1. The beer images and more details regarding these data are given in Giacalone et al. (2015).

3.2. Respondents' agreement

As indicated above, the data from respondent $i (i = 1, \cdots, 76)$ were converted into a binary matrix $\boldsymbol{X_i}$. These matrices are scaled by dividing them by their respective norm.

From the matrix, $\boldsymbol{S}$, containing the pairwise similarity coefficients, we computed the largest eigenvalue, which, once divided by $m = 76$ yields a homogeneity index $I = 46.8\%$. This highlights a rather fair agreement among the respondents. By performing the permutation test, this value turned out to be significant (p-value <0.001). Fig. 1 shows the agreement coefficient, α_i, associated to the individual respondents. As indicated above, these weights are the components of the eigenvector of matrix $\boldsymbol{S}$ associated with the largest eigenvalue. We can see that some subjects (*e.g.*, subjects 46 and 50) have very small weights indicating a very poor agreement with the panel.

Following the strategy of analysis called CATATIS (Llobell, Cariou, et al., 2019), the coefficients α_i are used to compute a weighted average $\boldsymbol{C} = \sum_{i=1}^{m} \alpha_i \boldsymbol{X_i}$, which stands as a compromise configuration. Thereafter, $\boldsymbol{C}$ is submitted to Correspondence Analysis (CA; Greenacre, 2007). Fig. 2 gives the biplot depicting the beers and the attributes. The first axis, which explains up to 85% of the total variation, shows an opposition of contexts that call for a celebration (*e.g.*, fine dining, achievement, treat) to casual contexts (*e.g.*, fishing, rugby). Accordingly, the beers are positioned along the first axis depending on their association with the contexts reflected by the first axis. All in all, the findings from this analysis agree to a large extent to those obtained by Giacalone et al. (2015) and we refer to this paper for a more comprehensive discussion of these findings.

3.3. Attributes consistency

Each attribute was considered in turn and submitted to a permutation test to assess whether the respondents were (in)consistent in their evaluation. All the permutation tests were significant, implying that for none of the attributes, the panel of respondents were completely inconsistent. In parallel, we performed Cochran's Q test on the data associated to each attribute to assess whether these attributes were discriminant. All the attributes turned out to be discriminant.

3.4. CLUSCATA

The hierarchical algorithm of CLUSCATA suggests to consider four clusters (Fig. 3). Indeed, we can observe a significant jump *delta* of the

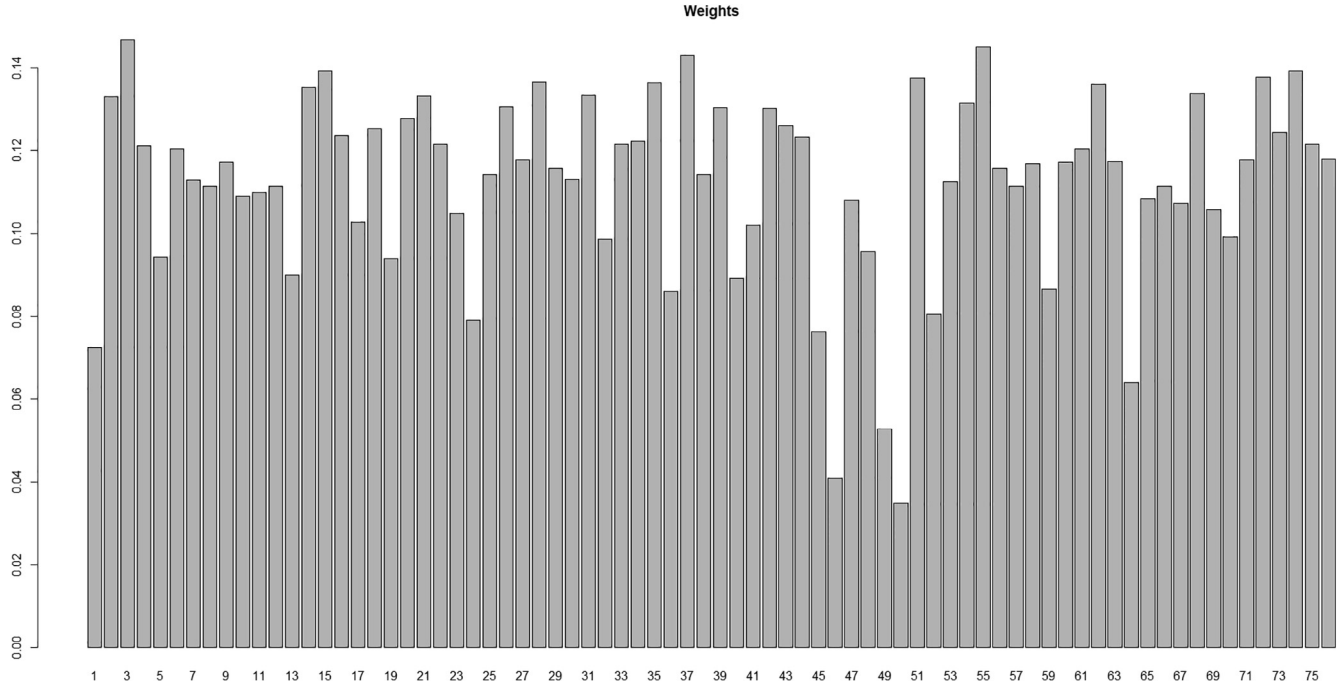

Fig. 1. Weight associated with each respondent.

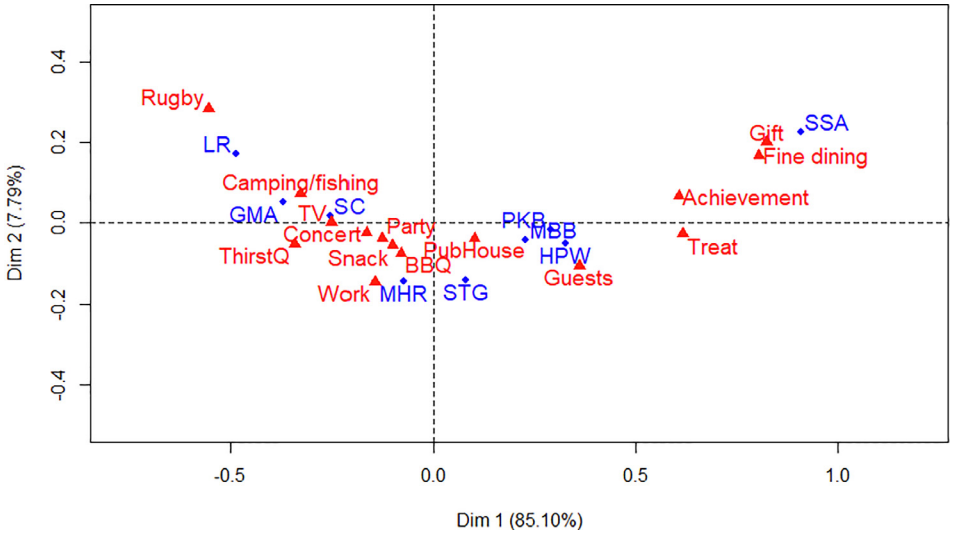

Fig. 2. CA on the weighted average of the beer data.

criterion H (Eq. (1)) when passing from a partition in four clusters to a partition with three clusters. This suggests that, at this step, heterogeneous clusters are being merged.

By way of comparing methods, we performed, in a first stage, the CLUSCATA partitioning algorithm without the option commending the detection of atypical subjects. Then, in a second stage, we run CLUSCATA including the "K+1" strategy. We computed the threshold parameter ρ from equation (4) and found 0.68. This resulted in 18 consumers being set in the noise cluster (Table 2) with an almost even repartition in the four clusters. By removing these respondents, the overall homogeneity increased from 54.0% to 61.7%.

Fig. 4 shows the graphical representation of the compromise of cluster 1 without identification of atypical subjects (left) and after the removal of atypical subjects (right). We can see that the removal of the atypical subjects resulted in some changes in the positions of the beers and the attributes. For instance, the beer LR appears as extreme along the second axis in the configuration on the right figure *(i.e.*, without the atypical respondents). The association of this beer with the attribute

Table 2
Size, homogeneity and overall homogeneity indices using CLUSCATA without and with the "K+1" strategy.

	CLUSCATA		CLUSCATA with "K+1"	
Clusters	Size	Homogeneity index	Size	Homogeneity index
1	15	48.9%	10	62.1%
2	28	60.3%	24	63.4%
3	21	48.3%	16	55.6%
4	12	55.6%	7	69.3%
Overall (weighted average)	76	54.0%	58	61.7%
The whole panel	76	46.8%		

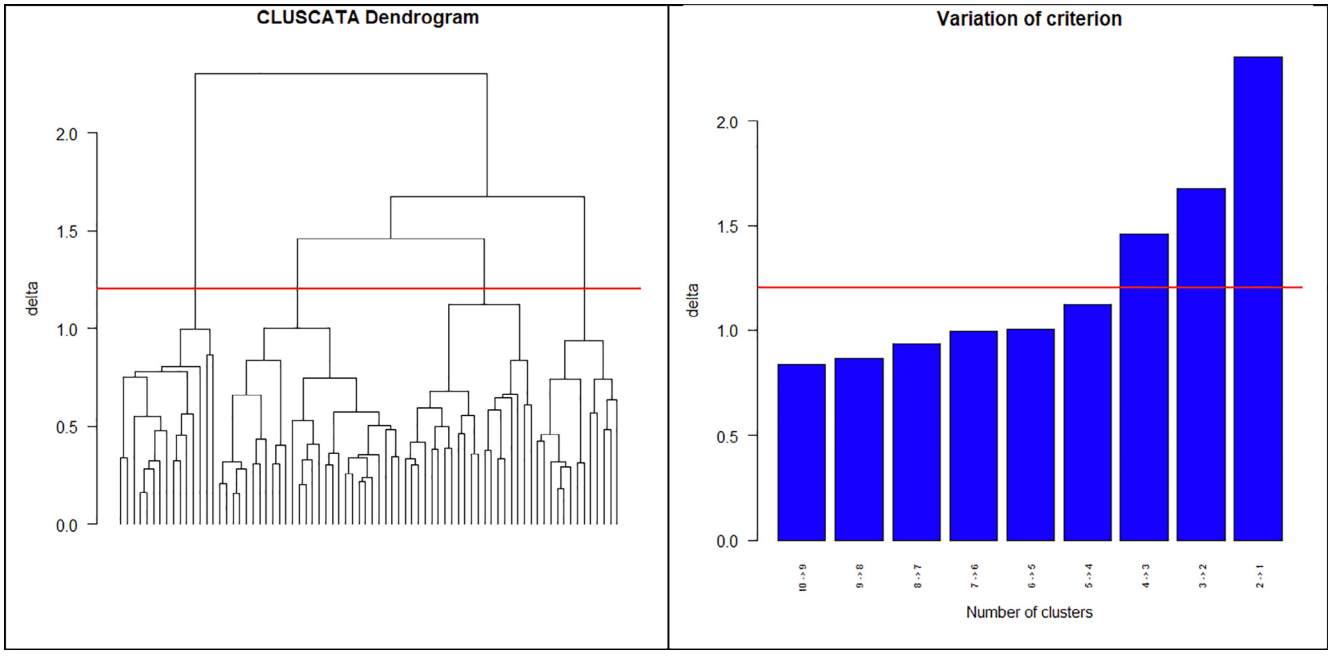

Fig. 3. Dendrogram given by the CLUSCATA hierarchical clustering analysis and variation of the criterion H.

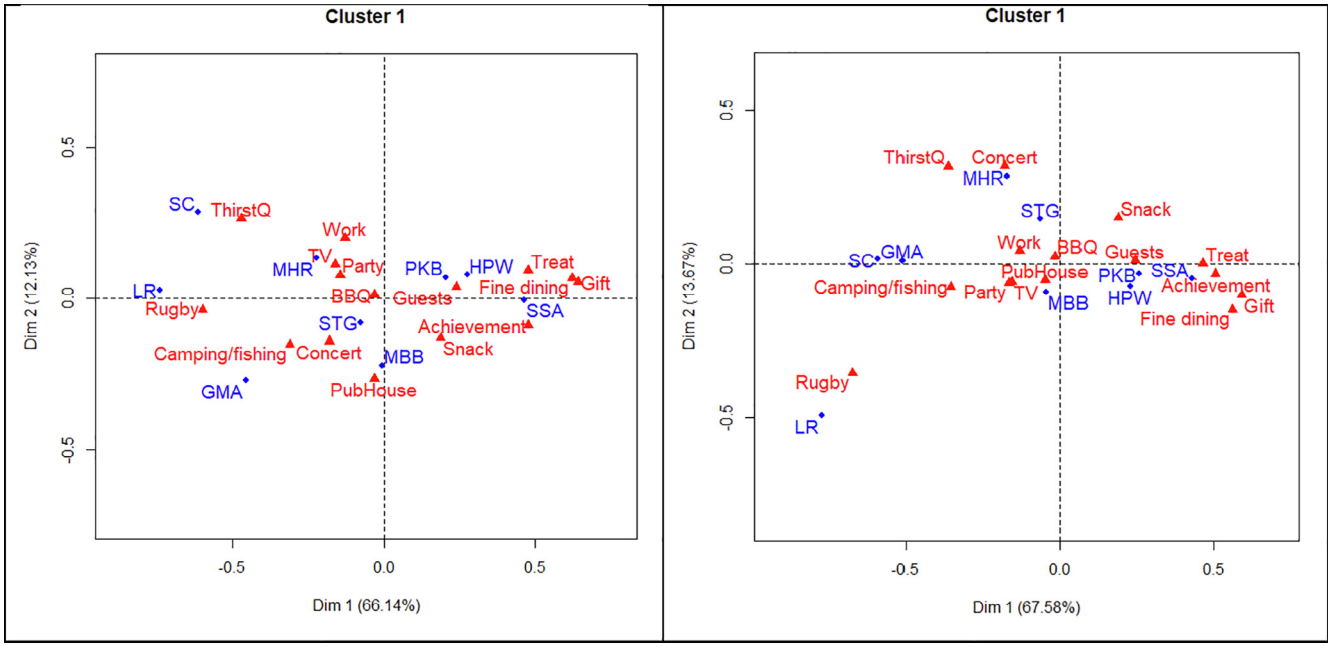

Fig. 4. The beers data: Representation of the products on the first two components of Cluster 1 with the atypical respondents included (left) and without these respondents (right).

"Rugby" is more emphasized than in the case when the atypical subjects were not removed.

3.5. Other case studies

By way of illustrating the strategy of analysis on a broader basis, we considered seven additional case studies from previously published CATA experiments to which we applied the general approach discussed in this paper. The objective is also to give orders of magnitude regarding the homogeneity indices, the number of attributes for which the respondents were not consistent, the gain in terms of homogeneity when we perform CLUSCATA, the number of respondents who are assigned to the "K + 1" cluster, *etc*. The characteristics of the case studies and the bibliographical references are shown in Table 3. We can see that the number of products varies from 4 to 12 and the number of respondents ranges between 73 and 154. In Table 4, we show the main results regarding the various steps of the strategy of analysis.

We can see that the homogeneity indices ranges between 21.3% for case study D1 to 70.7% for case study D4. However, if we exclude these two extreme case studies, the homogeneity of the other case studies is average to poor. The permutation tests applied to each case study in turn to assess the significance of the homogeneity or agreement indices lead to the conclusion that the null hypothesis stipulating that the respondents are in complete disagreement should be rejected.

As expected, the overall homogeneity increased after performing CLUSCATA and this increase was even more noticeable after the removal of those respondents who are deemed to be atypical. On average, one third of the respondents were set in the noise cluster.

The permutation tests regarding the consistency of the respondents for each attribute taken separately highlighted that in case studies D3, D4, D5 and D7, very few attributes revealed a consistency problem. However, more than 10 attributes were involved in a consistency problem for the remaining case studies (up to 40% of the evaluated attributes). For these case studies, the original number of attributes was relatively large and the agreement among the respondents as assessed by the homogeneity indices was poor. In parallel to the permutation test applied to the various attributes, we also run Cochran's Q test to assess their discrimination ability. The conclusions drawn from both tests (*i.e.*, the Cochran's and permutation tests) were the same for almost all 151 attributes across all the case studies. There were three exceptions: one attribute was assessed as non-discriminant but consistent, and two attributes were assessed as inconsistent but discriminant. These marginal differences can be explained by the fact the two hypothesis tests do not have exactly the same aim or can be attributed to the chance of making errors associated with a hypothesis test.

4. Conclusion

A strategy of analysis for the statistical treatment of CATA data was proposed. This strategy encompasses the evaluation of the agreement among the respondents at a global level and for each attribute considered separately, the clustering of the respondents while setting aside atypical respondents. The relevance of this analytical strategy was demonstrated on the basis of several case studies. The outputs from these

Table 3
CATA case studies.

Case study	Product category	*N* Products	*N* Attributes	*N* Respondents	Reference
D1	Beer	6	27	154	Giacalone, Bredie, and Frøst (2013)
D2	Beer	8	38	73	Reinbach, Giacalone, Ribeiro, Bredie, and Frøst (2014)
D3	Beer	9	15	97	Giacalone et al. (2015)
D4	Beer	4	10	87	Jaeger et al. (2013)
D5	Wine	12	17	112	Giacalone and Jaeger (2016)
D6	Coffee	6	30	83	Giacalone et al. (2018)
D7	Rye bread	6	14	134	Giacalone (2018)

Table 4
Main results of the various strategies of analysis.

Case study	Homogeneity index	*N* of inconsistent attributes (%)	*N* of clusters	Overall homogeneity without "K + 1" cluster	Threshold parameterρ	*N* of respondents in "K + 1" cluster (%)	Overall homogeneity with "K + 1" cluster
D1	21.3%	10 (37%)	3	24.8%	0.45	53 (34.4%)	30.5%
D2	38.2%	11 (28.9%)	3	42.7%	0.61	20 (27.4%)	48.2%
D3	45.5%	0 (0%)	4	51.8%	0.67	33 (34.0%)	60.1%
D4	70.7%	1 (10%)	3	74.7%	0.83	22 (29.9%)	79.9%
D5	48.5%	1 (5.9%)	2	51.1%	0.66	39 (34.8%)	62.8%
D6	31.8%	12 (40%)	3	35.8%	0.55	22 (26.5%)	40.7%
D7	31.7%	3 (21.4%)	3	37.3%	0.55	43 (32.1%)	44.1%

investigations showed cases with poor agreements among the respondents and cases with a relatively high degree of agreement. In all the cases, the clustering of the respondents yields an improvement of the overall agreement, which was even more noticeable when a noise cluster was added to the main clusters. The introduction of a noise cluster entails the selection of a noise parameter, ρ. We have proposed an analytical expression for this parameter which, on the basis of the presented case studies, seems appropriate. It is clear that this parameter depends on the number of clusters, on the structure of the data (*e.g.,* within and between variation associated with the clusters), on the number of datasets in each cluster, etc. Further developments regarding these aspects are needed.

This analytical strategy highlights an important issue related to the CATA experiment, that is, the number of attributes that are checked by each respondent for the various products. We proposed a standardization which consists in dividing each binary dataset by the square root of the total number of checked attributes for all the products. Notwithstanding, it still remains that those respondents who have a tendency to check more attributes will have a higher contribution to the analyses than those who check attributes more parsimoniously. A way to counteract this problem may be to instruct the respondents to check for each product the most salient attributes up to a pre-specified number of attributes (see *e.g.* Campo, Do, Ferreira, & Valentin, 2008).

Yet another issue concerns the number of attributes which can cause fatigue to the respondents when it is large. This problem is even more acute when the number of products to be assessed is itself large. Indeed, we can remark from the case studies discussed above that as a general rule the agreement among the respondents is rather poor when the number of attributes is high.

Some caution regarding the number of respondents involved in a CATA experiment should be taken when segmenting the subjects. For instance, take the scenario of the first case study where 76 respondents were involved in a CATA experiment. The segmentation of these respondents led to four clusters. It would be unreasonable to draw strong conclusions regarding the clusters with relatively small sample sizes. The findings from the cluster analysis should be taken for what they are worth, that is, indications about the presence of segments and some sources of variation between these segments. In any case, further actions such as increasing the number of respondents are needed.

In the future, more investigations will be dedicated to the tricky problem of determining the appropriate number of clusters when performing CLUSCATA. So far, the decision has been made on the basis of a visual examination of the dendrogram but a more formal choice based, for instance, on a hypothesis testing framework would be more indicated.

The analyses discussed in this paper with the exception of the permutation tests are available in the R package ClustBlock (Llobell, Vigneau, Cariou, & Qannari, 2019).

References

Berget, I., Varela, P., & Næs, T. (2019). Segmentation in projective mapping. *Food Quality and Preference, 71*, 8–20.

Carr, B. T., Dzuroska, J., Taylor, R. O., Lanza, K., & Pansini, C. (2009). *Multidimensional Alignment (MDA): A simple numerical tool for assessing the degree of association between products and attributes on perceptual maps. 8th Rose-Marie Pangborn Sensory Science Symposium*.

Campo, E., Do, B. V., Ferreira, V., & Valentin, D. (2008). Aroma properties of young Spanish monovarietal white wines: A study using sorting task, list of terms and frequency of citation. *Australian Journal of Grape and Wine Research, 14*, 104–115.

Cochran, W. G. (1950). The comparison of percentages in matched samples. *Biometrika, 37*, 256–266.

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

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th Ed.). Wiley Series in Probability and Statistics.

Giacalone, D. (2018). Product performance optimization. In G. Ares, & P. Varela (Eds.). *Methods in consumer research* (pp. 159–185). Elsevier.

Giacalone, D., Bredie, W. L., & Frøst, M. B. (2013). "All-In-One Test" (AI1): A rapid and easily applicable approach to consumer product testing. *Food Quality and Preference, 27*(2), 108–119.

Giacalone, D., Degn, T. K., Yang, N., Liu, C., Fisk, I., & Münchow, M. (2018). Common roasting defects in coffee: Aroma composition, sensory characterization and consumer perception. *Food Quality and Preference*.

Giacalone, D., Frøst, M. B., Bredie, W. L., Pineau, B., Hunter, D. C., Paisley, A. G., & Jaeger, S. R. (2015). Situational appropriateness of beer is influenced by product familiarity. *Food Quality and Preference, 39*, 16–27.

Giacalone, D., & Jaeger, S. R. (2016). Better the devil you know? How product familiarity affects usage versatility of foods and beverages. *Journal of Economic Psychology, 55*, 120–138.

Greenacre, M. J. (2007). *Correspondence analysis in practice.* Boca Raton, FL: Chapman & Hall/CRC.

Jaeger, S. R., Giacalone, D., Roigard, C. M., Pineau, B., Vidal, L., Giménez, A., & Ares, G. (2013). Investigation of bias of hedonic scores when co-eliciting product attribute information using CATA questions. *Food Quality and Preference, 30*(2), 242–249.

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

Llobell, F., Cariou, V., Vigneau, E., Labenne, A., & Qannari, E. M. (2019). A new approach for the analysis of data and the clustering of subjects in a CATA experiment. *Food Quality and Preference, 72*, 31–39.

Llobell, F., Vigneau, E., Cariou, V., & Qannari, E.M. (2019). ClustBlock: Clustering of datasets. R package version 1.0.0. https://CRAN.R-project.org/package = ClustBlock.

Llobell, F., Vigneau, E., & Qannari, E. M. (2019). Clustering datasets by means of CLUSTATIS with identification of atypical datasets. Application to sensometrics. *Food Quality and Preference*. https://doi.org/10.1016/j.foodqual.2019.02.017.

Meyer, C. D. (2000). *Matrix analysis and applied linear algebra.* Siam71.

Meyners, M., Castura, J. C., & Carr, B. T. (2013). Existing and new approaches for the analysis of CATA data. *Food Quality and Preference, 30*(2), 309–319.

Meyners, M., Castura, J. C., & Worch, T. (2016). Statistical evaluation of panel repeatability in Check-All-That-Apply questions. *Food Quality and Preference, 49*, 197–204.

Ochiai, A. (1957). Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions. *Bulletin of Japanese Society of Scientific Fisheries, 22*, 526–530.

Reinbach, H. C., Giacalone, D., Ribeiro, L. M., Bredie, W. L., & Frøst, M. B. (2014). Comparison of three sensory profiling methods based on consumer perception: CATA, CATA with intensity and Napping®. *Food Quality and Preference, 32*, 160–166.

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

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*, 54–63.

Worch, T., & Piqueras-Fiszman, B. (2015). Contributions to assess the reproducibility and the agreement of respondents in CATA tasks. *Food Quality and Preference, 40*, 137–146.