

Interuniversity Master in Statistics and Operations Research

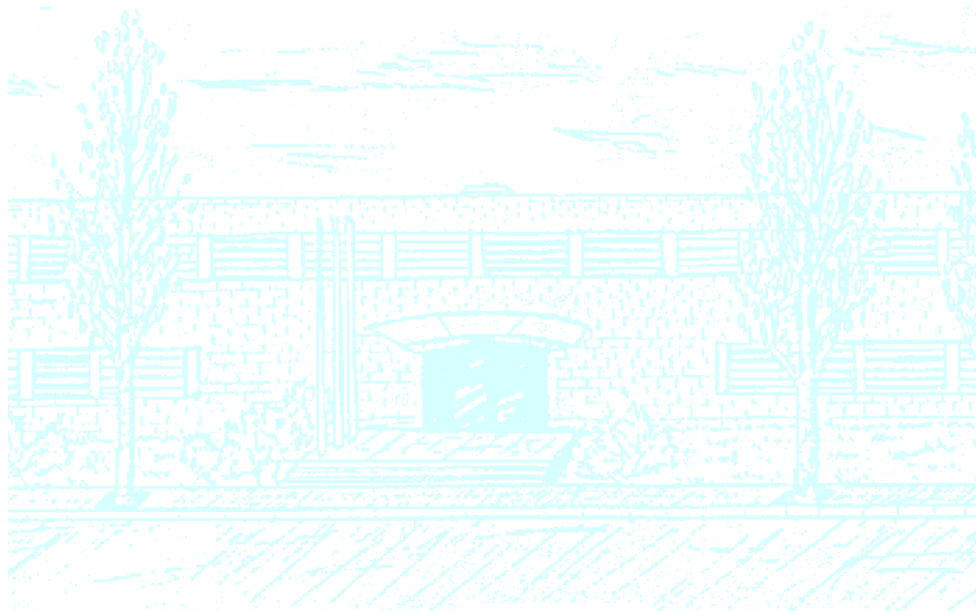
Title: Categorized napping, a sensometric tool for food and beverage industry. An application to a hall test session

Author: Belchin Adriyanov Kostov

Advisor: Mónica Bécue Bertaut

Department: EIO

University: UPC



Facultat de Matemàtiques
i Estadística

UNIVERSITAT POLITÈCNICA DE CATALUNYA



Thanks to everyone who helped me in some way to make this thesis. Specially my advisor, Mónica Bécue, for her explanations of statistical methods, my friend, Deniz Kutlu, for proofreading in English and my family for their moral support during all these years.

Resumen

E-mail del autor: belchin3541@gmail.com

Palabras clave: Análisis sensorial, datos de cata, cava, napping, free sorting task, napping categorizado, Análisis de Correspondencias Múltiple, Análisis Factorial Múltiple, Análisis Factorial Múltiple Jerárquico, sensometría

1. Introducción

El análisis sensorial trata del análisis normalizado de los alimentos realizado con los sentidos. Bajo el nombre de sensometría se reagrupan los métodos estadísticos que tratan este tipo de datos. Una importante área de aplicación es la industria del vino. Las grandes empresas empiezan a ver el potencial del análisis sensorial y cada vez organizan más catas con el fin de conocer sus productos y usar los resultados para mejorar su producción y marketing. No obstante, la utilización de estos métodos están todavía muy lejos del deseado por la falta de conocimiento general de ellos.

2. Objetivos

Es este estudio se analizan los datos de una cata con los siguientes objetivos

- I. Encontrar las similitudes entre diferentes tipos de cava a partir de los resultados de cata (método napping)
- II. Observar si las diferentes características de los cavas marcan alguna diferenciación entre ellos.
- III. Determinar si las diferencias observadas entre cavas tiene alguna relación con su composición química o perfil aromático.
- IV. Comparar las percepciones de los dos tipos de catadores, enólogos de Freixenet y estudiantes de Agrocampus.
- V. Ver si la variedad Chardonnay marca alguna diferenciación entre los cavas.
- VI. Estudiar el interés de los comentarios libres como datos sensometricos específicos.

3. Material y Métodos

La fuente principal de información de este proyecto está formada por los datos recogidos durante la jornada de cata de cavas realizada el 10 de febrero de 2009 en Freixenet.

Sobre 10 cavas escogidas por Freixenet, se aplica el napping categorizado.

- Napping^(R): Los catadores deben situar las copas de cava sobre un “mantel”, de tamaño 40cm x 60cm, de tal forma que dos cavas parecidos (según los propios criterios de cada catador) estén cerca y dos cavas distintos estén distanciados.
- Categorización (free sorting task): Los catadores deben marcar “clases de cavas” sobre los manteles. En el caso de 10 cavas, tenían que formar al menos 2 clases y como mucho 9 clases de cavas. Después deben de describir cada grupo por un conjunto de palabras.

La información recogida en la jornada de cata se complementa con otros 3 conjuntos de datos: descripción libre, análisis químico y análisis cromatografico. Para analizar los resultados se usan métodos estadísticos multidimensionales como análisis de correspondencias múltiple (ACM), análisis factorial múltiple (AFM) y análisis factorial múltiple jerárquico (AFMJ).

Además, en este trabajo se busca una metodología para cuantificar el consenso de un panel. Debido a la ausencia de proposiciones anteriores, esta parte es original.

4. Conclusión

Se observa que los expertos y los estudiantes tienen maneras de trabajar bastante distintas. Esta diferencia se manifiesta una vez separada el análisis global, obtenido mediante la aplicación del AFMJ, en subanálisis.

Abstract

E-mail of author: belchin3541@gmail.com

Keywords: Sensorial analysis, hall test, cava, napping, free sorting task, categorized napping, Multiple Factor Analysis, Hierarchical Multiple Factor Analysis, sensometrics

1. Introduction

Sensory analysis is the standard analysis of foods made with the senses. Under the name of sensometrics regroup statistical methods address to this type of data. An important application area of sensory analysis is the wine industry. The companies are beginning to see the potential of sensory analysis and they are organizing more hall test sessions every passing day to learn about their products and use these results to improve their production and marketing. However, yet these methods are not used very much because of their little knowledge between wine industry companies.

2. Objectives

In this study are analyzed dates of a hall test session with objectives

- I. Find the similarities between different types of cavas using results of hall test session (napping method)
- II. Observe if different features of cavas make some differentiation between them
- III. Determine if the differences between cavas have any relation with their chemical composition or flavor profile
- IV. Compare the perceptions of the two type of tasters: oenologist and students of Agrocampus
- V. See that if Chardonnay variety is a factor which separate the cavas
- VI. Study the interest of free comments as specific sensometrics data.

3. Material and Methods

The main source of information on this project consists of the data collected during the hall test session on 10th February 2009 at Freixenet.

Categorized napping is applied over 10 selected cavas.

- Napping^(R): Tasters should put the cavas on a "tablecloth", 40cm x 60cm size, so that two similar cavas (depending on own criteria of each taster) are close and two different cavas are spaced.
- Categorization (free sorting task): Tasters should make "clusters with cavas" on the tablecloths. In the case of the ten cavas, they should make at least two clusters and not more than nine. Then they should describe each cluster with some words.

The information collected on the hall test session is complemented by three sets of data: free descriptions, chemical analysis and chromatographic analysis. To analyze the results are used multidimensional statistical methods as multiple correspondence analysis, multiple factor analysis and hierarchical multiple factor analysis.

In addition, in this thesis is tried to establish a methodology to quantify the consensus of a panel. Due to the absence of previous proposals, this part is original.

4. Conclusions

It is noted that experts and students have very different ways of working. This difference manifests when the overall analysis, obtained by means of HMFA, is separated in sub-analysis.

Index

INTRODUCTION	6
DATA COLLECTION	7
1.1 HALL TEST SESSION	7
1.1.1 <i>Products</i>	7
1.1.2 <i>Preparation of the hall test session</i>	8
1.2 HALL TEST SESSION DATA.....	10
1.2.1 <i>Napping and free sorting task</i>	10
1.2.2 <i>Data table</i>	11
1.3 CHEMICAL AND CHROMATOGRAPHIC DATA	12
1.4 WORDS FREQUENCIES TABLE	13
STATISTICAL METHODS	14
2.1 FACTORIAL ANALYSIS	14
2.2 MULTIPLE CORRESPONDENCE ANALYSIS (MCA)	15
2.3 MULTIPLE FACTOR ANALYSIS (MFA)	16
2.3.1 <i>Data table</i>	16
2.3.2 <i>Balancing the sets of variables</i>	16
2.3.3 <i>MFA as a general factor analysis</i>	17
2.3.4 <i>Superimposed representation of the J clouds of individuals</i>	17
2.3.5 <i>Restricted transition formula</i>	17
2.3.6 <i>Global similarity between axial representations of the clouds N_j^j</i>	17
2.3.7 <i>Analysis in R^2 : Representation of the sets</i>	18
2.4 HIERARCHICAL MULTIPLE FACTOR ANALYSIS (HMFA)	19
GLOBAL ANALYSIS	21
3.1 DATA STRUCTURE.....	21
3.2 EIGENVALUES	22
3.3 CONFIGURATION OF CAVAS	22
3.4 REPRESENTATION OF COLUMNS	24
3.4.1 <i>Napping axes</i>	24
3.4.2 <i>Categorizations (free sorting task)</i>	24
3.4.3 <i>Chemical parameters and chromatographic variables</i>	25
3.4.4 <i>Words</i>	27
3.5 REPRESENTATION OF THE SETS	27
3.5.1 <i>First hierarchical level</i>	27
3.5.2 <i>Second hierarchical level</i>	28
3.5.3 <i>Third hierarchical level</i>	28
3.6 CLUSTER.....	29

3.7	CONCLUSIONS OF HMFA.....	30
SEPARATE ANALYSES		32
4.1	FREE SORTING TASK.....	32
4.1.1	<i>Eigenvalues.....</i>	32
4.1.2	<i>Configuration of cavas and description of clusters</i>	32
4.1.3	<i>Conclusions from free-sorting task results</i>	35
4.2	NAPPING	35
4.2.1	<i>Data structure and analysis method</i>	35
4.2.2	<i>Individual nappes</i>	36
4.2.3	<i>Eigenvalues structure in students and experts sets</i>	37
4.2.4	<i>Configurations of cavas.....</i>	37
4.2.9	<i>Supplementary elements: chemical parameters.....</i>	38
4.2.9	<i>Supplementary elements: chromatographic variables.....</i>	40
4.2.9	<i>Supplementary elements: words</i>	40
4.2.9	<i>Canonical correlation coefficients and inertia ratio</i>	41
4.2.9	<i>Representation of sets</i>	42
4.2.10	<i>Conclusions from napping results.....</i>	43
CONSENSUS AMONG THE PANELLISTS		45
5.1	QUANTIFICATION OF THE CONSENSUS LEVEL.....	45
5.2.1	<i>Statistical Test</i>	45
5.2.1	<i>Random panel.....</i>	45
5.2	CLUSTERING THE PANELLISTS.....	46
5.2.1	<i>Methodology.....</i>	46
5.2.3	<i>First method: Similarity graph</i>	47
5.2.3	<i>Second method: Hierarchical clustering</i>	48
5.3	CONCLUSIONS OF CONSENSUS LEVEL AND CLUSTERING	49
CONCLUSIONS.....		50
BIBLIOGRAPHY		51

Introduction

Statistics play a relevant and increasing role in many scientific and industrial fields. One of these fields is food and beverage industry, in particular to analyse data issued from hall test sessions. These sessions allow collecting sensorial data, which is about the perceptions of the products from vision, odour, taste and touch points of view. The analysis of these data bring answers to questions such as “*Are the products perceived as equal or different? Which are the most notable differences between them? Which are the characteristics that define each product better? Which are the preferences of the consumers and/or experts? Do typologies of consumers exist?*”

It is important to know that only one person’s opinion is not enough even if s/he is an expert. There is a great variety between individual opinions. So, it is necessary to define and collect the information such as to make possible its posterior statistical analysis.

The methods used to collect and make statistical analysis of information about the sensory aspects of the products are grouped under the name of *sensometrics*. Thus, sensometrics belongs to statistics. Presently, it follows a growing process, given that constantly new problems and new statistical methods appear.

The study that we present in this work corresponds to a hall test session organised in Freixenet, S.A.

In the first chapter we present the hall test session. The second chapter summarises the statistical methods that are used to analyse data. The results are presented in the third (global analysis) and fourth (comparison of the trained and untrained panels) chapters. In chapter five, we tackle the study of the homogeneity of the panels in an original way, looking for clustering the panellists depending on the consensus of their evaluations. Finally, we present conclusions and perspectives.

CHAPTER 1

DATA COLLECTION

In this chapter, we present the hall test session in a detailed way. The following are introduced; the products, the protocol, the two panels, experts and students, that have participated in the session. We also interpret the external data (chemical and chromatographic data) that were previously collected.

1.1 Hall test session

1.1.1 Products

The hall test session took place in San Sadurní d'Anoia (Barcelona) on 10th February, 2009. Products to taste were ten different cavas (*table 1.1*)

PRODUCT	LIQUOR TYPE	BRAND	YEAR	VARIETIES	WINES ORIGIN	SPECIAL ELABORATIONS
1(BA6CHC)	Brut (7 g/l)	A	2006	MA/XA/PA/10%CH	C	
2(NA5CHC)	Nature (< 3g/l)	A	2005	MA/XA/PA/10%CH	C	
3(BA6C)	Brut (9 g/l)	A	2006	MA/XA/PA	C	
4(NB4CHF)	Nature (< 3g/l)	B	2004	MA/PA/5%CH	F	
5(NB5CHF)	Nature (< 3g/l)	B	2005	MA/XA/PA/10%CH	F	
6(NB5F)	Nature (< 3g/l)	B	2005	MA/XA/PA	F	Fermented wines in barrel
7(NC5CHF)	Nature (< 3g/l)	C	2005	MA/XA/PA/5%CH	F	
8(BC3F)	Brut (4 g/l)	C	2003	MA/XA/PA	F	Cork
9(BC4F)	Brut (4 g/l)	C	2004	MA/XA/PA	F	
10(BC5F)	Brut (4 g/l)	C	2005	MA/XA/PA	F	

Macabeo (MA), Xarel·lo (XA), Parellada (PA), Chardonnay (CH)

Table 1.1. Description of cavas

Six variables define each cava.

1. *Liquor quantity*. The cavas are separated into groups depending on the sugar quantity they contain. In this study, only brut and nature cavas are considered. Brut cavas are sweeter than nature cavas.
2. *Brand*. The 10 cavas are produced from three different brands: *A*, *B* and *C* (original names of brands are not published due to a “confidentiality agreement” with Freixenet).
3. *Production year*. The cavas were produced between 2003 and 2006.
4. *Varieties*. The cavas are made of four varieties: *Macabeo*, *Xarel-lo*, *Parellada* and *Chardonnay*.
5. *Production unity*. Two different production unities: *C* and *F* (original names of production unity are not published for “confidentiality agreement”)
6. *Special elaboration features*. Only two cavas present special elaboration features. One cava has suffered *fermentation in barrel*; another cava is placed on the top of the bottle with cork during the second fermentation (traditional way).

In this study, each cava is identified through a label, which summarizes its characteristics. The first character indicates the liquor type (B: Brut, N: Nature); the second corresponds to the brand (A,B or C); the third comes from the year (3: 2003, 4: 2004, 5: 2005 and 6: 2006), “CH” means that cava contains some proportion Chardonnay variety (if not, no Chardonnay is included in the blend) and the last letter indicates the production unity (F or C). These labels make easier the graphics and numerical results readings.

1.1.2 Preparation of the hall test session

A hall test session is organized to evaluate some characteristics of a product or a group of products. A panel, which includes a group of panellists, evaluates these characteristics. The objective of the hall test session is to detect concordances and differences between the products and to determine which characteristics can explain the different perceptions.

The preparation of a hall test session is a delicate process because of many factors that have an influence on the results. For example, the panellists (experts or no experts), trial conditions (number of samples, preparation and presentation of the samples) and also the methods that are used to collect the information and determine its subsequent analysis.

Panellists: There is an important variability among the panellists. It is possible to build a panel of experts, non-experts or mix depending on the hall test objectives. In our

case, twenty panellists constituted have intervened. Ten were students of *Agrocampus-Ouest* and other ten experts (oenologist). One of the main objectives of the hall test was to compare the both of subpanels.

Installations: The installations of the hall test sessions are normalized. These installations are divided into two parts. One part is dedicated to the preparation and the other part consists of separated tables or individual booths for tasters. The separation of these two parts is important to prevent possible factors which could influence the tasters' opinion.

The installations should have a light air pressure to prevent arrival of smells from other places. All samples must be prepared before the starting of the hall test session. Tasters need to have drinkable water to rinse the mouth between sample tasting.

As it is recommended, tables should be easy to clean and have pleasant colours like neuter light gray. Lighting is another important factor. It must be uniform, enough but not intense to influence appearance of products.



Figure 1.1. Example of hall test session installations

Hour: The period of the day the test done is also important. Before meals the sensibility is higher but it is easy to take hasty decisions too. After meals, the sensibility is significantly reduced. So, it is important to avoid extreme schedule. In our case, the session began at 12 a.m.

Codification and presentation order of samples: It has been observed that the tasters have more strict judgments in the case of the first samples. So, the presentation order of the samples plays an important role in the results. To determine the presentation order, an experimental design (for example Latin squares) has to be used. Each cava has to be codified in such a way where, no information is provided to the taster about its identification. Usually a three-digit code is used to identify each cava.

Glass: The “*flute*” is the more often-used kind of glass to drink cava. This glass is known to be long and narrow. It keeps the temperature steady and the liquid fizzy. Its only disadvantage is it is difficulty to smell the cava.



Figure 1.2. Glass “flute”

1.2 Hall test session data

1.2.1 Napping and free sorting task

Ten cava samples were simultaneously submitted to each taster, who was asked to position them on a large sheet of blank paper, size of 40 cm x 60 cm corresponding to the standard size of the hall test booth.

They were asked to evaluate the similarities (or dissimilarities) between the ten cava samples according to their own criteria, those that are important to them. Criteria are implicit.

Cava samples had to be positioned on the tablecloth in such a way that two cava samples were very close if they seem alike and far from each other if they seem different. Once the operation was completed, they wrote down on the sheet the number of the cava and the place that it occupied.

The napping data are coded into a table indicating, for each cava, its x-axis and its y-axis on the sheet. The origin can be placed anywhere (the left bottom corner is easy).

A small example of napping data is shown in *figure 1.3*. Each taster provides a tablecloth like this.

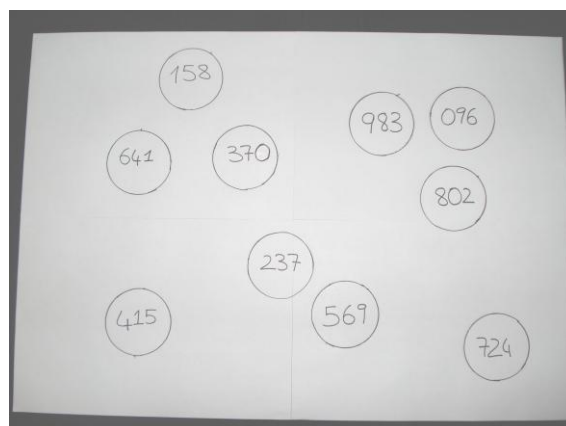


Figure 1.3. Example of napping

After having performed the napping, the tasters were requested to gather the cava samples into clusters. They had to make at least two clusters and a maximum of nine. They were also asked to write some words to describe each cluster. The name of this

method is categorisation or *free sorting task*. The joint implementation of napping and free sorting task is called *categorized napping*. An example of the categorized napping of one particular taster is displayed at *figure 1.4*.

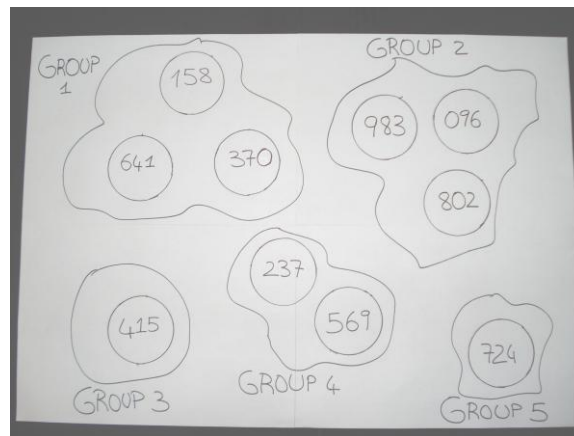


Figure 1.4. Example of categorized napping

1.2.2 Data table

Napping data jointly with free sorting task are filed into a multiple table (*figure 1.5*). For each cava i and each taster j , we have the x-axis x_{ij} and the y-axis y_{ij} (issue from napping data) and the cluster w_{ij} in which the taster j has included the cava i (free sorting task). It is important to remember that the cavas belonging to the same cluster are characterized by the same words.

	Taster 1				Taster 20		
Cava 1	$x_{1,1}$	$y_{1,1}$	$w_{1,1}$		$x_{1,20}$	$y_{1,20}$	$w_{1,20}$
...
Cava 10	$x_{10,1}$	$y_{10,1}$	$w_{10,1}$		$x_{10,20}$	$y_{10,20}$	$w_{10,20}$

Figure 1.5. Database (napping data, x_{ij} and y_{ij} , plus free sorting task w_{ij})

1.3 Chemical and Chromatographic Data

Chemical data: Eleven chemical variables were measured on the cavas:

- Alcoholic grade: Percentage of alcohol at total volume.
- PH: A measure of acidity or basicity of a solution.
- Total acidity (H₂SO₄): Sulphuric acid (gram/litre).
- Free sulphur dioxide: portion of sulphur dioxide (milligram/litre).
- Total sulphur dioxide: sum of free sulphur dioxide and bound sulphur dioxide (milligram/litre). It has not got any sensorial repercussion. Is an additive using as antioxidant or antiseptic.
- Total sugar: Portion of sugar at composition (gram/litre).
- D.O 420nm: Measure of optical density of yellow. When D.O. 420nm is higher it means cava is more yellow, more evolved and oxidized too.
- Malic acid: Portion of malic acid (gram/litre).
- Lactic acid: Portion of lactic acid (gram/litre). Grapes has only malic acid. Lactic comes from the transformation of malic through lactic bacteria. Cavas are softened when this fermentation occurs.
- Glycerol: Portion of compound glycerol (gram/litre). Glycerol gives stickiness and volume in mouth.
- Dry extract: The powder that is left when cava is placed in a centrifuge and all of the water is removed . Dry extract gives body and width.

Chromatographic data: Gas chromatography (GC) is a separation technique that can be used for both the qualitative and quantitative identification of materials. It relies on the selective adsorption and desorption of volatile components on a stationary phase. The components are carried through the column by an inert gas to a detector. Common detectors for gas chromatography include flame ionization (FID), thermal conductivity (TCD), and mass spectrometry (MS). Components are identified based on retention time, and, where available, mass spectrum.

chemical or chromatographic variable k

	Variable 1	Z_{ik}	Variable K
Cava 1	$z_{1,1}$		$z_{1,K}$
Cava i
Cava 10	$z_{10,1}$		$z_{10,K}$

Figure 1.6. Database (Chemical and chromatographic data)

Z_{ik} is the value of the variable k for cava i

In this study is disposed information of more or less fifty compounds that are obtained from chromatographic analysis. There is an example of chromatographic analysis at *figure 1.7*. Each peak represents a compound. The height of peaks is related with quantity of compound in cava. So the most important compounds correspond to highest peaks

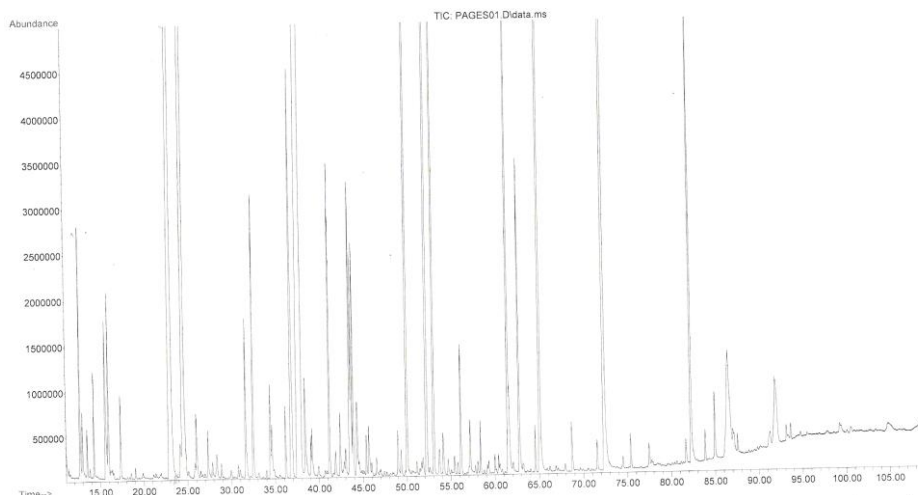


Figure 1.7. Example of chromatographic analysis

1.4 Words frequencies table

A words frequencies table has been built from the words used to characterize each cluster. For each cava i , the frequency with which each word is used to describe it is counted. Words with a total frequency (sum frequency for all cavas) smaller than 2 have been eliminated from the table.

	Word 1		Word T
Cava 1	$f_{1,1}$	f_{it}	$f_{1,T}$
Cava i
Cava 10	$f_{10,1}$		$f_{10,T}$

Figure 1.8. Database (words frequency)

f_{it} is the frequency of the word t for cava i

CHAPTER 2

STATISTICAL METHODS

In this chapter, we recall the general factorial analysis and then the three particular methods that are used in the further chapters: Multiple Correspondence Analysis (MCA) for free sorting task data, Multiple Factor Analysis (MFA) for napping data and Hierarchical Multiple Factor Analysis (HMFA) for categorized napping data.

2.1 Factorial Analysis

Given a table \mathbf{Z} with I rows and K columns, two clouds are built: the cloud of the I row-points N_I in R^K , and the cloud of K column-points N_K in R^I .

	Row-points cloud	Column-points cloud
Space	R^K	R^I
Metric	\mathbf{M}	\mathbf{D}
Data Matrix	\mathbf{Z}	\mathbf{Z}'
Weights	\mathbf{D}	\mathbf{M}
Axes of inertia	\mathbf{U}	\mathbf{V}
Equation	$\mathbf{Z}'\mathbf{DZMU} = \mathbf{U}\Lambda$ (Eq. 1)	$\mathbf{ZMZ}'\mathbf{DV} = \mathbf{V}\Lambda$ (Eq.2)
Orthonormality	$\mathbf{U}'\mathbf{MU} = \mathbf{Id}$	$\mathbf{V}'\mathbf{DV} = \mathbf{Id}$
Principal components	$\mathbf{F} = \mathbf{ZMU}$	$\mathbf{G} = \mathbf{Z}'\mathbf{DV}$
Equation	$\mathbf{ZMZ}'\mathbf{DF} = \mathbf{F}\Lambda$ (Eq.2bis)	$\mathbf{Z}'\mathbf{DZMG} = \mathbf{G}\Lambda$ (Eq.1bis)
Orthogonality	$\mathbf{F}'\mathbf{DF} = \Lambda$	$\mathbf{G}'\mathbf{MG} = \Lambda$
Equation (symmetrical form)	$\mathbf{M}^{1/2}\mathbf{Z}'\mathbf{DZM}^{1/2}\tilde{\mathbf{U}} = \tilde{\mathbf{U}}\Lambda = \mathbf{M}^{1/2}\mathbf{U}\Lambda$ (Eq.1ter)	$\mathbf{D}^{1/2}\mathbf{ZMZ}'\mathbf{D}^{1/2}\tilde{\mathbf{V}} = \tilde{\mathbf{V}}\Lambda = \mathbf{D}^{1/2}\mathbf{V}\Lambda$ (Eq.2ter)
Transition relations between the principal components in both spaces	$\mathbf{F} = \mathbf{ZMG}\Lambda^{-1/2}$	$\mathbf{G} = \mathbf{Z}'\mathbf{DF}\Lambda^{-1/2}$

Table 2.1. General scheme shared by the classical principal axes methods.

\mathbf{D} and \mathbf{M} are diagonal matrices.

The objective of the factorial analysis is to look for orthogonal axes which maximize the inertia of, respectively, clouds N_I and N_K as projected on these axes, called *principal axes*. In other words, factorial analysis aims to visualize the proximities between the variables, on the one hand, and between the individuals, on the other hand, as well as

the relationships between individuals and variables by representing both clouds on a series of axes that retain greater inertia.

The rows and columns can be weighted. The weights of the rows are filed into diagonal matrix **D** (general term d_i) and the weights of the columns are filed into diagonal matrix **M** (general term m_k).

The general factorial analysis is summarised in Table 2.1. Eq.1 and Eq.2 give the expressions of the matrixes to be diagonalized. The expression of the principal components and the relationships between Eq.1 and Eq.1bis, on the one hand, and between Eq.2 and Eq.2bis, on the other hand, show that either only Eq.1 or only Eq.2 have to be solved for computing both series of principal components. Alternatively to equation Eq.1 (respectively Eq.2), Eq. 1ter (respectively, Eq.2ter) can be solved, taking advantage of the symmetrical form of the matrix.

Specific computing of **Z**, **M** and **D** from the data lead to the classical principal axis methods, such as principal component analysis, correspondence analysis and multiple correspondence analysis (Lebart et al. 2004; Escofier & Pagès, 1988-2008).

2.2 Multiple Correspondence Analysis (MCA)

MCA is the particular factorial analysis used to tackle a table with I individuals and Q qualitative variables. These data sets are coded into a complete disjunctive table **Y** with K columns corresponding to the K categories of the Q qualitative variables. Matrix **Y** represents the set of categorical variables. Each categorical variable is expressed as a group of indicator (0,1) variables (a binary variable has two columns, a nominal three-level variable has three columns, etc., each column representing one category from one variable).

$y_{ik} = 1$ if individual i belongs to category k , $y_{ik} = 0$ if not. We note I_k the number of individuals belonging to category k . From matrix **Y**, the proportion matrix **F** is built up with general term $f_{ik} = \frac{y_{ik}}{IQ}$. The marginal terms of this table are filed, respectively,

in matrices **D** (general term $f_i = \frac{1}{I}$) and **M** (general term $f_{.k} = \frac{I_k}{IQ}$).

The matrix **Z** is built up with general term $z_{ik} = \frac{f_{ik}}{f_i \cdot f_{.k}} - 1 = \frac{Iy_{ik}}{I_k} - 1$

MCA considers three series of objects: individuals, variables and categories. Two individuals are similar if they share a great number of categories. Two categories are similar if they are frequently chosen by the same individuals.

2.3 Multiple Factor Analysis (MFA)

2.3.1 Data table

Multiple Factor Analysis (MFA; Escofier & Pagès 1988-1998 ; Pagès 2002) deals with multiple table in which a set of individuals is described by several sets of variables. Within one set, variables must present the same type (quantitative or categorical) but set of variables can belong to different types.

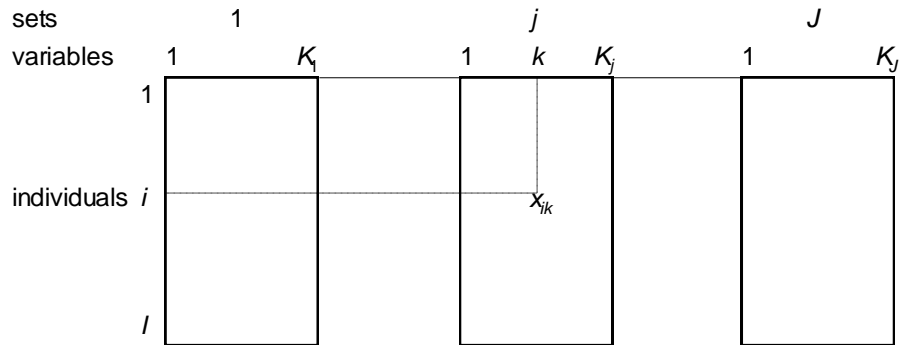


Figure 2.1. Data table.

x_{ik} : value of variable k for individual i . If k is a continue variable, x_{ik} is a real number ; if k is a categorical variable, x_{ik} is a number of category. The j^{th} set is denoted by j or K_j .

Individuals, noted i ($i=1, \dots, I$), constitute the cloud N_I in the K -dimensional space R^K ; the K variables, noted k ($k=1, \dots, K$) constitute the cloud N_K in the I -dimensional space R^I .

If we consider the only (sub-)table j , individuals are noted i^j ($i=1, \dots, I$) and constitute the cloud N_I^j in the K_j -dimensional space R^{K_j} ; the K_j variables constitute the cloud $N_{K_j}^j$ in the I -dimensional space R^I .

2.3.2 Balancing the sets of variables

The global analysis, where several sets of variables are simultaneously introduced as active, requires balancing the influences of the sets of variables. The influence of one set j derives from its structure (of the two clouds N_I^j and $N_{K_j}^j$ it induces) in the different space directions. If a set presents a high inertia in one direction, this direction will strongly influence the first axis of the global analysis. That suggests normalising the highest axial inertia of each set which is done by weighting each variable of the set j by $1/\lambda_j^1$, being λ_j^1 the first eigenvalue issued from the factor analysis applied to set j . Thus, MFA weighting normalises each of these two clouds by making its highest axial inertia equal to 1.

This weighting does not balance total inertia of the different sets. Thus, a set with a high dimensionality will contribute to numerous axes.

2.3.3 MFA as a general factor analysis

The basic principle of MFA is a general factor analysis applied to the multiple (global analysis). MFA works with continuous variables as principal component analysis does, the variables being weighted; MFA works with categorical variables as multiple correspondences analysis does, the variables being weighted.

MFA provides the classical outputs of general factor analysis:

- Co-ordinates, contributions and squared cosines of individuals
- Correlation coefficient between factors and continuous variables
- For each category, co-ordinate of the centroid of the individuals belonging to this category

2.3.4 Superimposed representation of the J clouds of individuals

We associate the cloud N_i^j of individuals in the space R^{K_j} to each set j . This “partial” cloud, is analysed in the factor analysis restricted to set j ; it contains “partials” individuals, noted i^j (individual i according to the set j).

To determine the resemblances, from one cloud to another, among distances between homologous points, the clouds N_i^j are projected upon the axes of the global analysis, as illustrative elements. The co-ordinate of i^j along axis s is denoted: $F_s(i^j)$.

2.3.5 Restricted transition formula

The co-ordinate $F_s(i^j)$ can be calculated from the coordinates of the variables $G_s(k)$, $k \in K_j$, by the way of the following relationship:

$$F_s^j(i) = F_s(i^j) = \frac{1}{\sqrt{\lambda_s}} \frac{1}{\sqrt{\lambda_1^j}} \sum_{k \in K_j} x_{ik} G_s(k)$$

We recognise here the usual transition formula but restricted to the variables of the set K_j .

2.3.6 Global similarity between axial representations of the clouds N_i^j

When the different sets induce similar structures on individuals, homologous points $\{i^j, j=1, \dots, J\}$ are close one another. This global property is measured, per axis, through the ratio computed as explained hereafter.

All the points of all the clouds N_i^j ($j = 1, J$) are considered. A partition of these $I \times J$ points in I classes is performed, such as the J homologous points $\{i^j, j=1, J\}$ corresponding to the same individual i belong to the same class. When axis s brings

out a structure common to the different sets of variables, the homologous points i^j , corresponding to the same individual i , are close one to the other and this partition has a low within-inertia (along axis s). The ratio (*between-inertia*) / (*total-inertia*) can be calculated for each axis. This ratio is close to 1 when the axis represents a structure common to the different sets.

2.3.7 Analysis in R^2 : Representation of the sets

MFA also visualizes the proximities between the sets, as represented, each of them by a unique point. In this visualization, two sets are close one another if they induce similar structures on the individuals.

Each set of variables K_j , is represented by the $I \times I$ matrix W_j of scalar products between individuals ($W_j = X_j X_j'$). Each scalar product matrix W_j can be represented by one point in the I -dimensional Euclidean space (denoted R^I). Thus, in this space, one set is represented by one point: the J points constitute the set cloud, denoted N_J . In this cloud N_J , the distance between two points W_j and W_l decreases as the similarity between the structures (defined upon individuals) induced by the sets K_j and K_l increases. For this reason, it is interesting to get a representation of the cloud N_J .

The representation provided by MFA is obtained by projecting N_J upon vectors (in R^I) induced by I -factors of global analysis (one factor may be considered as a set including a single variable; it is possible to associate to this set a scalar product matrix and thus a vector in R^I).

The normalised factor of rank s in R^K , previously denoted z_s , induces $w_s = z_s z_s'$ in R^I . Some properties of z_s induce corresponding properties for w_s :

$$z_s' z_t = 0 \Rightarrow \langle w_s, w_t \rangle = 0$$

$$\|z_s\| = 1 \Rightarrow \|w_s\| = 1$$

The main interest of this projection space is that its axes (upon which N_J is projected) are interpretable and, above all, possess the same interpretation that axes of global analysis (in the same manner, due to factor analysis duality, axis of rank s upon which individuals are projected and axis of rank order s upon which variables are projected possess the same interpretation).

This representation has the following property: it can be shown (Escofier & Pagès 1998 p 167) that co-ordinate of set j upon axis of rank s is equal to $L_g(z_s, K_j)$.

Thus:

- Set co-ordinates are always comprised between 0 and 1;

- A small distance between two sets along axis s means that these two sets include the structure expressed by factor s each one with the same intensity. In other words, set representations shows which ones are similar (or different) from the point of view of global analysis factors.

2.4 Hierarchical Multiple Factor Analysis (HMFA)

Hierarchical Multiple Factor Analysis (HMFA) extends the principles of MFA to multiple tables presenting a hierarchical structure on the variables. HMFA uses a sequence of MFA analyses in a sequential way to obtain a set of column weights to be used in a weighted and nonstandardized PCA global analysis that will balance the effects of the different sets of variables at every level of the hierarchy and within hierarchies.

HMFA provides graphical displays, which highlight the relationships among the individuals, on the one hand, and sets of variables, on the other hand, according to the various levels of the hierarchy. From the PCA performed on the whole data set, it is possible to depict the relationship among individuals on the basis of the first principal components. It is also possible to have partial representations on these individuals that are representations on the basis of a subset of variables. In HMFA, as in other statistical methods dealing with several data tables, there are as many partial representations for each individual as there are data tables. An interesting feature of the analysis is that the partial representation of each individual at each node is at the centroid of the partial representation of this individual associated with the various subsets of variables nested within this node.

HMFA is performed through the following steps:

Step 1. At the lowest level of the hierarchy, HMFA performs step 1 of MFA. The first eigenvalues at this step are named $\lambda_1^{h,j}$, where $h=1$ and $j=1,2,\dots,g_1$ (where $g_1 = J_q + J_c$ is the number of set of variables at this level).

Step 2. At the next higher level of the hierarchy, HMFA performs step 1 of MFA again within each of the high level set, obtaining a new set of g_2 (number of sets at the high level) eigenvalues $\lambda_1^{h,j}$, where $h=2$ and $j=1,2,\dots,g_2$, g_2 being the number of sets at the second level. If the hierarchy includes more than two levels (for example, p levels), this step is repeated to obtain p sets of eigenvalues according to the number of sets at each level.

Step 3. A global weighted and nonstandardized PCA on the whole \mathbf{X} matrix is then performed using $1/I$ as every row weight (each entry has equal weight) and the product of calculated column weights across the hierarchy as the weight column:

$$\prod_{h=1}^p \frac{1}{\lambda_1^h} \text{ for columns in the } X_j \text{ matrices (quantitative variables)}$$

$$\frac{w_{kj}}{Q_j} \prod_{h=1}^p \frac{1}{\lambda_1^h} \text{ for columns in the } Y_j \text{ matrices (categorical variables)}$$

$$Q_j = \sum_{k \in K_j} w_{kj} \quad w_{kj} = \sum_{i \in I} P_i z_{ikj}$$

Pagès (2004) proposed a procedure for measuring the contribution of one original variable (j_q if continuous, j_c if categorical) to the variability of a new axis v . This author showed that the total variability explained by one variable (from the mixture of continuous j_q and categorical j_c variables) on the new axis v could be expressed as

$$\sum_{j \in J_q} r^2 \langle \mathbf{X}, \mathbf{v} \rangle + \sum_{j \in J_c} \eta^2 \langle \mathbf{X}, \mathbf{v} \rangle = 1$$

where r is the correlation coefficient between each original variable and the new axis, and η is the correlation coefficient between the set of k_j indicator variables associated with each categorical variable and the new axis. Using these concepts, HMFA allows measuring the contribution of each variable and each set of variables to each of the new principal axis obtained in the final result.

HMFA provides a representation of the nodes involved in the hierarchy. The principle of this representation is similar to that of MFA: for each set of variables the index L_g between this set and each principal component is computed. This index reflects the extent to which the set of variables and the principal component under consideration are related. It ranges between 0 and 1. It is equal to 1 if the first principal component derived from HMFA is equal to the first principal component of the set of variables. On the contrary, this index is equal to 0 if the first principal component from HMFA is uncorrelated with any variable in the set.

There is an example of HMFA structure in figure 2.2. Single difference from previous graphic (figure 2.1), now we take account hierarchical structure on the data.

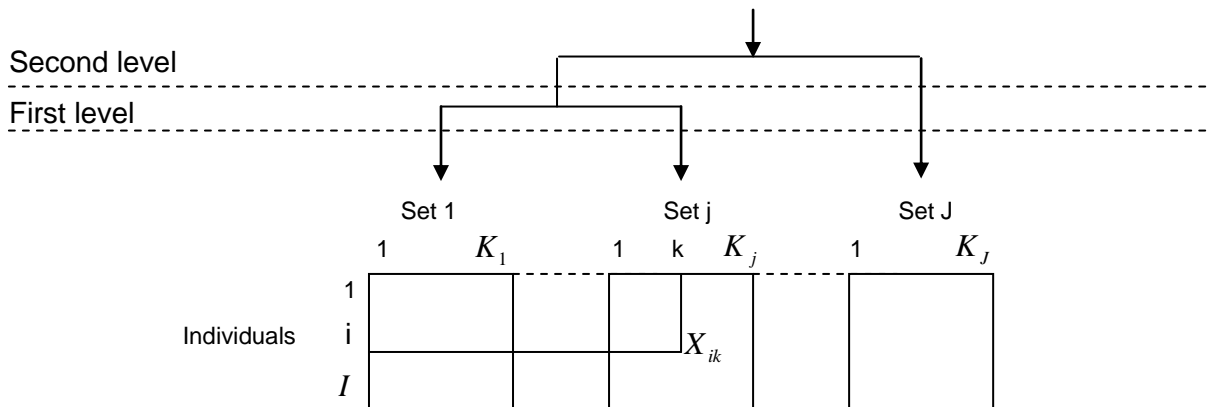


Figure 2.2. Hierarchical structure of the data table

Chapter 3

GLOBAL ANALYSIS

The first step when analyzing the categorized napping data collected at Freixenet consists in a global analysis, taking into account the whole of the data. HMFA was applied to categorized napping at the end of the hall test session. The objective of this analysis was to give a summary of the information obtained from the hall test session. Principally, the organizers would like to know about:

- I. Similarities and differences between cavas
- II. Similarities and differences between students and experts
- III. The most important factors which differentiate the cavas

In this chapter, we detail how the data have been coded. First, we present the results obtained at the end of the hall test session through performing HMFA. Then, we show how these results have been enriched by using a chemical and chromatographic description of the cavas, on the one hand, and by using the free-text description of the cavas. Finally, clustering the cavas has allowed for a synthetic summary.

3.1 Data structure

The multiple table that we analyse (chapter 1.2.2) present a hierarchical structure in three levels on the columns. At the third level, the tasters are divided into two sets depending on they are experts or students. At the second level, every third level set, expert and student, is divided into ten set-tasters (there were ten students and ten experts). Finally, the first level splits napping data and free sorting task of each taster into two sets (*figure 3.1*). HMFA realizes a subanalysis at each level starting with the lowest level until the highest. In this way are obtained a) a global representation of the cavas b) partial representations of the cavas and c) a representation of the sets at each level of the hierarchy.

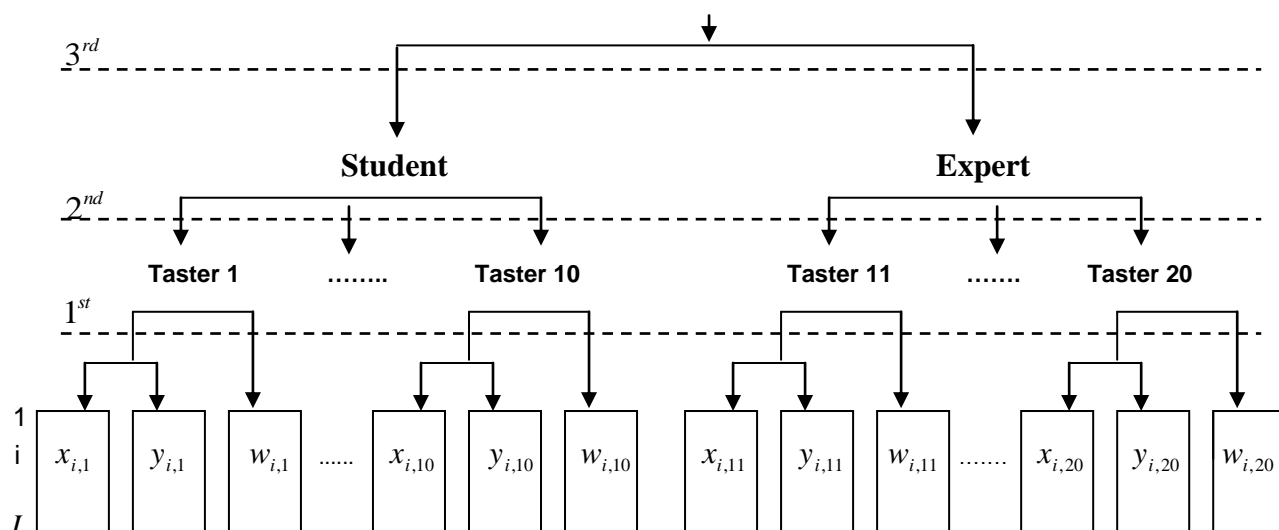


Figure 3.1. Hierarchy of data set. X_{ij} is x-axis of napping, Y_{ij} is y-axis of napping and W_{ij} is free sorting task (cava i and taster j)

3.2 Eigenvalues

The first eigenvalue of HMFA has a value comprised between 1 and the number of sets at the highest level of the hierarchy. If it is nearby to the number of sets at the highest level it means that the sets of the highest level are similar. In this study, the first eigenvalue is 1.87 (*table 3.1*) and it is nearby to 2 (number of sets at third level). So, according to HMFA it is possible to say that the sets of the third level, expert and student, have similar representations of cavas. The eigenvalues decrease slowly and it is necessary to take into account the third and fourth dimensions to keep more than half of the variance between cavas (cumulative percentage of variance for fourth dimension is equal to 58,09%)

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	Dim 6	Dim 7	Dim 8	Dim 9
Eigenvalue	1,87	1,71	1,46	1,30	1,19	1,04	0,92	0,82	0,60
Percentage of Variance	17,11%	15,68%	13,41%	11,89%	10,92%	9,56%	8,44%	7,48%	5,52%
Cum. percentage of variance	17,11%	32,79%	46,20%	58,09%	69,01%	78,57%	87,01%	94,48%	100%

Table 3.1. Eigenvalues, percentage of variance and cumulative percentage of variance of HMFA

3.3 Configuration of cavas

The importance of cavas on axes is determined by its contributions on these axes. Two special cavas dominate on the first factorial plane (first and second axis): 2(*NA5CHC*) and 6(*NB5F*). The sum of their contributions on the first dimension is more than 80% of the total contribution for all cavas on this dimension (*table 3.2*).

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
1(<i>BA6CHC</i>)	3.141	0.603	30.744	0.359	9.206
2(<i>NA5CHC</i>)	28.977	22.883	23.342	9.930	0.213
3(<i>BA6C</i>)	1.058	0.000	10.408	0.972	17.784
4(<i>NB4CHF</i>)	0.106	4.845	3.382	1.120	0.688
5(<i>NB5CHF</i>)	0.224	0.023	0.510	58.405	7.852
6(<i>NB5F</i>)	53.242	34.210	0.208	1.413	0.477
7(<i>NC5CHF</i>)	1.527	0.245	1.440	12.742	0.033
8(<i>BC3F</i>)	6.860	17.623	15.858	1.323	0.073
9(<i>BC4F</i>)	1.894	17.337	0.157	3.328	6.701
10(<i>BC5F</i>)	2.971	2.230	13.953	10.408	56.974

Table 3.2. Contributions of cavas for the first five dimensions

Characteristics of these cavas

6(*NB5F*) is the only cava which has a part of fermentation in barrels. 2(*NA5CHC*) had a cork defect. This defect was detected by the most of the tasters.

The second axis opposes these two cavas to 4(*NB4CHF*), 8(*BC3F*) and 9(*BC4F*) which are the oldest cavas (*figure 3.2*).

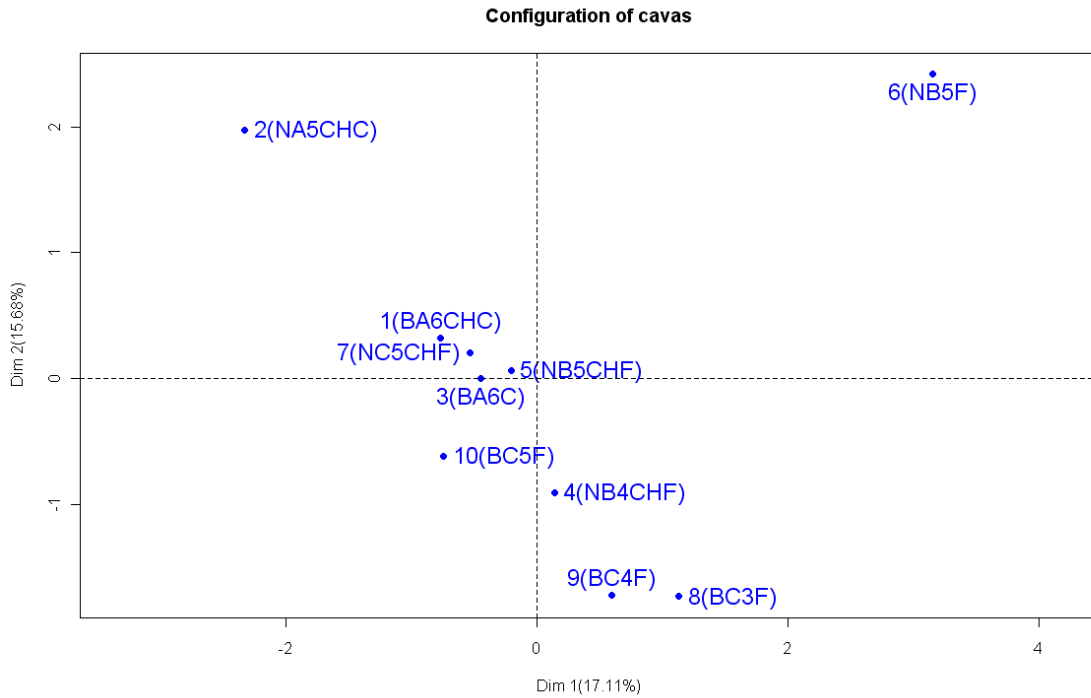


Figure 3.2. Configuration of cavas on the first factorial plane (first and second axis)

The third axis opposes the youngest cavas *1(BA6CHC)*, *3(BA6C)* and *10(BC5F)* to *2(NA5CHC)*, *4(NB4CHF)* and *8(BC3F)*. This axis makes an approximate ordination of the cavas depending on the production year (*figure 3.3*). The coordinate of *2(NA5CHC)* on this dimension indicates that cork defect gives an old cava perception. The fourth axis is built by *5(NB5CHF)* which has more than half of the total contribution of this dimension. Fifth and sixth axes are also dedicated specially to one cava.

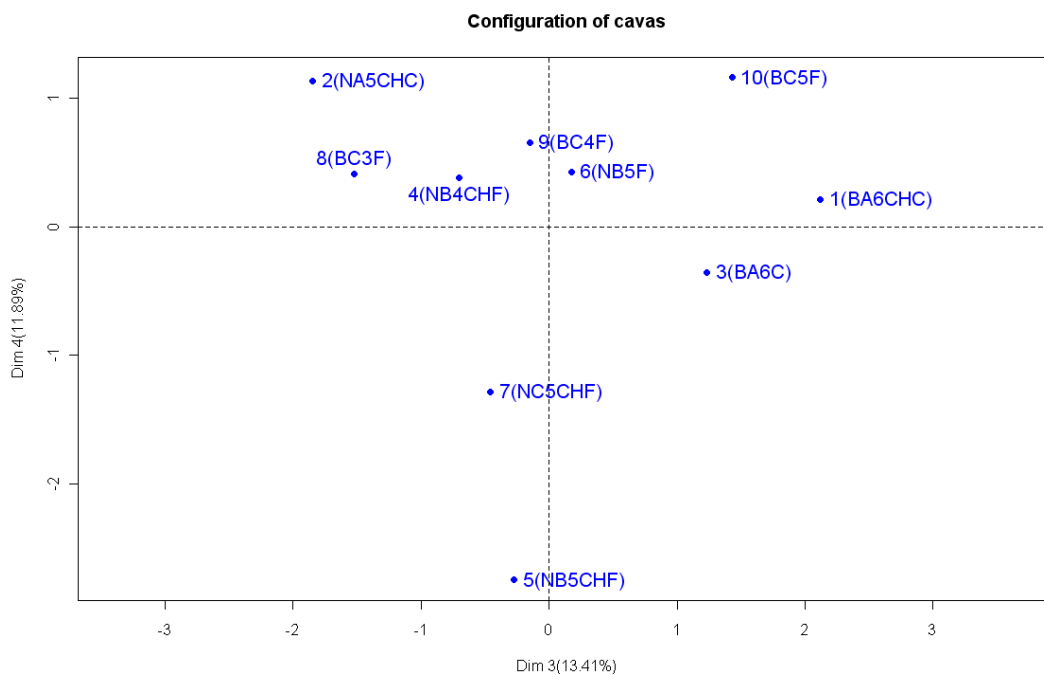


Figure 3.3. Configuration of cavas on the third and fourth axis

3.4 Representation of columns

3.4.1 Napping axes

The interpretation rules are similar to principal components analysis (PCA) interpretation rules: the coordinates of the variables are the correlations with the principal factors. X6, X8, X9 and X14 (X-axis of nappes of tasters 6,8,9 and 14) have a significant positive correlation with first dimension (*figure 3.4*). It means the configurations of the cavas provided by these tasters on their x-axis is very similar to the configuration of the cavas on the first axis issued from the global analysis (specially concerning the opposition between 2(NA5CHC) and 6(NB5F)). It is not possible to draw general conclusions from this information because each taster has used her/his own criteria to define each axis of its nappe.

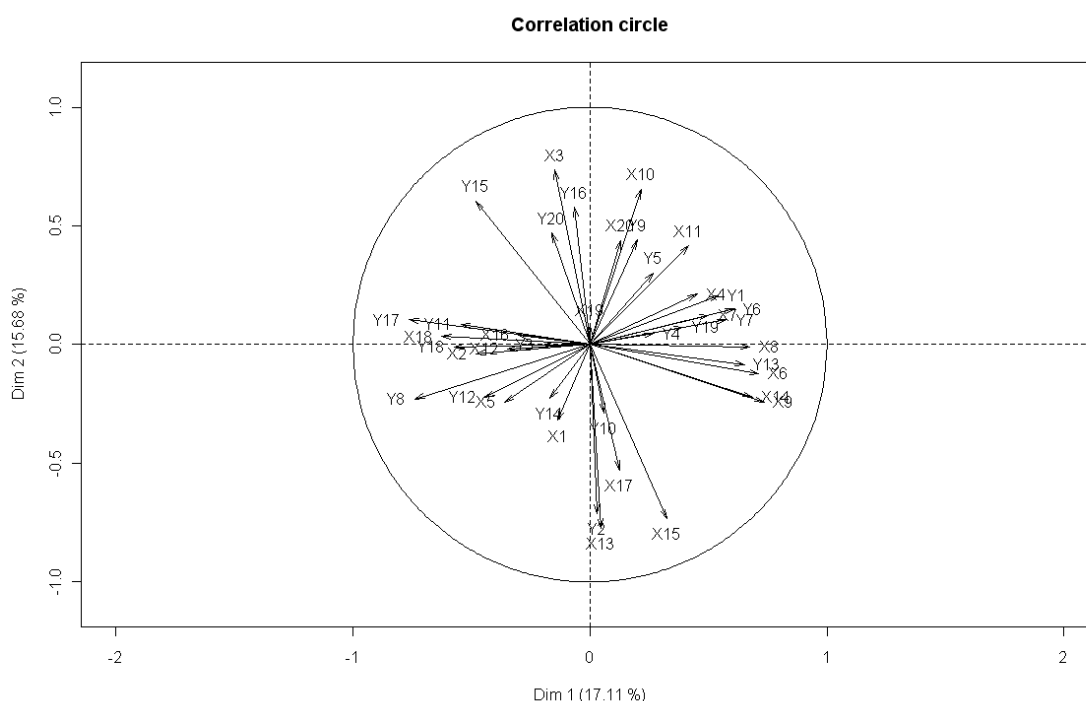


Figure 3.4. Napping axes

3.4.2 Categorizations (free sorting task)

As explained before (chapter 1.2.1) categorization performed by each taster is coded into a categorical variable, whose categories are the sets (*figure 3.5*). Every category is labelled by using the words written to describe it. In HMFA, the categories are represented as in MCA at the centroid of the individuals that present this category. The point category is represented by using the label (*figure 3.5*).

Close to 2(NA5CHC), we find labels mentioning cork defect ("tap", "TCA", "bouchon", "trichloroanisol", "tapón"). 6(NB5F) is clearly defined by its special characteristic, fermentation in barrels ("fusta", "barrica", "boisé"). The oldest cavas are labelled by

Trichloroanisol (TCA): a substance resulting from the degradation of trichlorophenol (or TPA) which in turn comes from the union of phenols cork with dissolved chlorine particles in the air. This degradation occurs in humid environments and is caused by a variety of fungi. The trichloroanisol is responsible for the odor and taste wine cork or cork that was not treated properly during their production or that the bottle has not been maintained under appropriate conditions of temperature and humidity.

terms such as *evolution*, *oxidation*, *aging*, *intense yellow colour* and *toasted* (“oxidación”, “evolución”, “criança”, “color groc intens”, “tostado”).

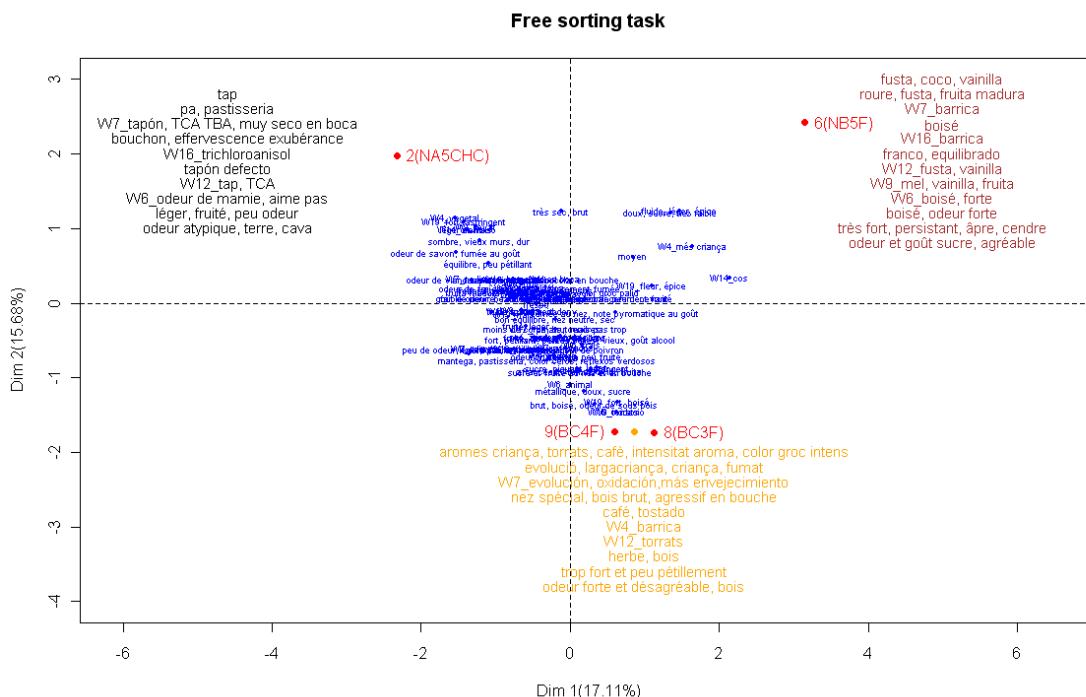


Figure 3.5. Qualitative variables (free sorting task)

3.4.3 Chemical parameters and chromatographic variables

Chemical parameters and chromatographic variables are used as supplementary information. Hereafter, we comment the variables with a high correlation with the HMFA principal axes. Concretely, we have retained the variables that present correlations higher (respectively, smaller) than 0.5 (-0.5) in the case of the chemical parameters and higher (respectively, smaller) than 0.7 (-0.7) in the case of the chromatographic variables (table 3.3).

	Dim 1	Dim 2
<i>Alcoholic Grade</i>	0.717	0.380
<i>Acetil furan</i>	0.841	-0.081
<i>Acetoina (2)</i>	0.712	0.147
<i>Ac. Isovalérico</i>	-0.168	0.702
<i>Alc. 2-feniletilo</i>	-0.026	0.858
<i>g-Decalactona</i>	0.252	0.848
<i>Acetato etilo (2)</i>	0.491	0.713
<i>Ac. Caproico</i>	0.109	-0.706
<i>Ac. Caprílico</i>	-0.003	-0.726
<i>D.O 420nm</i>	0.428	-0.676
<i>Glycerol</i>	0.256	0.552

Table 3.3. Correlations of chemical and chromatographic variables with first and second dimension

We note that the supplementary variables seem to be more related with the second axis than with the first. This result was expected because the first axis is very particular

built up from only special cavas. Thus, only the supplementary variables for which these special cavas present particular values have an important correlation with the first axis.

"Alcohol Grade" is highly correlated with the first bisector, due to the fact that 6(NB5F) has the highest alcohol grade among all cavas (figure 3.6). The chromatographic compounds "Acetoina" and "Acetil Furan" are highly correlated with the first axis because of 6(NB5F) presenting high values for these variables. Chemical parameter "Glycerol" and chromatographic compounds "Alc. 2-feniletilo", "Ac. Isovalérico", "g-Decalactona" and "Acetato etilo" are highly correlated with the second axis due to the fact that 8(BC3F) and 9(BC4F) have the smallest values for these variables. "D.O. 420nm" highly correlated with the second bisector and the third axis ($r=-0.6760$) because of its relationship with aging.

The high correlation between the third axis and "Total Sugar" due to the high contribution of the youngest bruts (1(BA6CHC), 3(BA6C) and 10(BC5F)) which are the cavas with the highest value of total sugar.

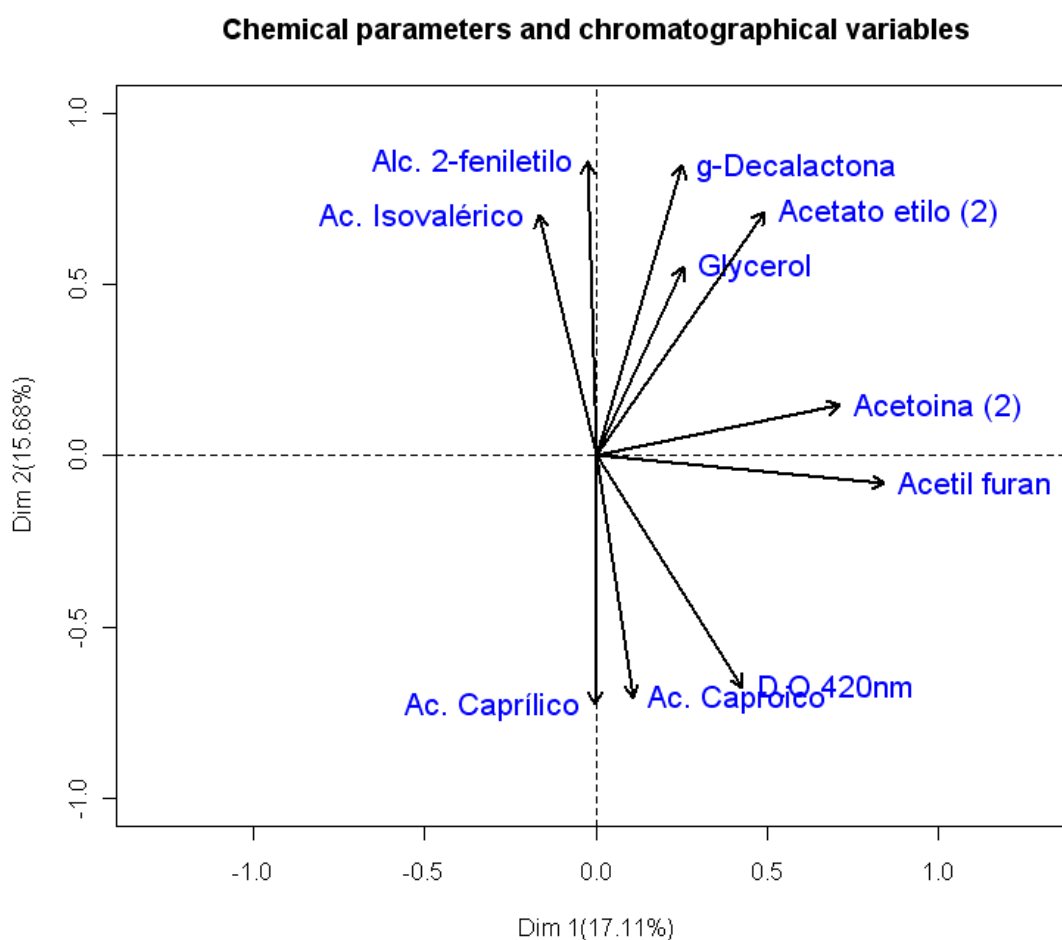


Figure 3.6. Chemical parameters and chromatographical variables

3.4.4 Words

The words columns are used as supplementary information and projected on the principal axes as supplementary frequency columns. The words that are more used to describe on cava are situated close to it. "cork" (tap) and "TCA" appear close to 2(NA5CHC) and "vanilla" close to 6(NB5F). Words as "Barrel" (barrica) and "wood" (fusta), which describe the most important characteristic of 6(NB5F), are not so close to this cava because they are also employed to define other cavas (*figure 3.7*). The oldest cavas are very related with "oxidation" (oxidació), "evolution" (evolució), "aroma" and "toasted" (torrats). On the third axis, appear words "young", "fresh", "fruit" and "floral" which are very related with young cavas and words "oxidation", "evolution", "aging" and "toasted" very related with the oldest cavas.

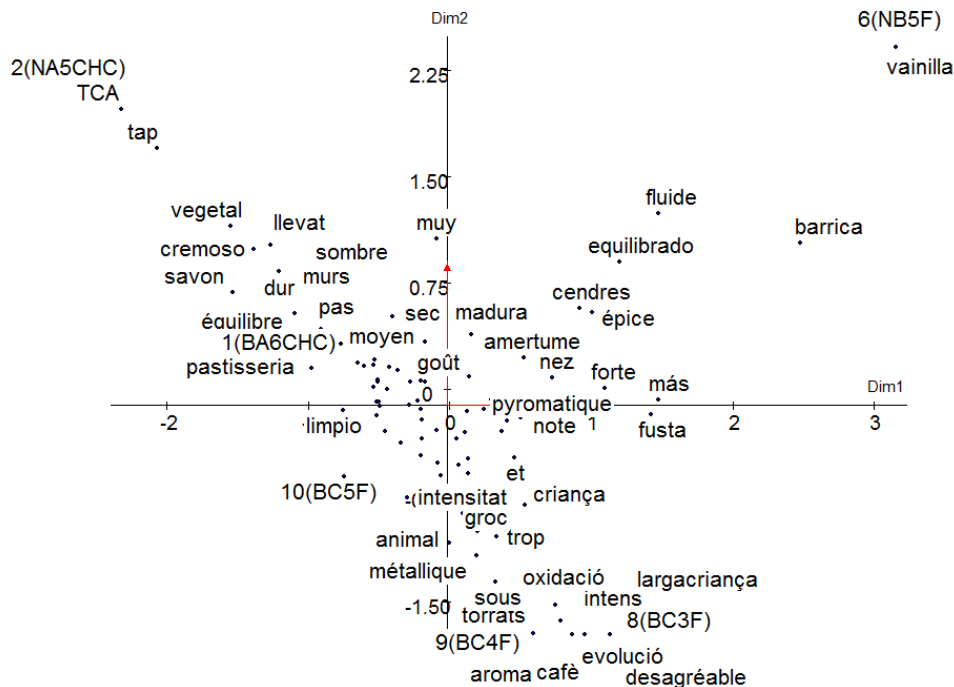


Figure 3.7. Configuration of words

3.5 Representation of the sets

In the case of HMFA, the sets can be represented at every level of the hierarchy.

3.5.1 First hierarchical level

There is a clear separation between napping and free sorting task sets (*figure 3.8*). In fact, napping consists of quantitative variables while free sorting task columns are qualitative columns. In this case, the qualitative variables –which are more simple– appear to be more related with the global analysis dimensions than the quantitative variables are. There are only little differences between students and experts napping, on the one hand, and between students and experts free sorting task, on the other hand: at thus level, the points representing these sets are mixed quite a lot between

them. Only the point “free sorting task-taster 1” (FST1), who is a student, lies far from the others; it shows few relationship with the global analysis dimensions (consulting the data, we can see that the free sorting task of this student consists only in four clusters and, furthermore, one cluster includes 6 cavas, the highest number of cavas in a cluster among all the free sorting task data).

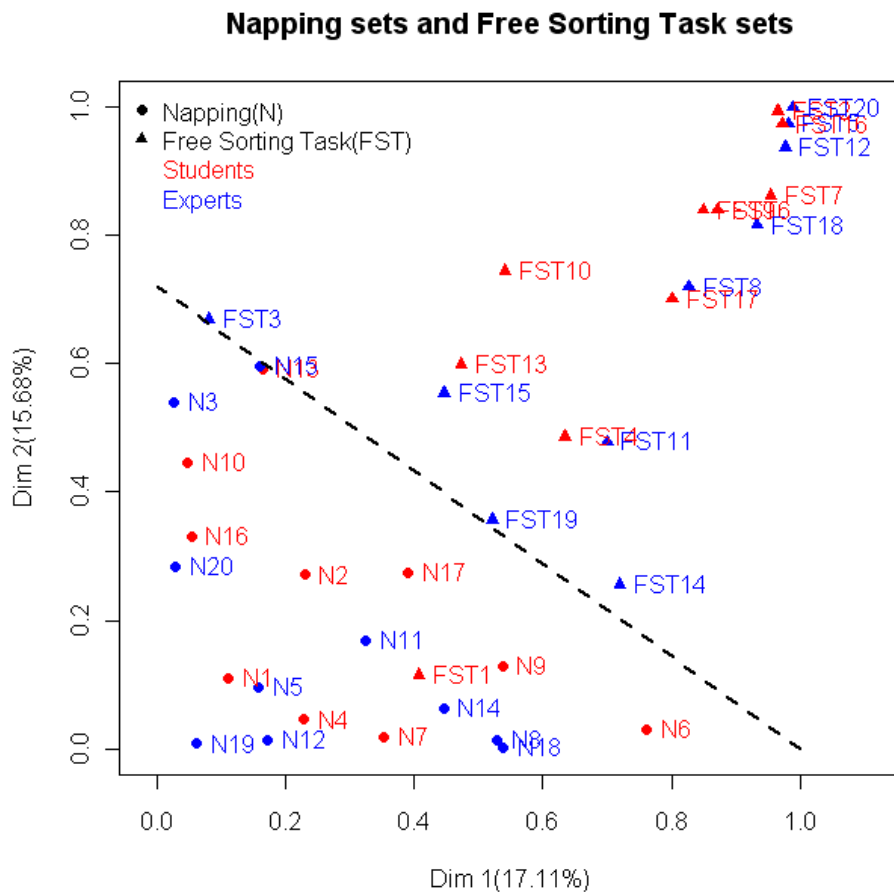


Figure 3.8. Sets representation (napping and free sorting task)

3.5.2 Second hierarchical level

The experts are globally more related with the first dimension of the global analysis than students. There are not important differences between students and experts on the second dimension (*figure 3.9*). Generally, experts are more much closer between them than students. Only two experts (E4 and E14) are situated a little far from the other experts because of a low relationship with the second dimension of the global analysis. On the other hand, students form small subgroups, some of them (S6, S8 and S17) close to experts and other (for example, S1, S19 and S3) far from the experts.

3.5.3 Third hierarchical level

At the third level, the global representation of students and experts are similar. The third level students point summarizes well the global analysis.

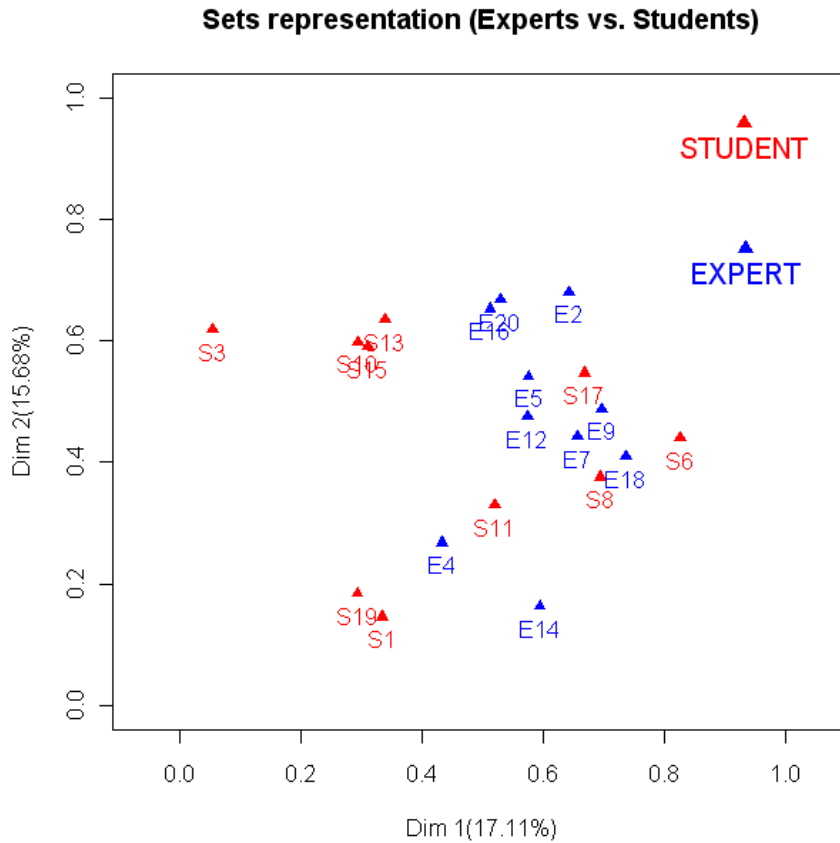


Figure 3.9. Sets representation (experts vs. students)

3.6 Cluster

Clustering of the cavas is performed from their coordinates on the first four principal axes of HMFA. A hierarchical algorithm is used. Ward generalized criterion to compute the proximity between individuals or nodes. The partition into 6 clusters is retained. Three of these clusters consist in only one cava (*figure 3.10*).

Every cluster is described by its characteristics (words and variables). The characteristics words are significantly overused to describe the cavas of the clusters (Lebbart et al., 2000). Characteristic quantitative variables, chemical and chromatographic, are those that have a mean within the cluster significantly different, from the global mean as computed on all the cavas.

The partition summarizes the information provided by HMFA graphics. The cluster composed of 6(*NB5F*) is described through words such as “vanilla”, “wood” and “barrel”. Cluster 5, composed of 2(*NA5CHC*), is described by “cork” and “TCA”. We observe that the youngest cavas are gathered into on cluster (cluster 1) as well as the oldest cavas (cluster 4).

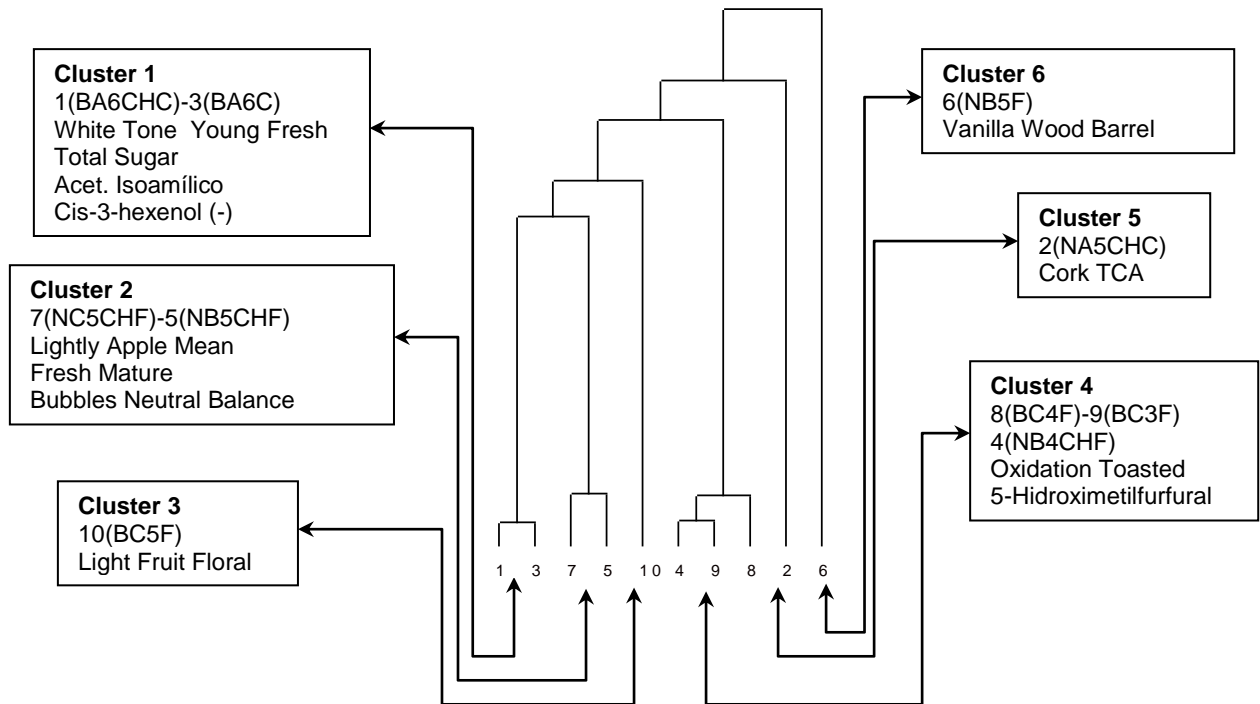


Figure 3.10. Hierarchical Clustering (partition at 6 clusters)

3.7 Conclusions of HMFA

The most important conclusions drawn from HMFA are

- The first axis of the global analysis (HMFA) opposes two special cavas: 2(NA5CHC) with cork defect (“cork” and “TCA”) and 6(NB5F) fermented in barrel (“vanilla”, “wood” and “barrel”).
- The third axis is very related with aging. The youngest and the oldest cavas are opposed on this axis. The youngest are described as “fresh”, “floral”, “fruit”, “young”, “white” while the oldest cavas are associated to “oxidation”, “evolution”, “aging”, “toasted” and “cafe”.
- Aging is confused with wood (some descriptions of the oldest cavas include “wood”).
- Free sorting task results are more much related with the components of the global analysis than napping (first hierarchical level).
- Free sorting task of the first taster, a student, is very different from the free sorting task of the other tasters.
- Individually, experts are more much related between them than students. They are more much related with the first dimension of global analysis than the students (second hierarchical level).

- Globally, students and experts lie very close on the first factorial plane of the global analysis. So partial results of students and experts are similar (third hierarchical level).

The global analysis results could lead to conclude that the configurations of cavas provided by the students and by the experts are similar. However, we can observe that the representation of the tasters as sets at the second level—students and experts—are not totally similar. Experts appear to be closest one to another (indicating consensual judgement) while the students present a higher dispersion (*figure 3.9*). This shows that there are differences between both students and experts panels that are somewhat masked in the HMFA results.

We explore this question in the following chapters, looking for answering the one of the initial objectives of this project. We will analyze separately napping data and free sorting task of students and experts.

Chapter 4

Separate analyses

In this chapter, the results of the free sorting tasks –analysed via multiple correspondence analysis (MCA)– and the results of the napping –analysed via multiple factor analysis (MFA)– are presented. In both cases, students and experts configurations are separately analysed.

4.1 Free sorting task

In this section, we separately analyse the free-sorting tasks performed either by the students or the experts. The results are presented in a parallel way.

4.1.1 Eigenvalues

Students

The percentage of variance explained by the first four axes only slowly decreases (table 4.1). The first factorial plane (first and second dimensions) explains only one third part of all variance; four dimensions are needed to keep more than 50% of the total variance.

Experts

The percentage of variance explained by the first factorial plane is very similar to the percentage of variance explained by the first factorial plane of students (35.05% vs. 32.24%). The variance explained by the third and fourth axes is slightly higher than the homologous axes issued from MCA applied to students free sorting task.

	STUDENTS				EXPERTS		
	Eigenvalue	Percentage of variance	Cumulative percentage of variance		Eigenvalue	Percentage of variance	Cumulative percentage of variance
dim 1	0.625	16.443	16.443	dim 1	0.916	18.319	18.319
dim 2	0.600	15.800	32.243	dim 2	0.844	16.886	35.205
dim 3	0.486	12.799	45.042	dim 3	0.753	15.068	50.273
dim 4	0.444	11.695	56.737	dim 4	0.705	14.109	64.382
dim 5	0.424	11.165	67.902	dim 5	0.550	11.004	75.386
dim 6	0.377	9.910	77.812	dim 6	0.441	8.827	84.213
dim 7	0.356	9.377	87.189	dim 7	0.335	6.691	90.904
dim 8	0.307	8.082	95.271	dim 8	0.276	5.528	96.432
dim 9	0.180	4.729	100	dim 9	0.178	3.568	100

Table 4.1. Eigenvalues, percentage of variance and cumulative percentage of variance

4.1.2 Configuration of cavas and description of clusters

Two individuals are close one another in MCA plane when they are frequently assigned to the same categories. In the case of this study, two cavas are close when they are assigned to the same cluster by many tasters; they are far if they belong to the same clusters only in rare occasions.

The clusters, identified through the words that describe them, are situated at the centroid of the cavas which belong to them.

Students

The first axis opposes two special cavas, 2(*NA5CHC*) and 6(*NB5F*). Jointly, their contributions to the first axis are higher than 70% of the total cavas contributions (table 4.2). Four students made a single cluster with 6(*NB5F*) and three students made a single cluster with 2(*NA5CHC*), which causes that these cavas are put to the fore, as different from the others.

The second axis opposes these two special cavas to 8(*BC3F*) and 10(*BC5F*). 8(*BC3F*) and 10(*BC5F*) are *brut* and the former is the oldest cava (figure 4.1).

	STUDENTS					EXPERTS				
	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
1(BA6CHC)	3.21	3.05	0.18	49.36	12.51	3.16	<0.01	<0.01	38.40	8.71
2(NA5CHC)	41.14	12.96	0.41	14.18	6.89	16.11	59.48	7.43	4.11	2.68
3(BA6C)	2.03	0.09	3.29	7.33	5.15	3.13	0.03	0.96	12.05	13.69
4(NB4CHF)	0.19	1.22	<0.01	17.74	13.01	0.42	3.53	0.03	1.37	7.71
5(NB5CHF)	7.86	1.46	54.57	0.25	14.03	3.05	0.42	18.69	14.53	10.00
6(NB5F)	29.55	38.53	8.92	0.04	6.67	67.80	16.09	4.79	0.71	0.03
7(NC5CHF)	4.36	1.93	5.62	2.76	5.90	1.01	3.80	13.56	14.92	<0.01
8(BC3F)	1.74	16.85	0.47	0.75	0.43	3.92	6.72	39.19	4.42	2.41
9(BC4F)	4.39	8.26	3.88	0.30	8.63	0.27	6.22	12.83	0.32	0.08
10(BC5F)	5.52	15.64	32.30	21.12	26.79	1.13	3.72	2.52	9.18	54.70

Table 4.2. Contributions of the cavas to the first five dimensions (<0.01: contribution smaller than 0.01%)

The descriptions of the clusters made by four students who put 6(*NB5F*) describe this cava as *strong* (*forte*) and *wood* (*boisé*). In fact, this cava is barrel-aged. The students who put 2(*NA5CHC*) in a single cluster describe it as *old smell* (*odeur de mamie*) and *atypical smell* (*odeur atypique*). It can be thought that these students have noticed the cork defect of this cava, but without identifying the reason. 8(*BC3F*) and 10(*BC5F*) were defined from their *sparkling characteristics* (*pétillant, pétillément*). However, some students have noted a lack of sparkling character.

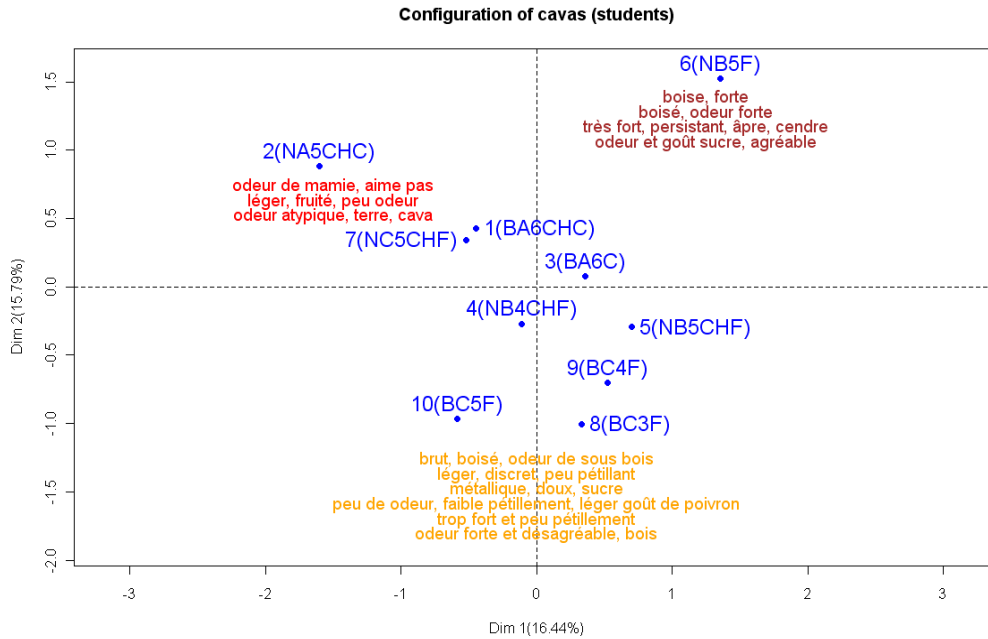


Figure 4.1. Configuration of cavas (students)

Experts

The first two axes issued from MCA are built by only two cavas, 2(NA5CHC) and 6(NB5F). Their joint contribution is greater than 75% for each of both axes.

Eight out of ten experts made a single cluster with 6(NB5F) and seven made a single cluster with 2(NA5CHC). Thus, more much experts than students isolated 2(NA5CHC) or 6(NB5F) (8 vs. 4 for 6(NB5F) and 7 vs. 3 for 2(NA5CHC)).

The other cavas are ranked on the second bisector according to their production year. The youngest cavas are situated on the top part of this bisector and the oldest are on the bottom part (figure 4.2).

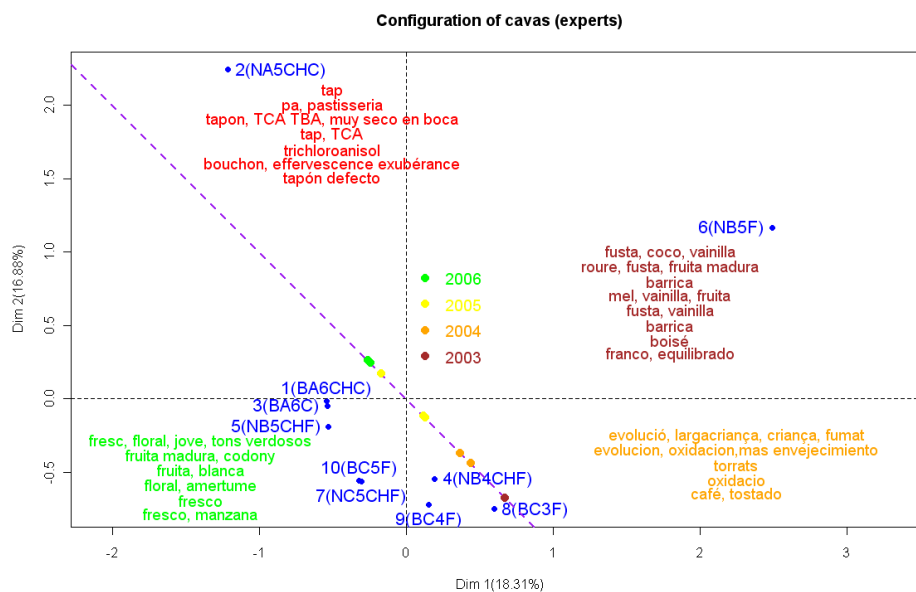


Figure 4.2. Configuration of the cavas (experts)

6(NB5F) is described with words such as *wood* (fusta) due to the special elaboration of 6(NB5F), barrel-aged cava, clearly identified by the experts. Concerning 2(NA5CHC) description, two words are frequently repeated: *cork* (tap, tapón) and *tricoloroanisol*. The special elaboration and the cork defect seem to be the two main criteria for experts to isolate these two cavas into single clusters.

Close to the youngest cavas we frequently find *fresh* (fresc, fresco), *fruit* (fruita), and *floral*. The oldest cavas are related with *oxidation* (oxidació, oxidación), *evolution* (evolució, evolució), *toasting* (torrats, tostado) and *aging* (envejecimiento).

4.1.3 Conclusions from free-sorting task results

2(NA5CHC) and 6(NB5F) are very different one another and from the other cavas:

The most of the experts considered 2(NA5CHC) and 6(NB5F) clearly different from the other cavas. Eight experts made a single cluster with 6(NB5F) and seven experts with 2(NA5CHC). In the case of the students, this opinion is not so shared: only four students made a single cluster with 6(NB5F) and three students with 2(NA5CHC).

Differentiation criteria: The descriptions of the clusters indicate that the experts who putted 2(NA5CHC) and 6(NB5F) into single clusters are able to give the reason. In this case, they are related with the characteristics of the cavas (barrel-aged or cork defect). In the case of the students, those who put these special cavas into single clusters noted their difference but they were not able to identify the reason.

Bisector related with production year: In the case of the experts, the second bisector of the first plane ranks the cavas according to their year.

Confusion between ageing and wood: Some students confused ageing effect with wood. So, they described some old cavas by using the word *wood*.

4.2 Napping

In this section, we separately analyse the nappings performed either by the students or the experts. The results are presented in a parallel way.

4.2.1 Data structure and analysis method

MFA is a weighted PCA. Each variable is weighted by the inverse of the first principal component inertia computed in the separate PCA applied to its subgroup (chapter 2.3). This reweighting induces balanced contributions of all of the tasters to the first factorial axis.

MFA allows for answering to the following important questions:

1. Which cavas are similar and which are different according to the information collected by napping?

II. Which are the most important differences and similarities between students and experts opinions?

Each taster j (student or expert) generated a *nappe* (tablecloth). The configuration provided by one *nappe* is coded through the coordinates of every cava on the x-axis and y-axis.

So, napping provides two data sets, one for students and other for experts. Every set is composed of twenty columns (x-axis column and y-axis column of the 10 nappes) and 10 rows (10 cavas). Chemical parameters (variables as alcoholic grade, ph, total acidity, etc...), chromatographic data and the words frequency table were added to napping data as supplementary information (figure 4.3).

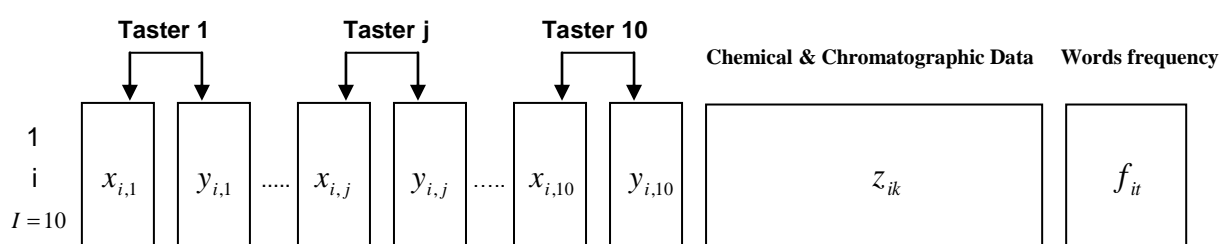


Figure 4.3. Data structure. x_{ij} is x-axis of napping, y_{ij} is y-axis of napping and supplementary information (chemical, chromatographic and words frequency)

4.2.2 Individual nappes

The structure of the individual nappes (1 nappe = 1 set) can be disclosed by performing separate analyses of each of them (non-standardized PCA for each set). In particular, the proportions of variance explained by two axes shows the ability of the tasters for using or not both dimensions

In the case of the experts, the results are very similar. The minimum of variability explained by the first axis is 65.2% (E16) and the maximum is 80.3% (E14). Most of this variability explained ranks between 70% and 80% (table 4.3).

In the case of the students, more variability is observed, from 59% (S6) and 90% (S8).

An explained variability by the first axis about 50% means that this taster uses both dimensions in the same way to discriminate the cavas while an explained variability by the first axis about 90% means that the taster uses only one dimension to discriminate the cavas (the differences between cavas are only reflected on one dimension).

	Eigenvalues S1		Eigenvalues S3		Eigenvalues S6		Eigenvalues S8		Eigenvalues S10	
Dim1	377.172	79.48%	637.421	89.74%	203.180	59.00%	282.038	90.00%	306.649	75.52%
Dim2	97.340	20.52%	72.883	10.26%	141.162	41.00%	31.403	10.00%	99.396	24.48%

	Eigenvalues S11		Eigenvalues S13		Eigenvalues S15		Eigenvalues S17		Eigenvalues S19	
Dim1	352.57	66.95%	287.038	73.25%	264.403	83.67%	204.371	62.40%	281.556	82.35%
Dim2	174.09	33.05%	104.807	26.75%	51.597	16.33%	123.186	37.60%	60.326	17.65%

	Eigenvalues E2		Eigenvalues E4		Eigenvalues E5		Eigenvalues E7		Eigenvalues E9	
Dim1	301.425	65.37%	257.697	75.65%	321.075	70.83%	246.822	76.63%	296.305	74.61%
Dim2	159.650	34.63%	82.945	24.35%	132.205	29.17%	75.292	23.37%	100.825	25.39%

	Eigenvalues E12		Eigenvalues E14		Eigenvalues E16		Eigenvalues E18		Eigenvalues E20	
Dim1	206.688	79.37%	308.068	80.30%	91.069	65.20%	299.016	68.41%	329.454	70.48%
Dim2	53.732	20.63%	75.582	19.70	48.596	34.80%	138.062	31.59%	137.986	29.52%

Table 4.3. Eigenvalues of separated group analysis (PCA). Students (S) and Experts (E)

4.2.3 Eigenvalues structure in students and experts sets

The global analysis performed by MFA on the students napping provide a first plane that explains 44.78% of the variability, while, the first plane issued from the experts napping explains 49.80% of the variability. In both cases, it is possible to explain about 75% of the total variability with the first four axes of MFA (*table 4.4*).

	MFA GLOBAL (Students)				MFA GLOBAL (Experts)		
	Eigenvalue	Percent	Cumulative Percent		Eigenvalue	Percent	Cumulative Percent
Dim 1	3.165	23.663	23.663	Dim 1	3.709	26.824	26.824
Dim 2	2.824	21.117	44.781	Dim 2	3.178	22.981	49.805
Dim 3	2.364	17.678	62.458	Dim 3	1.957	14.152	63.957
Dim 4	1.614	12.066	74.524	Dim 4	1.451	10.493	74.450
Dim 5	1.469	10.985	85.509	Dim 5	1.307	9.449	83.899
Dim 6	0.899	6.723	92.232	Dim 6	0.912	6.597	90.496
Dim 7	0.467	3.493	95.725	Dim 7	0.554	4.003	94.499
Dim 8	0.358	2.677	98.402	Dim 8	0.403	2.912	97.411
Dim 9	0.214	1.598	100	Dim 9	0.358	2.589	100

Table 4.4. Eigenvalues of global analysis for students and for experts

4.2.4 Configurations of cavas

Students

The cavas with the highest contributions to the first and second axis are 2(*NA5CHC*), 6(*NB5F*) and 9(*BC4F*) (*table 4.5*). The first axis opposes *brut* to *nature* cavas with exception of 4(*NB4CHF*), which is a *nature* cava lying close to *brut* cavas (*figure 4.4*). The second axis opposes 2(*NA5CHC*) to 6(*NB5F*). 6(*NB5F*) is the barrel-aged cava while 2(*NA5CHC*) presents a cork defect, also called TCA. So the second axis can be defined as a particular characteristics dimension.

Experts

Compared to students, the experts underline the specificity of 6(*NB5F*) did not consider 2(*NA5CHC*) so different from the others. The second bisector is more interesting than the first dimensions because cavas are ranked on the second bisector depending on the production year.

	STUDENTS					EXPERTS				
	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
1(BA6CHC)	1.39	1.66	14.58	4.14	0.18	4.08	15.36	20.85	8.12	1.65
2(NA5CHC)	18.16	35.16	1.11	14.72	0.73	0.53	6.19	7.27	1.93	52.31
3(BA6C)	2.49	8.46	0.62	39.14	8.38	7.59	17.80	3.12	25.73	3.43
4(NB4CHF)	6.12	0.09	10.22	0.60	0.28	2.90	4.02	19.79	16.83	8.05
5(NB5CHF)	7.21	4.13	4.83	0.82	36.88	13.85	11.35	0.09	1.69	3.27
6(NB5F)	27.16	30.20	1.60	2.61	13.11	29.94	24.43	3.41	0.01	0.45
7(NC5CHF)	3.22	3.24	10.35	3.03	12.98	5.16	2.38	6.70	0.05	8.15
8(BC3F)	1.70	13.55	24.34	13.71	9.66	28.57	6.60	4.08	13.29	0.04
9(BC4F)	28.02	3.00	0.05	0.12	17.79	1.80	8.55	9.75	17.36	22.52
10(BC5F)	4.523	0.51	32.30	21.12	0.01	5.58	3.31	24.95	14.98	0.13

Table 4.5. Contributions of cavas

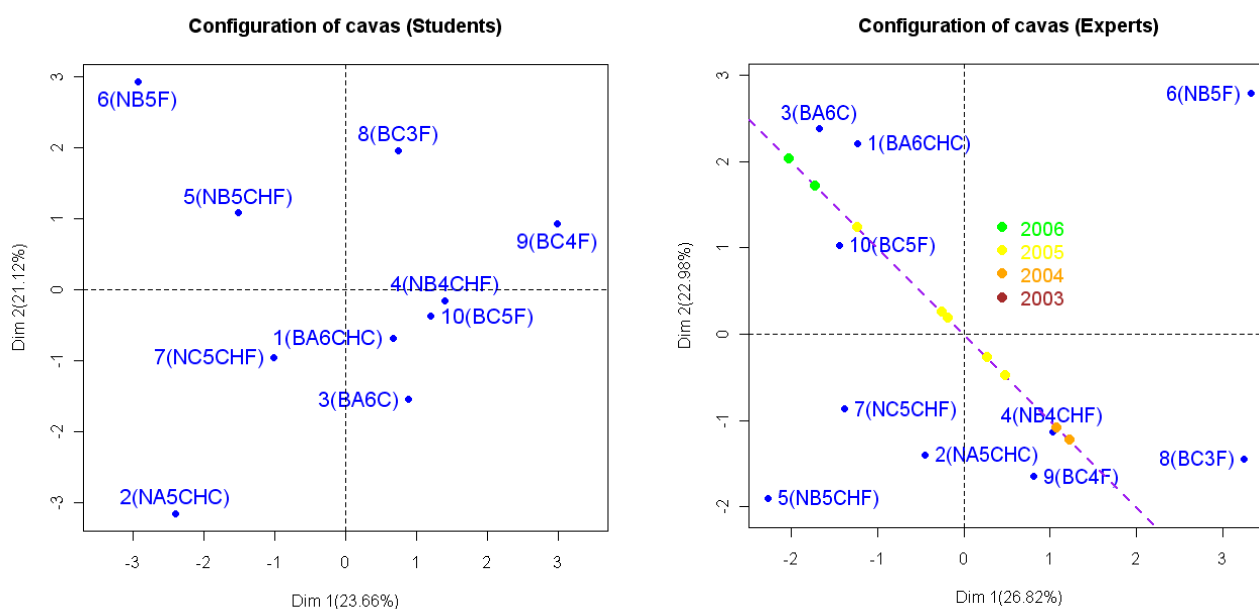


Figure 4.4. Configurations of cavas

4.2.9 Supplementary elements: chemical parameters

Students

“Total Sugar” has a high correlation with the first axis because of the separation between brut and nature cavas on this axis (brut cavas have more sugar than nature ones). “Alcohol Grade” is correlated with the second bisector (*6(NB5F)* has the highest alcohol grade). “Glycerol” and “Malic Acid” have high correlations with the first axis and “D.O. 420nm” with the first bisector. Remember that D.O. 420 nm measures optical density of yellow and when it is higher it means cava is more yellow, more evolved and oxidized too.

Experts

“D.O. 420nm” presents a high correlation with the second bisector because of the ranking of the cavas on this bisector depending on their production year (this variable is very related with ageing). “Glycerol” and “Alcohol Grade” have high correlations with the first bisector (*6(NB5F)* has the highest alcohol grade and glycerol). We also note the high correlation of “Dry Extract” with the first axis.

	STUDENTS					EXPERTS				
	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
Alcoholic Grade	-0.267	0.462	-0.122	0.113	0.667	0.647	0.369	0.397	0.225	-0.100
PH	0.354	0.248	-0.036	0.098	0.586	0.096	0.003	0.497	0.345	-0.580
Total Acidity	0.028	-0.134	-0.007	-0.082	-0.233	0.117	-0.078	-0.635	-0.087	0.363
F. SO2	0.031	-0.407	0.081	0.033	-0.102	-0.201	-0.147	-0.867	0.052	0.250
T. SO2	0.105	-0.165	-0.056	-0.131	0.150	0.176	-0.050	-0.696	0.023	0.143
Total Sugar	0.613	-0.017	0.201	0.428	-0.262	-0.114	0.440	-0.526	0.243	-0.171
D.O. 420nm	0.453	0.512	-0.506	-0.231	0.217	0.524	-0.519	-0.021	0.123	-0.530
Malic A.	0.466	0.332	-0.034	-0.057	0.513	0.291	-0.287	0.256	0.491	-0.410
Lactic A.	0.132	-0.389	-0.269	0.219	-0.335	-0.136	0.182	-0.362	0.179	0.350
Glycerol	-0.361	0.033	-0.085	0.030	0.256	0.452	0.375	0.086	0.107	0.501
Dry Extract	0.184	0.140	-0.357	-0.076	0.763	0.653	0.076	0.161	0.452	0.058

Table 4.6. Correlations of chemical parameters with global analysis dimensions

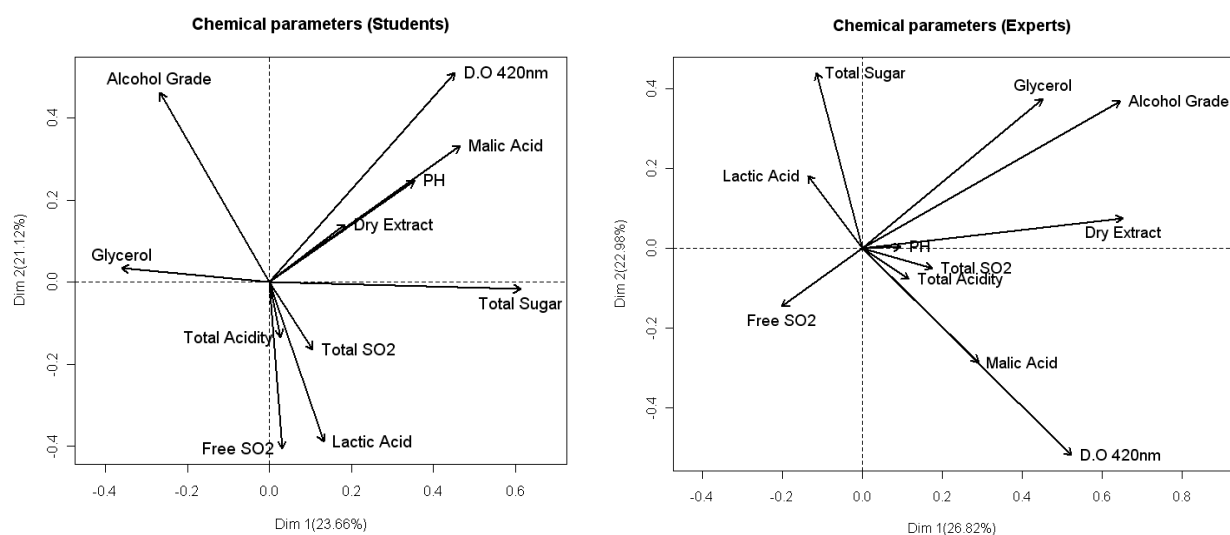


Figure 4.5. Chemical parameters from their covariances with the first two global axes issued from MFA

4.2.9 Supplementary elements: chromatographic variables

Students

Acetil furan and Furfural indicate aging (they increase with aging time). We previously observed that the oldest cavas were up and the youngest cavas were down on the second global axis (*figure 4.4*). So interpretation of Acetil furan and Furfural agrees with results of analysis. Furfural also is related with barrel. So it has a high correlation with second axis where 6(NB5F) has an important contribution on this axis.

Experts

Acet. Isoamílico is a typical aroma of a fruity fermentation and it gives banana taste. It decreases with aging time and so normally only young cavas have high concentrations of Acet. Isoamílico (it has a great correlation with second bisector where cavas were ordered by means of their years). Acetil furan, furfural and hidroximetilfurfural are related with aging (high correlation with second bisector) and succinato dietil is related with barrel.

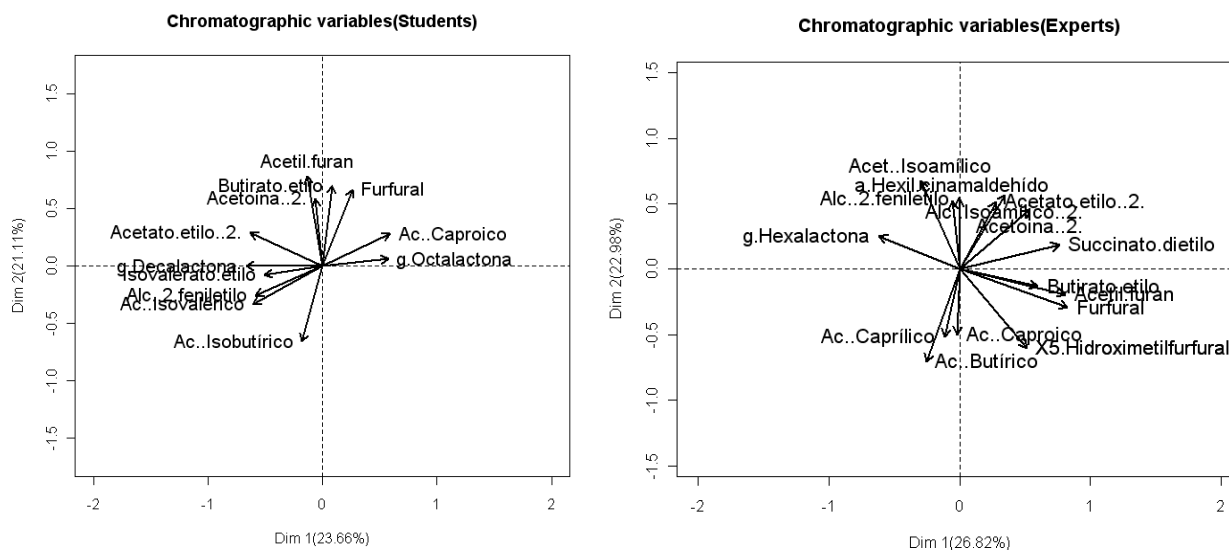


Figure 4.6. Chromatographic variables from their covariances with the first two global axes issued from MFA

4.2.9 Supplementary elements: words

Students

The oldest cavas are defined with negative words as *hot spices* (piquant), *metallic* (métallique), *animal* and *disagreeable* (désagréable). Some students confused ageing perception with *wood* (bois, boisé). There is not any characteristic word concerning 2(NA5CHC) and 6(NB5F).

cavas, is a common factor for sets S6, S8 and S17. In other words, the difference between special cavas and the rest is very well represented on nappes of students S6, S8 and S17.

The first principal component of global analysis of MFA for experts is a common factor for sets E4, E9 and E14. Second principal component is a common factor for E9, E12 and E20.

A global measurement to define similar structures between different sets is inertia ratio. If there is a common structure between sets, then the points which represent the same individual at different sets are near among them (low inertia-intra). Ratio is calculated as inertia-inter/inertia total. If it is near to 1, the principal component represents a common structure for all sets.

The highest ratio of all dimensions for students and experts is 0.41 (first dimension of experts). So ratios are low and there are not common structures for the most of the sets (table 4.8).

INERTIA INTER/INERTIA TOTAL (students)						INERTIA INTER/INERTIA TOTAL (experts)					
DIMENSION 1 TO 5						DIMENSION 1 TO 5					
FAC.	1	2	3	4	5	FAC.	1	2	3	4	5
	0.33	0.33	0.25	0.19	0.19		0.41	0.37	0.23	0.17	0.20

Table 4.8. Ratio (Inertia-inter/Inertia-total)

4.2.9 Representation of sets

The coordinate of a set on each axis of the global analysis is interpreted as the accumulated inertia of the variables of the set on this axis. The reweighting of the variables makes that the coordinates of the sets vary between 0 and 1. Higher is the coordinate, more the set is related with the corresponding axis.

	STUDENTS						EXPERTS				
	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5		Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
S1	0.448	0.212	0.004	0.356	0.035	E2	0.286	0.298	0.359	0.083	0.336
S3	0.340	0.078	0.096	0.008	0.351	E4	0.698	0.007	0.161	0.026	0.165
S6	0.145	0.735	0.160	0.072	0.222	E5	0.328	0.472	0.315	0.036	0.125
S8	0.001	0.669	0.223	0.039	0.068	E7	0.041	0.171	0.284	0.568	0.091
S10	0.426	0.030	0.027	0.420	0.169	E9	0.780	0.380	0.021	0.075	0.039
S11	0.442	0.264	0.353	0.250	0.015	E12	0.159	0.784	0.060	0.002	0.050
S13	0.660	0.133	0.168	0.104	0.046	E14	0.882	0.023	0.027	0.028	0.134
S15	0.448	0.282	0.169	0.057	0.113	E16	0.229	0.306	0.195	0.264	0.034
S17	0.196	0.411	0.352	0.235	0.300	E18	0.284	0.125	0.462	0.282	0.124
S19	0.059	0.009	0.812	0.072	0.149	E20	0.022	0.613	0.072	0.086	0.209
Che*	0.024	0.045	0.005	0.022	0.025	Che*	0.034	0.012	0.544	0.003	0.028
Chr**	0.193	0.216	0.076	0.069	0.227	Chr**	0.100	0.276	0.222	0.034	0.304

Table 4.9. Sets coordinates (*:Chemical, **: Chromatographic)

Students

The relationships between the sets and the first two dimensions of the global analysis are low (*table 4.9*). Only S13 has a significant relationship with the first dimension as well as S6 and S8 with the second. Chemical parameters and chromatographic variables are not related with the first two dimensions of the global analysis. S19 is totally different from the global analysis for the first factorial plane (its coordinates are very low on these axes).

Experts

There are more experts than students who have a significant relationship with the first dimensions of the global analysis. E4, E9 and E14 present a high relationship with the first dimension and E12 and E20 with the second. Chemical parameters and chromatographic variables are not related with the dimensions of the global analysis.

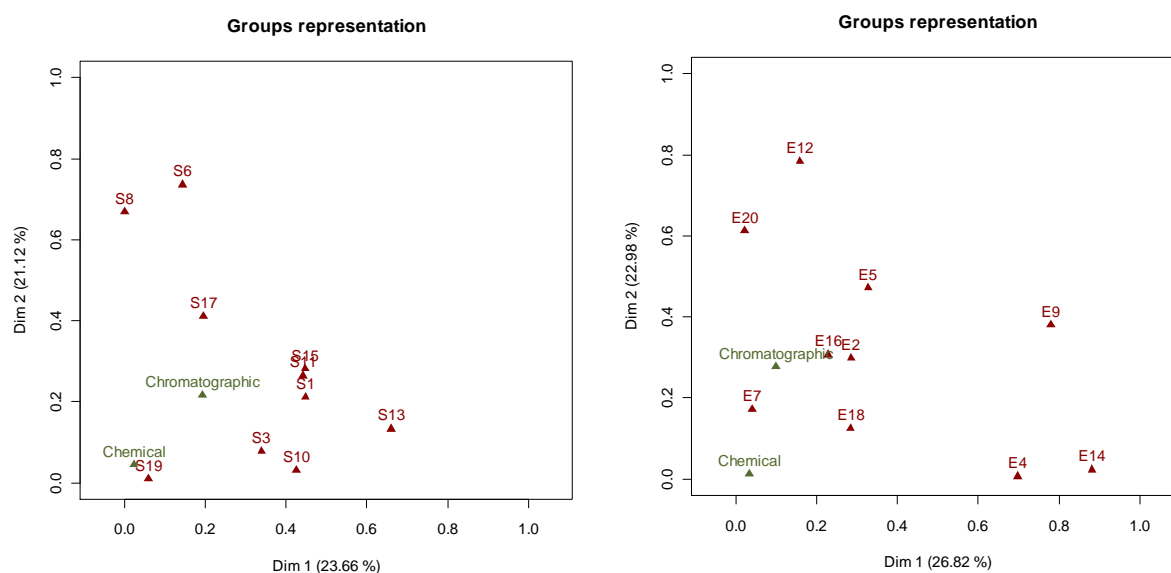


Figure 4.8. Sets representation

4.2.10 Conclusions from napping results

Special cavas: *2(NA5CHC)* and *6(NB5F)* appear as very specific cavas, with a high contributions on the first axes in the case of the students. In the case of the experts, *2(NA5CHC)* is not perceived to be very different from the others. In fact, the experts are sensible to the cork defect, as they show in their free sorting task, but they try to see the other characteristics to place it among the others.

Brut versus nature cavas: In the case of the students, the first axis of the global analysis opposes brut and nature cavas with the exception of *4(NB4CHF)*, a nature closes to brut cavas. So the students use sugar as a criterion.

Bisector related with year: As in MCA, the second bisector of the global analysis for experts ranks the cavas according to their year production, which reveals itself as a criterion for the experts.

Greater consensus among experts: Descriptions and words used by the experts to characterize the clusters of cavas show a higher consensus than in the case of students. The descriptions are also more precise and more related to the real characteristics of the cavas in experts as compared to students.

D.O. 420nm and total sugar: *D.O. 420nm* is highly related with the production year of the cavas. So, this variable has a high correlation with the second bisector of global analysis in the case of the experts. In a similar way *Total Sugar* has a high correlation with the first axis of the global analysis in the case of the students because this axis separates brut and nature cavas.

Students and experts are different: There are important differences between nappes of students and experts such as

- Although both experts and students note the cork defect of 2(*NA5CHC*) –as shown by free sorting task– the experts take into account its other characteristics to place it among the other cavas
- The experts rank the cavas according to their year production.
- In the case of the experts, the global axes are more much related with the vocabulary use describe the clusters than in the case of the students.

We can say that there are some important differences between students and experts, concerning the perception of sensorial attributes (for example, sweetness versus ageing) and also the description of these attributes, richer and more precise in the case of the experts.

Chapter 5

Consensus among the panellists

In this chapter, we quantify the consensus level by comparing napping of panels and random generate napping (section 5.1). We also look for measuring the similarities between the panellists, tasters, either in the case of the students or in the case of the experts. Clustering allows for summarizing the proximities between panellists (section 5.2), we also cluster the panellists.

5.1 Quantification of the consensus level

The comparison of the separate analyses led to conclude that there is a higher consensus between experts than students. *“Is it possible to quantify the consensus level?”*. To answer the question, we compare the results of both panels, separately with randomness.

5.2.1 Statistical Test

The hypotheses are:

H_0 : The panel works at random (no consensus)

H_1 : The panel does not work at random (consensus)

For this contrast, we use as statistic the first eigenvalues issued from MFA of panels of 10 panellists. Higher is the first eigenvalue more consensus exists between the tasters. Thus, this statistic can be considered as a consensus measure. We build the reference distribution of this statistic under H_0 as explained hereafter.

5.2.1 Random panel

We have generated 1000 panels of 10 panellists at random. Every nappe is generated by locating every cava at random on the nappe. MFA is applied to each virtual panel and the first eigenvalue of each analysis is saved. The first eigenvalue of MFA for students and the first eigenvalue of MFA for experts are located on the distribution of 1000 generated panels and thus a p-value is computed (*figure 5.1*). All the panellists are of both panels are considered.

Then, respectively, the two students and the two experts with a lower relationship with the mean configuration (as issued from MFA) are excluded.

In any case, the null hypothesis is rejected, but it is clear that the experts panel present more consensus than the students panel.

We can consider that this test gives a pessimistic result. In any case, we have to enlarge this study. Among the experts, the repeatability of napping is posed but not yet systematically studied.

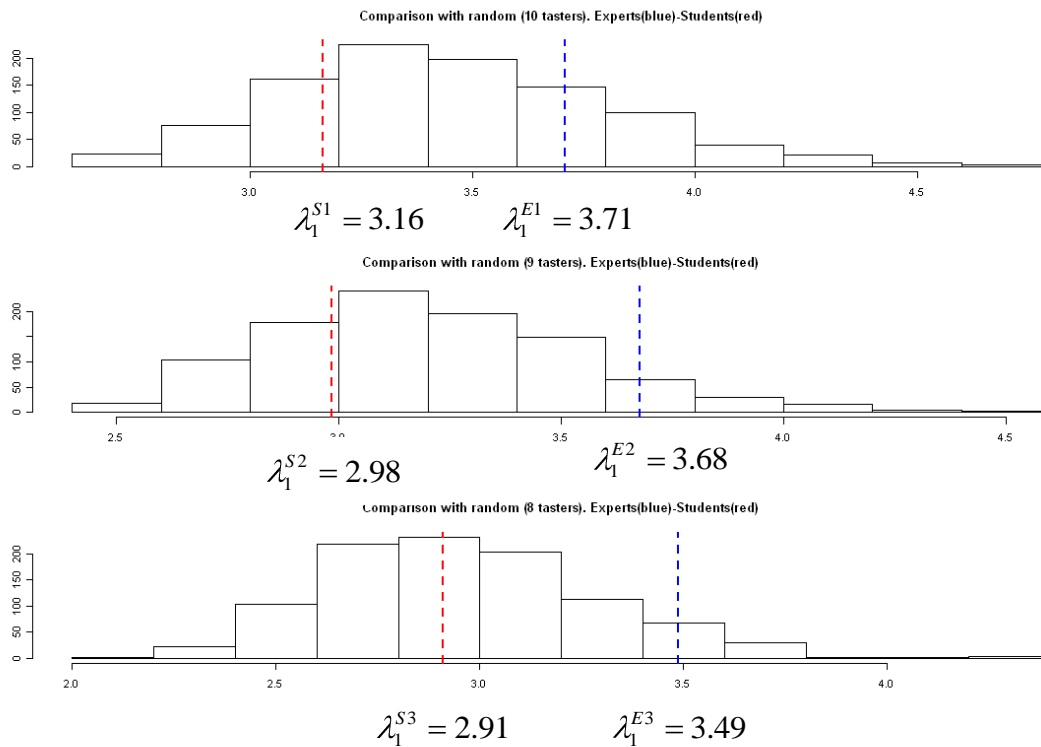


Figure 5.1. Comparison with random panel

5.2 Clustering the panellists

5.2.1 Methodology

We want to see the similarities between panellists. So we have to define a distance between the panellists from the configurations that they produce. After, clustering will allow for summarizing the proximities and determining if clusters of panellists exist. The composition of the clusters will also underline (or not) the differences between experts and students.

Lg coefficient is a measure of the similarity between panellists (chapter 2.3.7). We know that $2^*(1-RV)$ give an euclidean distance and we also know that there is a positive lineal relation between Lg and RV. So we suppose that is possible consider Lg coefficients to make clustering although it is not a distance. These coefficients are obtained from MFA applied to napping of all tasters.

We have considered two different methods to make a clustering of panellists.

- I. Similarity graph
- II. Hierarchical clustering

5.2.3 First method: Similarity graph

A similarity graph is computed on the set of the panellists, from the proximities between their nappings, as computed from *Lg* coefficient (Escofier & Pagès, 1988-2008).

10 shortest edges lead to two clusters. The first cluster was composed of 7 out of 10 experts and one student. The second cluster was composed of only 2 experts and 2 students (*figure 5.2*).

Two graph 20 shortest edges were kept, leading to two clusters of connected panellists (*figure 5.3*). The first was composed of 8 out of the 10 experts, strongly interconnected and 4 out of the 10 non-experts. The second cluster was composed of 2 experts and 4 non-experts. 2 non-experts did not present any connection.

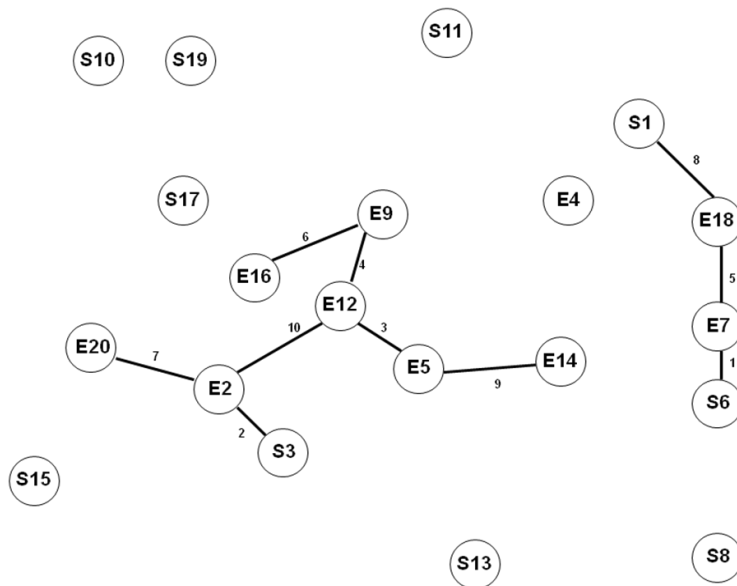


Figure 5.2. Ten highest *Lg* coefficients

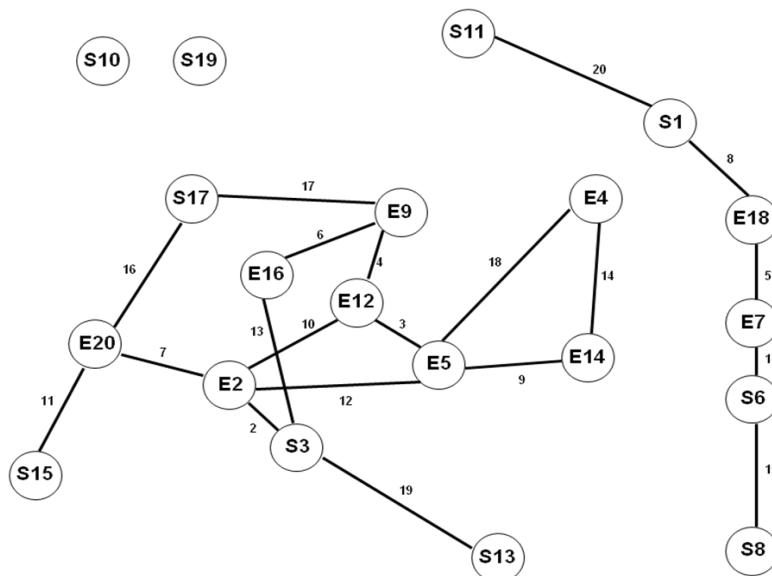


Figure 5.3. Twenty highest *Lg* coefficients

5.2.3 Second method: Hierarchical clustering

A second approach to the homogeneity of the panels is through clustering. We propose to use a hierarchical algorithm. For that purpose, a distance (or similarity) has to be defined between panellists, on the one hand, and between clusters of panellists, on the other hand.

We wish to cluster the panellists from their napping. Thus, we use the Lg coefficient as proximity index between individuals. When two panellists (or more) are gathered into a cluster, this cluster is represented by the mean configuration of the cavas obtained through MFA of the napping performed by these panellists. Then, the proximity between two clusters (or between one cluster and one panellist) is computed through the maximum Lg coefficient between the nappings of the panellists of one cluster and those of the other, which corresponds to apply the “minimum salt” distance between clusters.

This rationale is applied to the 20 panellists. The complete hierarchy is built. Then a partition is chosen which forms three clusters.

The first cluster gathers seven experts and three students; the second cluster contains two students and, finally, the third cluster is composed of three experts and five students (*figure 5.4*).

However, we face the problem that the maximum of Lg coefficients are not always decreasing at successive aggregation levels. Thus, we have to improve this strategy to define a distance between clusters which does not present this drawback. We want to underline that there are no previous works on this problem.

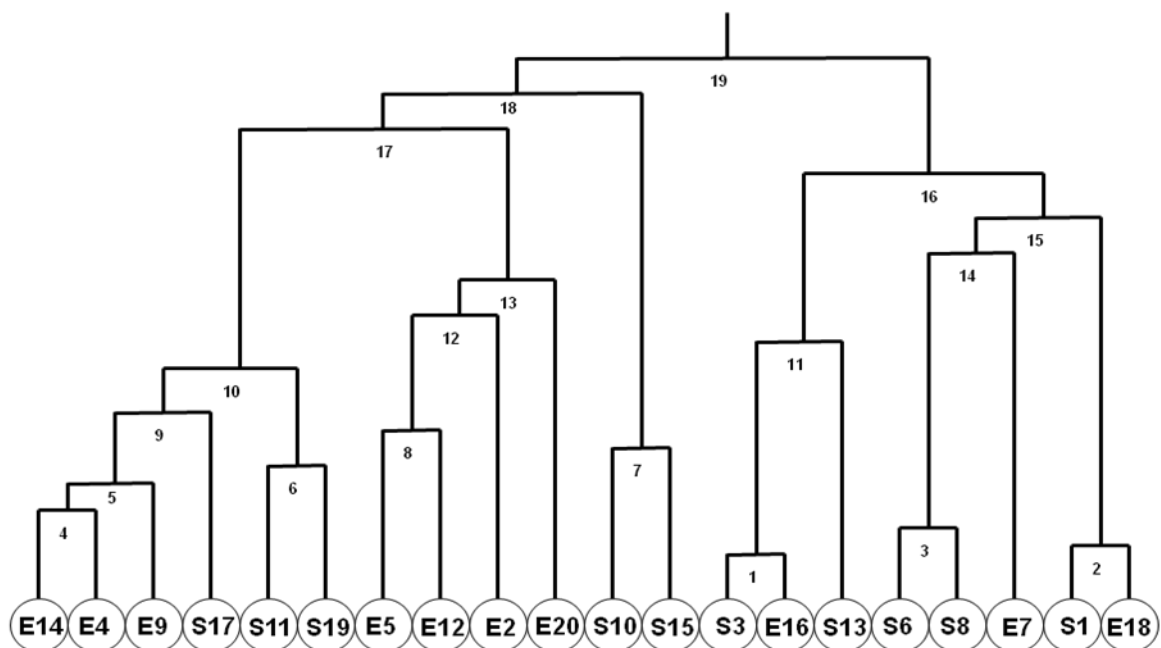


Figure 5.4. Hierarchical clustering

5.3 Conclusions of consensus level and clustering

The two clustering methods that we have used present weakness points. However, they provide clues about the consensus between the panellists that are useful.

Thus, we can conclude that there is a much higher consensus between experts than between students.

However, quantifying the consensus between panellists needs to be improved, looking for a global measure. This problem, not yet tackled by the experts in this field, goes beyond the framework of this project.

Conclusions

Some conclusions are extracted from the previous work:

- Two cavas have a strong influence on the results, which could have been very different if these two cavas would have not been included. In napping, the judgements are relative to the whole of the products that are tested.
- Students and experts work in different ways. Students penalize the cava with cork defect in napping step but, after, they gather the corresponding cava with others in the free sorting task step. Experts did not penalize the cork defect in napping step but, in free sorting task step, they frequently isolated the corresponding cava in a single cluster.
- Experts can explain the reasons of their perception (for example, in the case of the cork defect) by using words (in this case, as TCA and cork) while students cannot explain their perceptions through precise attributes.
- Generally, the experts describe the cavas in a precise way. For example, they associate to aging words as oxidation, evolution and toasted.
- There is much more consensus between experts than students, as seen by quantifying consensus level and by clustering the panellists.
- Students favour the gustatory aspects (such as sugar) while the experts favour the olfactory aspects (through the aromas).
- The global analysis through HMFA hides important differences between experts and students.

Bibliography

- ♣ Bécue-Bertaut,M.;Pagès,J.(2008). Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data. *Comput. Stat. Data Anal.*, 52, 3255-3268.
- ♣ Colmenares,B.(2009). Aplicación de métodos factoriales en los análisis sensoriales: caso de ocho vinos catalanes.
- ♣ Escofier,B.;Pagès,J.(1998). *Analyses factorielles simples et multiples*. Paris: Dunod.
- ♣ Franco,J.;Crossa,J.;Desphande,S.(2009). Hierarchical Multiple Factor Analysis for Classifying Genotypes Based on Phenotypic and Genetic Data. *Crop Science*, 50, 105-117.
- ♣ Husson,F.;Le,S.;Mazet,J.(2006). FactoMineR: Factor analysis and data mining with RR package version 1.02. <http://factominer.free.fr>
- ♣ Le Dien, S.;Pagès,J.(2003). Hierarchical multiple factor analysis: Application to the comparison of sensory profiles. *Food Quality and Preference*, 14, 397-403.
- ♣ Pagès,J.(2005a). Collection and analysis of perceived product interdistancies using multiple factor analysis: application to the study of 10 white wines from the Loire Valley. *Food Quality and Preference*, 16(7), 642-649.
- ♣ Pagès,J.;Husson,F.(2001). Inter-laboratory comparison of sensory profile: Methodology and results. *Food Quality and Preference*, 12, 297-309.
- ♣ Pagès,J.;Morand,E.(2006). Procrustes multiple factor analysis to analyse the overall perception of food products. *Food Quality and Preference*, 17, 36-42.
- ♣ Pagès,J.;Jourjon,F.;Asselin,C.;Maitre,I.;Symoneaux,R.;Perrin,L.(2008). Comparison of three sensory methods for use with the Napping procedure: Case of ten wines from Loire Valley. *Food Quality and Preference*, 19, 1-11.
- ♣ Pagès,J.; Le Dien, S.(2003). Analyse factorielle multiple hiérarchique. *Revue de statistique appliquée*, tome 51,no 2, 47-73.