Pecvnia Monográfico 2012, pp. 31-51

DATOS TEXTUALES COMO ELEMENTOS ACTIVOS EN SENSOMETRÍA / *TEXTUAL DATA AS ACTIVE ELEMENTS IN SENSOMETRY*

Ramón Álvarez-Esteban[1]
ralve@unileon.es
Pedro Aguado Rodríguez[2]
pedro.aguado@unileon.es

Universidad de León

Resumen

La utilización de datos textuales en estudios estadísticos sobre sensometría generalmente se ha realizado tratando de explicar e interpretar los resultados alcanzados a partir de datos cuantitativos. Este trabajo muestra una metodología que permite utilizar datos textuales como elementos activos. Dos catas de vinos ilustran el procedimiento.

Palabras clave: Datos textuales; Sensometría; Análisis de Correspondencias; Análisis Factorial Múltiple.

Abstract

The use of textual data in statistical studies into sensometric field has been conducted generally seeking to explain and interpret results obtained from quantitative data. This work shows a methodology that allows use textual data as active elements. Two wine tastings illustrate the procedure.

Keywords: Textual data; Sensometry; Correspondence Analysis; Factorial Multiple Analysis.

[1] Facultad de Ciencias Económicas y Empresariales, Departamento de Economía y Estadística, Área de Estadística e Investigación Operativa. Universidad de León, Campus de Vegazana, 24071-León (España).

[2] Escuela Superior y Técnica de Ingeniería Agraria. Departamento de Ingeniería y Ciencias Agrarias. Área de Ingeniería Agroforestal. Universidad de León.

1. INTRODUCCIÓN

El análisis estadístico de datos sensoriales es un instrumento necesario para conocer las preferencias de los consumidores y comprender cómo perciben los productos.

El estudio de los vinos en sensometría utiliza información que contiene un elevado número de variables.

Generalmente se utilizan formularios estandarizados en los que un panel de expertos analiza y puntúa aspectos visuales, aromáticos y gustativos. La puntuación media global se obtiene como suma de puntuaciones parciales.

Los métodos estadísticos utilizados en análisis sensorial plantean dificultades cuando se aplican a personas no expertas, especialmente en el caso del vino. Expertos y consumidores pueden utilizar diferentes variables en su decisión de compra o bien valorarlas con distinta importancia.

Es frecuente encontrar personas que no son capaces de describir con palabras un vino. Este hecho ha llevado a suponer que todo conocimiento requiere conocer su lenguaje asociado (Brochet y Dubourdieu, 2001). De esta forma, enólogos, sumilleres y aficionados han modelado un lenguaje del vino para describir las propiedades sensoriales. En la mayor parte de los estudios, esta descripción se realiza analizando la presencia o ausencia de las propiedades sensoriales, más que cuantificando éstas (por ejemplo, utilizando adverbios de cantidad "poco", "mucho", etc.).

Tradicionalmente el análisis estadístico de productos u objetos se ha realizado utilizando métodos multidimensionales a partir de matrices de proximidad entre objetos (Takane, 1980, 1982), pero sin tener

1. INTRODUCTION

Statistical analysis of sensory data is a necessary tool to discover consumer preferences and to understand how products are perceived.

Sensometric study on wine field demands interpreting information that involves a large number of variables.

The classical approach is based on the score from a panel of expert tasters analyzing visual, smell and taste aspects, using standardized forms. The overall score is obtained as a sum of weighted partial scores.

Statistical methods used in sensory analysis have difficulties when they are applied to non-experts, especially in the case of wine. Experts and consumers can use different variables in their purchase decision or evaluate them with different emphasis.

It is common to find people who are not able to describe a wine using words. This problem has raised speculation that all knowledge requires the knowledge of the language associated (Brochet and Dubourdieu, 2001). Thus, winemakers, sommeliers and amateurs have built a language to describe wine sensory properties. In most studies, this description is done by analyzing the presence or absence of sensory properties rather than by quantifying these properties (e.g. using adverbs of quantity "little", "a lot", etc.).

Often the analysis of consumer judgment of products or objects has been conducted using statistical analysis based on multidimensional matrices of proximity between objects (Takane, 1980, 1982) but regardless of the language, the descriptions. The results of these tests

en cuenta el lenguaje, las descripciones. Los resultados suelen mostrar solamente una dimensión predominante, una dimensión hedónica que agrupa las preferencias de gusto de los catadores (Berglund et al., 1973).

No obstante, estos resultados parecen estar más relacionados con la forma de recoger la información que con la existencia de una sola dimensión. Entre estos estudios destacan las ordenaciones de productos (*sorting task*) en función de preferencias, obteniendo similares valoraciones entre expertos y no expertos (Lelièvre et al., 2008).

En otros casos se solicita que el catador agrupe productos en función de sus percepciones, obteniendo para cada catador una tabla productos *x* productos en la que "1" indica que los productos *i* y *j* se perciben dentro del mismo grupo y "0" en caso contrario. En la tabla global la frecuencia *ij* indica el número de veces (catadores) que han señalado el producto *i* dentro del mismo grupo que el producto *j* (Abdi et al., 2007). Las agrupaciones suelen ser similares para expertos y no expertos, pero la descripción de éstas suele ser diferente (por ejemplo, los enólogos buscan defectos en los vinos, mientras que los sumilleres buscan virtudes).

El perfil convencional (*conventional profile*, basado en la norma ISO 11035) es uno de los métodos más clásicos utilizados en los estudios sensométricos. Requiere utilizar un lenguaje común para todos los catadores, construyendo una lista consensuada de atributos (Delarue y Sieffermann, 2004) identificando qué descriptores están presentes o no en cada producto, para lo que se necesita la participación de catadores con gran experiencia. A continuación, se cuantifica

often show only one dominant dimension, a dimension that brings hedonic variables that explains the preferences of the tasters (Berglund et al., 1973).

However, these results seem to be more related to how collect the information than the existence of a single dimension. These studies highlight the arrangement of products (sorting task) according to the preferences, obtaining very similar ratings between experts and nonexperts (Lelièvre et al., 2008).

In other cases each taster is asked to perform product groups according to their perceptions. A table products *x* products is obtained for each taster where "1" indicates that the *i* and *j* products are perceived in the same group and "0" otherwise. In the global table of all the tasters, frequency *ij* indicates the number of times (tasters) who have noted the product *i* within the same group as the product *j* (Abdi et al., 2007). Clusters are often similar for experts and non-experts, but the description of these groups is often different (e.g. winemakers said that they look for defects in wines, while sommeliers look for virtues).

The conventional profile (based on ISO 11035) is one of the most classical methods used in sensometric studies. It requires the use of a common language for all the tasters and building a unique list of attributes through consensus (Delarue and Sieffermann, 2004). First we have to identify which descriptors are present or are not in every product, so we need the participation of highly experienced tasters. Then the intensity of these descriptors is measured. Results of conventional profile seem to be more accurated when tested products are simple

la intensidad de los descriptores. El perfil convencional parece ofrecer mejores resultados cuando los productos evaluados son simples, mientras que en productos complejos, los resultados son peores (Lawless, 1995). Otra desventaja es el tiempo elevado que se necesita para el entrenamiento de los expertos.

El principal objetivo de este trabajo es evaluar si las técnicas textuales aplicadas como elementos activos (y no solamente como información suplementaria o ilustrativa) permiten obtener configuraciones de productos que sean estables, a pesar de las variaciones entre expertos señaladas.

Para ello, se han utilizado nuevas técnicas de recogida de información como el napping y las descripciones textuales (ver Campo et al., 2010, sobre una comparación de la frecuencia de citación y el análisis descriptivo en el caso del vino y Sauvageot et al., 2006, sobre grandes variaciones encontradas en descripciones textuales en expertos).

La sección segunda contiene el proceso de recogida de información con dos catas de vinos relacionadas. A partir de las descripciones textuales de ocho vinos y dieciocho catadores se construyen las tablas de frecuencias. Se aplicó un Análisis de Correspondencias (AC) con cada tabla. En la sección cuarta se realiza un Análisis Factorial Múltiple (AFM) para construir la configuración compromiso a partir de las configuraciones del AC con los distintos umbrales. En la sección quinta se analizan las diferencias de los factores y subespacios del AC en relación a esta nueva configuración compromiso. Finalmente se comparan los resultados de esta métodología con los obtenidos en la primera cata.

and worse using complex products (Lawless, 1995). Another disadvantage is the long time required for the necessary training of experts.

The main objective of this study is to assess if textual techniques applied as active data (and not only as supplementary or illustrative information) provides product stable configurations, in spite of the variations among experts pointed out.

New techniques for gathering information such as napping and textual descriptions have been applied (see Campo et al., 2010, for a comparison of descriptive analysis and frequency of citation in the wine case and Sauvageot et al., 2006, about the large variation in textual descriptions among experts).

Second section describes the process of gathering information through two related wine tasting. From textual descriptions of eight wines and eighteen tasters several contingency tables are built using different thresholds of lexical forms. Correspondence Analysis (CA) is applied for each table in section three. Multiple Factorial Analysis (MFA) in section four allows to obtain an average configuration from the coordinates of all CA analyzed. Section five measures the overall differences between the MFA average configuration and individual CA configurations. Finally the results of this methodology are compared with those obtained from the first tasting

2. MATERIAL Y MÉTODOS

Se han seleccionado ocho vinos tintos correspondientes a dos denominaciones de origen (Bierzo y Tierra de León). Los vinos del Bierzo (etiquetados con los números 1, 3, 6 y 7) son del tipo de uva Mencía y los de Tierra de León del tipo Prieto Picudo. Cuatro de estos son vinos jóvenes, dos con seis meses de madera y los otros dos con doce meses (Tabla 1).

2. MATHERIAL AND METHODS

Eight red Spanish wines from two AOC designations (Bierzo and Tierra de León) were selected. The wines of Bierzo are the type of grape Mencía (wines labelled 1, 3, 6 and 7) and the wines of Tierra de León are the type of grape Prieto Picudo. Four wines are young, two with six months in oak barrels and two with twelve months (Table 1).

Tabla 1. Etiquetas de los vinos / *Table 1. Labels of wines*

DO/*AOC*	Meses en barrica *Months in oak barrels*		
	0	6	12
Bierzo (uva Mencía) / (*grape* Mencía)	1; 7	3	6
Tierra de León (uva Prieto Picudo) / (*grape* Prieto Picudo)	4; 8	2	5

Se conocen varios datos químicos: grado de alcohol, ph, acidez total, acidez volátil real, anhídrido sulfuroso libre y el anhídrido sulfuroso total. En esta experiencia han participado dieciocho personas, nueve enólogos, cuatro sumilleres y cinco aficionados con experiencia.

Cada catador tenía en su mesa ocho catavinos homologados, numerados del uno al ocho. El orden de los vinos fue diferente para catador, utilizando el diseño de cuadrados latinos de Williams (Williams, 1946) con objeto de que el orden de cata de los vinos no produzca efectos sobre los resultados finales.

Se realizaron dos catas sucesivas, con una separación de quince minutos.

2.1. Primera cata

Cada catador dispuso de una hoja de papel (mantel) de 60 centímetros de ancho por 40 centímetros de alto (Pagès, 2003, 2005). Las ocho copas con el vino se colocaron en la parte superior, fuera del

Chemical data are obtained: alcohol, pH, total acidity, volatile acidity, free sulfur dioxide and total sulfur dioxide.

In this experiment, eighteen people participated of which nine are oenologists, four sommeliers and five experienced amateurs.

Eight homologated wine glasses numbered from one to eight are placed in each table. Williams Latin-squares design (Williams, 1946) was used for assign the presentation order of wines for each taster looking for the order did not produce effects in final results.

There was two successive tastings, with a break of fifteen minutes between them.

2.1. First wine tasting

Each taster has a sheet of paper (tablecloth) of 60 cm wide by 40 cm high (Pagès, 2003, 2005). The glasses were placed in the top, outside the tablecloth (nappe). A taster will place next two glasses

mantel (*nappe*). Un catador situará dos vinos tanto más próximos cuanto más se parezcan, utilizando su opinión personal (los criterios que el catador considera, de forma global). De la misma forma, dos vinos se encontrarán tanto más alejados cuanto más diferentes sean percibidos.

Dos catadores tendrán configuraciones distintas si no han considerado los mismos descriptores o no los han ponderado de la misma manera. No hay respuestas buenas ni malas, no hay una configuración de mantel a la que haya que aproximarse. Tampoco deben justificar el motivo por el que han posicionado los vinos sobre el mantel.

Esta disposición será registrada en coordenadas numéricas, obteniendo una configuración individual para cada catador.

Este método *napping* de recogida de información permite obtener las diferencias entre vinos y/o catadores, así como la configuración conjunta, pero es necesario utilizar información adicional para conocer los motivos por los que cada catador decide posicionar las copas, para explicar el perfil sensorial de cada catador. Con el fin de obtener esta información, se solicitó que cada catador realizara una breve descripción de cada vino (*ultra-flash profile*) y la escribiera sobre el mismo mantel.

Por último, los catadores dibujaron en el mantel tantos círculos o elipses como desearon, agrupando los vinos que consideraron parecidos en función de sus percepciones.

Para cada catador se construye una tabla vinos *x* vinos que contiene "1" si los dos vinos han sido agrupados juntos y "0" en caso contrario. Tendremos tantas tablas como catadores (18 tablas). Si sumamos todas las tablas podemos establecer un indicador de la distancia entre cada pareja de vinos.

in the tablecloth if he perceives the wines as resemblance, using his personal opinion (the criteria considered as a whole). In the same way, two wines will be so much far away the more different they are perceived.

Two tasters will build two different configurations if they do not consider the same descriptors or they weight them in a different way. There are not right or wrong answers, there is not a reference tablecloth setting to approach. The tasters should not justify the reasons for locating each glass in the tablecloth.

This layout is recorded in numerical coordinates, obtaining individual settings for each taster.

Napping method of gathering information allows obtaining the differences between the wines and/or tasters as well as a joint configuration, but it is necessary to use additional information to know why each taster has placed the glasses in his tablecloth, to explain the sensorial profile of each taster. To obtain this information, each taster was asked to conduct a brief description of each wine (ultra-flash profile) and write it on the same tablecloth.

Finally, the tasters drew so many circles or ellipses in the tablecloth as they would wish, grouping the similar wines depending on their perceptions.

For each taster we build a table wines *x* wines. Each cell contains "1" if the two wines have been grouped together and "0" otherwise. We will have as many tables as tasters (18 tables). If we add all the tables we can obtain an indicator of the distance between each pair of wines.

2.2. Segunda cata

Se construyó el vocabulario de las palabras elegidas en la primera cata. No se incluyeron artículos, preposiciones, etc. Las formas en masculino y femenino son agrupadas, lo mismo que singulares y plurales (Labbé, 1990). Algunos términos que los catadores consideran sinónimos en la primera cata se agrupan como un solo descriptor en la segunda cata. Por último, los términos que tienen una frecuencia baja no son incluidos en la lista ordenada alfabéticamente con la frecuencia de repetición de cada palabra. Esta lista se muestra a los catadores.

El objetivo es el de consensuar las palabras conservadas, permitiendo añadir nuevos términos para atributos relevantes, obteniendo una nueva lista de descriptores consensuados que puedan ser utilizados en la segunda cata.

La segunda cata fue realizada con los mismos vinos, presentados en un orden distinto, con el fin de que los resultados de la primera cata no pudieran condicionar los de la segunda. Cada catador caracterizó los ocho vinos, asociando los descriptores de la lista consensuada, teniendo en cuenta características visuales, aromas directos y retronasales, sensaciones en boca, etc. El objetivo es la obtención de información que permita explicar las configuraciones obtenidas, así como la variabilidad de los datos recogidos.

Una vez realizada la segunda cata se aplicó la lematización. Se añadieron a los términos los adjetivos y palabras cuantificadoras (e.g. "roble viejo", "roble francés", "roble nuevo", "fruta negra", "bien equilibrado", etc.) (Perrin y Pagès, 2009). En algunos casos descriptores con baja frecuencia se agruparon en otro más amplio para no perder información (e.g. "almizcle" es agrupado bajo el descriptor "animal").

2.2. Second wine tasting

A vocabulary was built from the words used in the first tasting. Articles, prepositions, etc. are not included. Masculine and feminine terms are grouped. The same is done for singular and plural forms (Labbé, 1990). Some words that the tasters consider synonymous are grouped as a single descriptor in the second tasting. Words with low frequency are not included in the agreed list because of their little relevance. The total vocabulary arranged alphabetically with the frequency of repetition of the words was given to each taster.

It was allowed to add new terms for relevant attributes. The goal is to agree the words that we must retain for obtaining a new list of agreed descriptors that can be used in the second tasting.

The second tasting was conducted with the same wines, presented to each taster with a different order, so that the results of the first tasting could not prejudge the outcome of the second. Each taster characterized eight wines, associating the descriptors of the agreed list, taking into account the visual characteristics, direct and retronasal aromas, sensations in the mouth, etc. The objective is to obtain information explaining the configurations drawn in the first tasting, as well as the data variability.

Once the second taste has been performed, lemmatisation is applied. Quantifiers and adjectives were added to terms (e.g. "old oak", "French oak", "new oak", "black fruit", "well balanced", etc.) (Perrin and Pagès, 2009). In some cases specific descriptors with low frequency are grouped into a broader descriptor in order to not lose information (e.g. "musk" was grouped under the descriptor "animal").

El nuevo corpus tiene 948 ítems, de los que 151 son diferentes y 44 son *hapax*. La distribución de los ítems para cada vino se muestra en la Tabla 2.

A continuación se muestran los treinta ítems más utilizados, indicando la frecuencia entre paréntesis: astringente (33), madera (32), lácteo (26), capa alta (24), violáceo (24), acidez (22), fruta (22), especias (21), rojo picota (19), capa media (18), alcohólico (17), carbónico (17), cereza (16), amargo (15), fruta roja (15), tánico (15), tostado (15), vainilla (15), capa media-alta (14), cereza picota (14), cuero (14), grosella (14), corto (13), equilibrado (13), seco (13), glicérico (12), ligero (12), mora (12), redondo (12) y teja (12).

Entre los ítems utilizados al menos tres veces, 52 están asociados a características positivas del vino, 17 a negativas y un tercer grupo de 20 palabras que son positivas o negativas dependiendo del contexto.

The new corpus has 948 total items, of which 151 items are different and 44 are hapax. Frequency citation analysis shows the distribution of items for each wine (Table 2).

The thirty most frequently used items, indicating the frequency in parentheses, are: astringent (33), wood (32), lactic (26), high layer (24), violet (24), acid (22), fruit (22), spice (21), picota-red colour (19), medium layer (18), alcoholic (17), carbonic (17), cherry (16), bitter (15), red fruit (15), tanic (15), roasted (15), vanilla (15), medium-high layer (14), picota-cherry (14), leather (14), redcurrant (14), short (13), balanced (13), dry (13), glyceric (12), light (12), blackberry (12), rounded (12) and russet (12).

Among the items used at least three times, 52 different items are associated with positive characteristics of wine, 17 to negative characteristics and a third group of 20 words that they depending on which context they are used.

Tabla 2. Distribución de los ítems / *Table 2. Distribution of items*

Vino / ***wine***	1	2	3	4	5	6	7	8	Total
Número de ítems */number of items*	111	118	122	110	125	127	112	123	948
% total	11.71	12.45	12.87	11.60	13.19	13.40	11.81	12.97	100
Media de ítems por vino y catador *Average items for wine and taster*	6.2	6.6	6.8	6.1	6.9	7.1	6.2	6.8	6,6
Número diferentes ítems del vino *Number of different items of wine*	57	53	55	54	54	59	48	55	-
% diferentes ítems del vino *% different items of wine*	51.35	44.92	45.08	49.09	43.20	46.46	42.86	44.72	-

3. AC SOBRE LA TABLA VINOS X PALABRAS

Entre las diversas formas de tratar la información recogida en la segunda cata, presentamos la realización del AC sobre la tabla de contingencia formada por los

3. CA FROM TABLE WINES X WORDS

Several ways of managing the information can be used. We present the CA implementation on the occurrence table wines *x* words obtained from the second tasting. Frequencies indicate the number

ocho vinos en fila y un número elevado de columnas (palabras), en la que las frecuencias indican el número de veces que cada palabra aparece en la descripción de un vino, como suma de las opiniones de los catadores. Este resultado puede enriquecerse utilizando información de tipo suplementario (e.g. diferenciando entre sumilleres, enólogos y amateurs).

El umbral de palabras escogido para realizar el AC determinará el número de columnas (Tabla 3).

of times that each word appears in the description of a wine, as the sum of the opinions of the tasters. Supplementary information can be added (e.g. differentiating between oenologists, sommeliers and amateurs).

The occurrence table has eight rows (wines) and a high numbers of columns (words). The threshold of words chosen to carry out the CA determines the number of columns (Table 3).

Tabla 3. Número de palabras diferentes y frecuencia total para los umbrales de dos a diez palabras / *Table 3. Number of different words and total frequency for thresholds from two to ten words*

Umbral/***threshold***	1	2	3	4	5	6	7	8	9	10
Palabras distitnas / *Different words*	151	107	89	79	69	59	51	43	39	36
Total palabras / *Total words*	948	904	868	838	798	748	700	644	612	585

La gran variabilidad de los catadores a la hora de describir los vinos (Labbè et al., 2004), la baja frecuencia de las palabras utilizadas y la posibilidad de trabajar con diferentes umbrales hacen que nos planteemos a priori la posibilidad de obtener resultados estables utilizando AC.

Lebart señala que hay que establecer un umbral para que los resultados del AC tengan sentido estadístico (Lebart et al., 1998). Además, estos resultados debieran ser consistentes con los obtenidos a partir de otras metodologías.

Hay que considerar dos aspectos para elegir el umbral con el que efectuar el AC. Primero, elegir un umbral suficientemente elevado para obtener estabilidad en los resultados (en este caso de los vinos). Segundo, elegir un umbral suficientemente pequeño para que el número de formas léxicas utilizadas sea suficiente para interpretar los resultados.

The great variability of the tasters when describing wines (Labbè et al., 2004), the high number of items, the low frequency of items and the possibility of working with different thresholds lead us to consider a priori the possibility of obtaining stable results using CA.

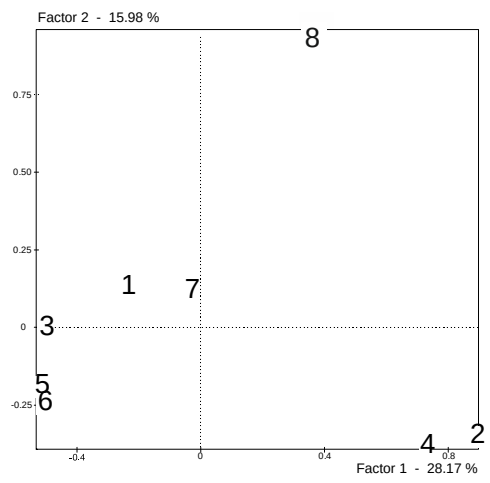
Lebart pointed out that a threshold must be established to achieve than CA results can be the most significant in a statistical sense (Lebart et al., 1998). Moreover, these results should be consistent with those obtained from other methodologies.

The choice of a threshold to carry out the CA can be split into two aspects. First, to choose a high enough threshold that it allows to obtain stability in the results of the CA (in this case of the wines). Secondly, to choose a threshold low enough that the number of used lexical forms is sufficient to interpret the results.

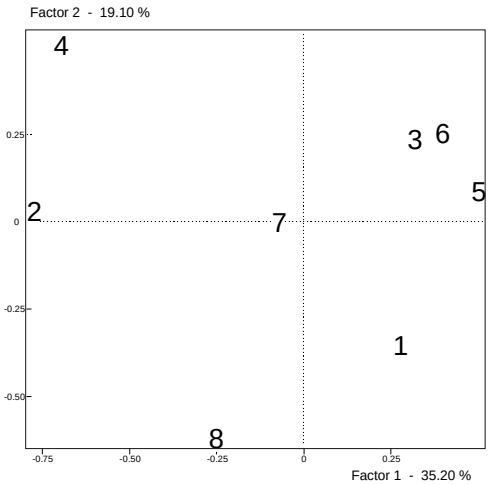
Evidentemente, no es posible comparar las coordenadas de las columnas (palabras) obtenidas de los diferentes AC ya que las palabras son diferentes, pero sí es posible comparar las configuraciones de los ocho vinos. La Figura 1 muestra el primer plano factorial para el AC con umbrales de dos y diez palabras.

Obviously it is not possible to compare the coordinates of the columns (words) obtained from different CA because the words are different, but it is possible to compare the configurations of the eight wines. Figure 1 shows the first factorial plane for CA with thresholds of two and ten words.

Figura 1. Plano factorial del AC para los umbrales de dos y diez palabras
Figure 1. First factorial plane of wines for CA with thresholds of two and ten words



Umbral de dos palabras / *Two words threshold*



Umbral de diez palabras / *Ten words threshold*

La comparación directa de resultados obtenidos a partir de perfiles léxicos diferentes plantea algunos problemas.

En primer lugar, los puntos (vinos) están centrados cuando se ponderan por su frecuencia marginal (número total de palabras que se han retenido para un vino teniendo en cuenta un umbral determinado). En segundo lugar, es posible encontrar reflejos (en la Figura 1 hay reflejos tanto en el eje1 como en el eje2). En tercer lugar, en ocasiones los ejes se intercambian (especialmente cuando los valores propios son muy parecidos). En cuarto lugar, es posible encontrar sub-

The direct comparison of these results obtained from different lexical profiles arise some problems.

First of all, the points (wines) are centered when they are weighted by their marginal frequency (number of words that have been retained for a wine considering a threshold). Secondly, it is possible to find reflections (in Figure 1 we can see reflections in axis 1 and axis 2). In the third place, sometimes the axes are interchangeable (especially when the eigenvalues are very close). Fourth, it is possible to find similar subspaces when using orthogonal rotations (e.g. the first factorial plane is

espacios factoriales similares cuando se utilizan rotaciones ortogonales. Por último, la dilatación de una de las figuras puede hacer disminuir las distancias entre los puntos de las dos configuraciones.

4. AFM DESDE EL AC

De cada AC realizado en el apartado anterior se obtiene una matriz de coordenadas de ocho vinos por siete factores. Esta matriz ha sido centrada por columnas, con el fin de dar a cada uno de los vinos la misma ponderación. Este proceso se repite para todos los umbrales analizados.

A continuación se ha aplicado un AFM (Escofier y Pagès, 1990, 1994), utilizando el paquete de R FactoMineR (Husson et al., 2007). En el AFM se realiza un Análisis de Componentes Principales ACP para cada matriz de coordenadas X_i (ocho vinos *x* siete dimensiones), obteniendo los valores propios λ_l^j, indicando "i" el número del valor propio y "j" el grupo (en nuestro caso el índice del umbral elegido).

Las nueve tablas X_i correspondientes a las coordenadas para los distintos umbrales (de dos a diez) se yuxtaponen en una tabla X, ponderando cada una por la inversa de la raíz cuadrada del primer valor propio λ_1^j procedente del ACP de las j tablas individuales

the same but the factors are not). Finally, the dilation of one configuration can decrease the distances between the points of the two configurations.

4. MFA FROM CA

From the coordinates of the CA in previous section, we will get a matrix of eight wines by seven factors. This matrix has been centered by columns, in order to give the same weight to each wine. This process is repeated for all the thresholds analyzed.

Next a MFA (Escofier and Pagès, 1990, 1994) has been performed using R FactoMineR package (Husson et al., 2007). In MFA, a Principal Component Analysis PCA is carried out for each coordinates matrix X_i (eight wines *x* seven dimensions), obtaining the eigenvalues λ_l^j, notating "i" the number of the eigenvalue and "j" the group (in our case the index of threshold chosen).

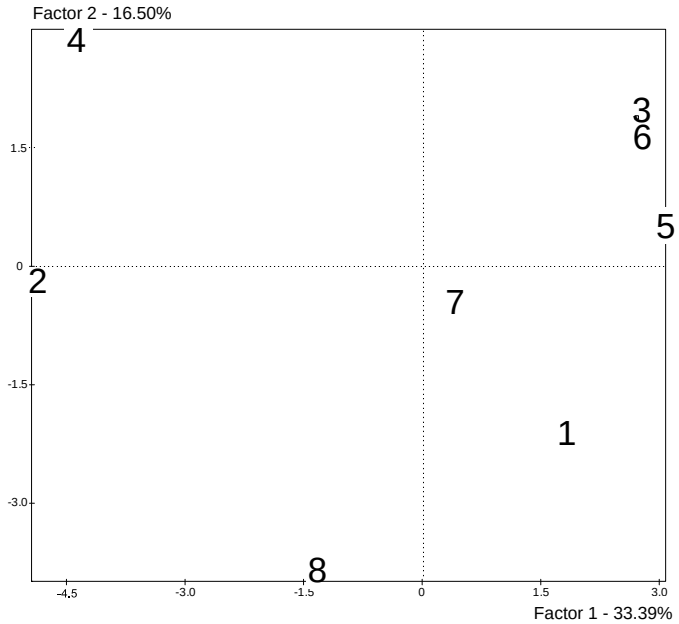
All the nine coordinates tables X_i for the different thresholds (from two to ten) are juxtaposed in a table X, weighting each one by the inverse of the square root of the first eigenvalue λ_1^j from the j PCA individual tables.

$$X = \left(\frac{1}{\sqrt{\lambda_1^1}} X_1, \frac{1}{\sqrt{\lambda_1^2}} X_2, \ldots, \frac{1}{\sqrt{\lambda_1^J}} X_J \right)$$

A partir de la diagonalización de la tabla *X'X* se obtiene la configuración compromiso del AFM (Figura 2).

MFA average configuration is obtained diagonalizing *X'X*. Figure 2 shows the first factorial plane.

Figura 2. Primer plano factorial para la configuración compromiso obtenida del AFM / *Figure 2. First factorial plane for MFA average configuration*



5. DISTANCIA INDIVIDUAL Y GLOBAL

La Figura 2 indica que tan solo hay pequeñas variaciones con las gráficas de la Figura 1 obtenidas para los umbrales 2 y 10.

Entre las medidas para comparar la similitud entre dos configuraciones (distancia global entre la configuración media del AFM y la del AC) se encuentra el coeficiente RV (Krzanowski, 1990), basado en la similaridad normalizada Procrustres.

Si X e Y son dos matrices con las coordenadas de los vinos (ocho) por las dimensiones (siete) y se da la misma ponderación a cada uno de los vinos, siendo *tr()* la traza, el coeficiente *RV* se define como:

5. GLOBAL AND INDIVIDUAL DISTANCE

There are only small variations between the MFA average configuration and the obtained for thresholds two and ten of CA (Figure 2).

Among the measures to compare the similarity between two configurations (global distance between MFA average configuration and CA configuration) we can use the *RV* coefficient (Krzanowski, 1990), based on the normalized Procrustres similarity.

If X and Y are matrices with the coordinates of wines (eight) by the dimensions (seven), giving the same weight to each wine, and notating *tr()* as the trace, *RV* coefficient can be computed as:

$$RV = \frac{tr(XX'YY')}{\sqrt{tr(XX'XX')tr(YY'YY')}}$$

RV varía entre 0 y 1 (Josse et al., 2008). Un valor de RV igual a 1 indica que las dos configuraciones son idénticas. Un valor de RV igual a 0 indica que cada punto de la primera tabla está incorrelado con los de la segunda.

RV varies between 0 and 1 (Josse et al., 2008). A RV value of 1 indicates that both configurations are identical. A RV value of 0 indicates that each point of the first table is uncorrelated to each point of the second table.

La Tabla 4 muestra los coeficientes RV entre las configuraciones para los diferentes umbrales (U) y la compromiso (AFM en itálica).

Table 4 shows the RV coefficient between configurations for the different thresholds (U) of CA and the MFA average configuration (italic letters).

Si solamente se deseara comparar un subespacio es necesario rotar previamente la tabla Y hacia la tabla X.

If only a subspace must be compared it is necessary to rotate the Y table toward X table.

Tabla 4. *RV* entre vinos de las configuraciones AFM y AC para umbrales dos al diez / *Table 4. RV between wines MFA and CA configuration for thresholds from two to ten*

	U2	U3	U4	U5	U6	U7	U8	U9	U10	*MFA*
U2	1.000	0.996	0.991	0.985	0.981	0.971	0.962	0.958	0.955	*0.986*
U3	0.996	1.000	0.998	0.993	0.991	0.981	0.973	0.968	0.967	*0.993*
U4	0.991	0.998	1.000	0.996	0.992	0.984	0.978	0.974	0.973	*0.995*
U5	0.985	0.993	0.996	1.000	0.995	0.991	0.981	0.978	0.977	*0.996*
U6	0.981	0.991	0.992	0.995	1.000	0.996	0.989	0.985	0.984	*0.998*
U7	0.971	0.981	0.984	0.991	0.996	1.000	0.990	0.985	0.984	*0.994*
U8	0.962	0.973	0.978	0.981	0.989	0.990	1.000	0.998	0.997	*0.992*
U9	0.958	0.968	0.974	0.978	0.985	0.985	0.998	1.000	1.000	*0.990*
U10	0.955	0.967	0.973	0.977	0.984	0.984	0.997	1.000	1.000	*0.989*
MFA	*0.986*	*0.993*	*0.995*	*0.996*	*0.998*	*0.994*	*0.992*	*0.990*	*0.989*	*1.000*

RV es muy alto para todos los casos, es decir, las configuraciones de los vinos obtenidas del AC utilizando diferentes umbrales son muy parecidas a la configuración compromiso del AFM. Además la compara-

RV is very high for all cases considered, that is to say, the wines configurations from CA using different thresholds are very similar to the average configuration of the MFA. Furthermore, the comparison

ción de vinos con diferentes umbrales (parte central de la Tabla 4) indica que las configuraciones son muy parecidas, independientemente del umbral elegido.

Altos valores *RV* indican que las configuraciones globalmente son muy parecidas, pero no que todos los factores sean iguales, no que todos los factores puedan ser seleccionados para su interpretación.

Por ello, se han calculado los coeficientes de correlación de Pearson entre cada factor del AFM con el correspondiente factor del AC (Tabla 5).

of wines configurations between different thresholds (middle part of Table 4) shows that the configurations are very close, regardless of the chosen threshold.

High *RV* values indicate that overall configurations are very similar, but not that the factors are equal, not that all factors can be selected for their interpretation.

Therefore, Pearson's correlation coefficients between each MFA factor and the corresponding CA factor have been computed (Table 5).

Tabla 5. Coeficientes de correlación de Pearson entre las coordenadas de los vinos del AFM y las coordenadas del AC para distintos umbrales. Los menores coeficientes están marcados en negro / *Table 5. Pearson's correlation between wines MFA coordinates and CA coordinates for different threshold. Lowest coefficients are pointed out in black*

Umbral \ *Threshold*	r_{11}	r_{22}	r_{33}	r_{44}	r_{55}	r_{66}	r_{77}
2	-0.996	-0.830	-0.699	0.422	0.616	0.937	0.940
3	0.995	0.472	0.524	0.855	0.906	-0.935	0.950
4	0.997	-0.770	-0.793	0.865	0.894	0.968	0.993
5	-0.999	0.786	-0.820	-0.956	-0.862	0.936	0.926
6	0.996	0.940	0.876	0.880	0.934	0.932	0.925
7	0.997	0.967	-0.642	0.591	0.886	0.974	-0.975
8	0.996	-0.979	0.951	0.946	-0.915	0.907	-0.845
9	0.995	0.981	0.929	0.942	0.897	-0.399	0.326
10	0.995	0.993	0.911	-0.925	0.905	-0.289	-0.230

La primera columna r_{11} muestra altas correlaciones entre las coordenadas del primer factor del AFM con las coordenadas del primer factor para el AC con diferentes umbrales. Esto indica que es un factor estable desde el punto de vista de los umbrales escogidos y es posible interpretar las configuraciones de los vinos utilizando las formas lé formas léxicas como descriptores.

First column r_{11} shows high correlations between the coordinates of the first MFA factor and the coordinates for first CA factor for the different thresholds. This means that it is a stable factor from the point of view of the thresholds chosen and it is possible to interpret the wine configuration using the lexical forms as descriptors.

La Tabla 6 muestra un resumen de las formas léxicas que caracterizan el primer factor del AC utilizando un umbral de 5 palabras. Las coordenadas negativas corresponden a características negativas de los vinos.

Para los factores segundo y tercero (Tabla 5) las correlaciones con los tres umbrales más grandes indican factores similares. Las correlaciones son más bajas para los umbrales pequeños (marcadas en negro), por lo que podríamos pensar que no hay estabilidad en el segundo y tercer factor, por lo que no debieran ser interpretados.

Las mayores diferencias entre los factores 2 y 3 del AFM (Figura 3a) con los del AC corresponden al umbral de tres palabras (Figura 3b) con correlaciones de 0.472 y 0.524.

The following table contains an excerpt from the lexical forms that characterize the first CA factor using a threshold of five words (Table 6). Negative coordinates are related to negative characteristics of wines.

For the second and third factors (Table 5), correlations with the three highest thresholds (8, 9 and 10) are very high, indicating similar factors. Smaller thresholds show low correlations (highlighted in black), so we might think that there is no stability in the second and third factors, so they should not be interpreted.

The biggest differences of MFA factors 2 and 3 (Figure 3a) with the CA correspond to three words threshold (Figure 3b) with correlations 0.472 and 0.524.

Tabla 6. Resumen de formas léxicas características del primer factor del AC con umbral de cinco palabras. / *Table 6. Excerpt of lexical forms for the first CA factor using a threshold of five words*

Coordenadas negativas / *negative coordinates* Características negativas / *negative characteristics* a) siempre / *always* b) depende del contexto / *depending on the context*	Coordenadas positivas / *positive coordinates* Características positivas / *positive characteristics*
a) capa media-baja / *medium-low layer*, evolucionado / *evolved*, oxidado / *oxidised*, licor / *liquor*, desequilibrado / *unbalanced*, sucio / *dirty*, herbáceo / *herbaceous*, corto / *short*, carbónico / *carbonic*, acidez / *acid*, ligero / *light*, madera vieja / *old oak*, capa media / *medium layer*, verde / *green*, desagradable / *unpleasant*, amargo / *bitter*, intensidad baja / *low intensity*, intensidad media / *medium intensity.* b) teja / *russet*, caramelo / *caramel*, cereza / *cherry*, color rojo cereza / *cherry-red colour*, fresco / *fresh*, fresa / *strawberry*, frambuesa / *raspberry.*	persistente / *persistent*, capa alta / *high layer*, madera nueva / *new oak*, tostado / *roasted*, redondo / *rounded*, *suave* / *smooth*, intenso / *intense*, fruta negra / *black fruit*, roble francés / *French oak*, rojo cereza / *red cherry*, tonos violáceos / *hints of violet*, color cereza picota / *picota-red colour*, madera / oak, estructurado / *structured*, largo / *long*, cálido / *hot*, torrefacto / *high-roast*, agradable / *pleasant*, vainilla / *vanilla*, glicérico / *glyceric*, toffee / *toffee*, equilibrado / *balanced*, animal /*animal*, violáceo / *violet*, tánico / *tannic*, mora / *blackberry*, capa medio-alta / *high-medium layer*, mineral / *mineral*, coco / *coconut.*

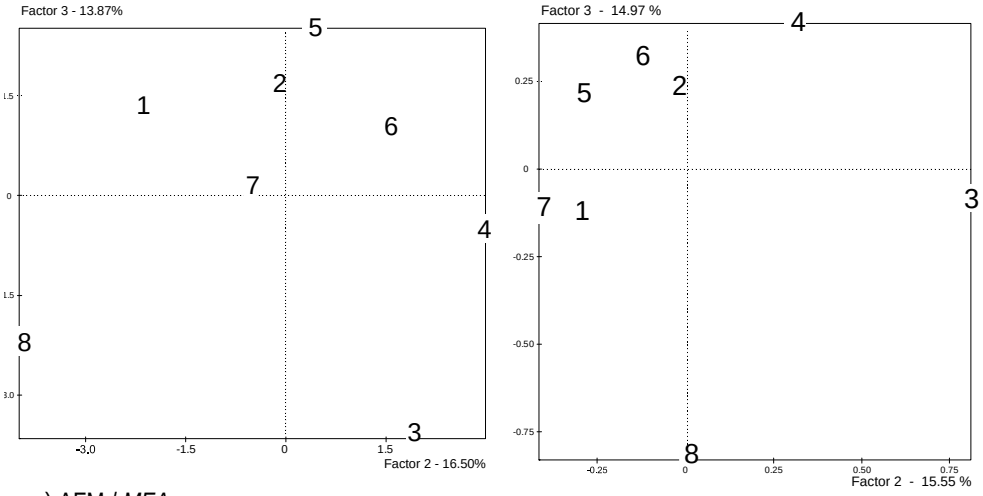
Como indicamos anteriormente, hay que considerar la posibilidad de que haya reflejos, intercambios, rotaciones y dilataciones cuando se comparan configuraciones, especialmente cuando los valores propios son cercanos (15.55% de la varianza del factor dos y 14.97% del factor tres de la Figura 3b).

Si X es la configuración de destino o referencia del MFA e Y es la configuración de origen CA, la matriz de rotación H puede ser obtenida minimizando $\|X\text{-}YH\|$.

As indicated above, there must be considered the possibility of reflections, interchanges, rotations and dilations when comparing configurations, especially when eigenvalues are close (15.55% of variance for factor two and 14.97% for factor three in Figure 3b).

If X is the target or reference MFA configuration and Y is the source CA configuration, H rotation matrix can be obtained minimizing $\|X\text{-}YH\|$.

Figura 3. Plano factorial para los factores dos y tres del AFM y el AC con umbral de tres palabras / *Figure 3. Factorial plane factors two and three from MFA and CA (threshold of three words)*



a) AFM / *MFA*

b) AC para tres palabas / *CA for a threshold of three words*

H se determina descomponiendo $X^{\mathrm{T}}Y$ en las matrices de vectores propios U y V, y la matriz de los valores S, estableciendo la relación $X^{\mathrm{T}}Y=USV^{\mathrm{T}}$. La matriz de rotación se define como $H=VU^{\mathrm{T}}$.

En ocasiones es necesario realizar un cambio de escala obteniendo la matriz dilatada como $Z = \mathrm{c}YH$, siendo *tr()* la traza, c es la

H is determined decomposing $X^{T}Y$ in U and V eigenvectors matrices and a S matrix of eigenvalues, establishing the equation $X^{\mathrm{T}}Y=USV^{\mathrm{T}}$. Rotation matrix is defined as $H=VU^{\mathrm{T}}$.

Sometimes is necessary to carry out a scale exchange. Dilated configuration Z is defined as $Z = \mathrm{c}YH$. Constant dilation c is

constante de dilatación:
$c=tr(YHX^T)/tr(YY^T)$.

Los factores en la Figura 3 obviamente no son los mismos, pero si realizamos una rotación ortogonal (Gower y Dijksterhuis, 2004) de AC hacia la configuración del AFM, las diferencias decrecen mucho (Figura 4).

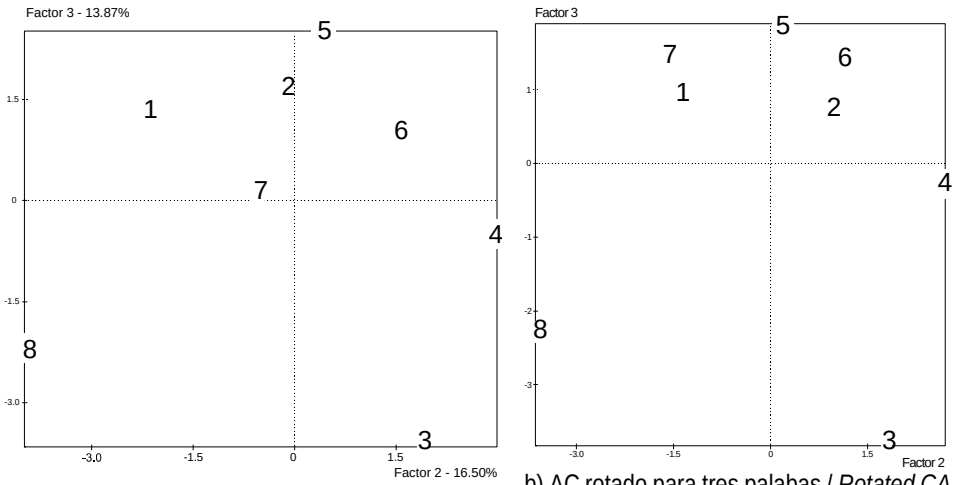
Estos resultados indican que hay diferencias entre los factores considerados uno a uno, pero hay estabilidad en el plano factorial formado por los factores dos y tres con distintos umbrales.

computed as $c=tr(YHX^T)/tr(YY^T)$ when *tr()* is the trace.

Factors in Figure 3 obviously are not the same, but if we perform an orthogonal rotation (Gower and Dijksterhuis, 2004) of CA to MFA configuration differences decrease a lot (Figure 4).

These results show that there are differences between the factors taken one by one but there is stability in the factorial plane made up by factors two and three using different thresholds.

Figura 4. Plano factorial para los factores dos y tres del AFM y el AC rotado con umbral de tres palabras / *Figure 4. Factorial plane factors two and three from MFA and rotated CA (threshold of three words)*



a) AFM / *MFA*

b) AC rotado para tres palabas / *Rotated CA for three words*

La configuración media del AFM ha sido comparada con los resultados del napping de la primera cata, obteniendo resultados consistentes (no incluidos en este trabajo).

The MFA average configuration has been compared with the results of napping from the first taste, getting consistent results (not included in this paper).

6. ANALYSIS CLUSTER

Con el fin de contrastar la consistencia de estos resultados se seleccionaron las coor-

6. CLUSTER ANALYSIS

In order to compare the consistency of these results MFA factorial coordinates of

denadas factoriales del AFM de la segunda cata a partir de los datos textuales considerados como elementos activos. Con estas coordenadas se realizó un análisis cluster con método el promedio entre grupos y la distancia euclídea al cuadrado (Figura 5.a). También se realizó un análisis cluster con la tabla agregada obtenida a partir de los clusters de los vinos creados por los catadores de la primera cata (Figura 5.b).

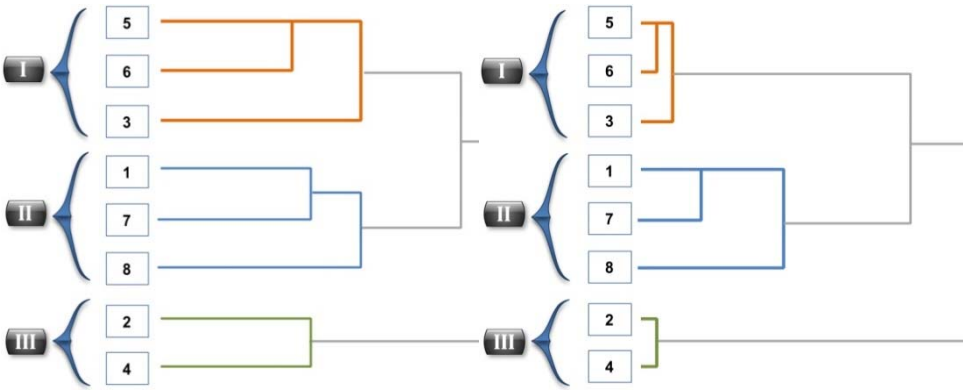
Como puede observarse en la siguiente figura, los tres grupos obtenidos son iguales utilizando las dos metodologías.

the second tasting from textual data considered as active data were selected. With these coordinates, a cluster analysis was performed using the average linkage between groups method and the squared Euclidean distance (Figure 5.a).

A cluster analysis with the aggregated table from the wine clusters created by the tasters in the first tasting was also carried out (Figure 5b).

As shown in the following figure, the three resulting groups are equal using the two methodologies

Figura 5. Análisis clusters utilizando a) las palabras como elementos activos y b) a partir de las agrupaciones de la primera cata / *Figure 5. Clusters analysis using a) words as active elements and b) from groups in the first tasting*



a) Clusters con los factores del AFM de la segunda cata de vinos
Clusters from MFA factors in the second wine tasting

b) Clusters de vinos para la primera cata *Clusters for first wine tasting*

7. CONCLUSIONES

La descripción de propiedades sensoriales de los productos no es una tarea fácil, incluso para los expertos, ya que pueden utilizar diferentes variables y ponderaciones. En la mayoría de los estudios, esta descripción se realiza mediante el análisis de la

7. CONCLUSIONS

Describing sensory properties of products is not an easy task, even for experts because they can use different variables or weighting. In most studies, this description is done by analyzing the presence or absence of sensory properties.

presencia o ausencia de las propiedades sensoriales.

En este trabajo, los datos textuales se aplican como datos activos, con el objetivo de construir configuraciones estables de productos, a pesar de las variaciones entre tres tipos de expertos.

Se realizaron dos sucesivas catas de vino. En la primera se utilizó el "napping" como método de recopilación de información y un "ultra-flash profile", creando un vocabulario y la lista de descriptores. Los expertos consensuaron una nueva lista a partir de este vocabulario.

Antes de aplicar el AC con las palabras de este segundo vocabulario es necesario fijar un umbral para la selección de palabras. Diferentes umbrales proporcionan diferentes configuraciones de coordenadas. Estas coordenadas del AC son seleccionadas para construir una configuración media utilizando el AFM. El coeficiente *RV* indicó que la configuración media del AFM era muy similar a las obtenidas para los diferentes umbrales.

Los resultados del AFM y del AC pueden ser comparados, pero es preciso utilizar rotaciones ortogonales. La metodología utilizada se ha mostrado útil para comparar configuraciones, factores individuales y subespacios.

La utilización de datos textuales como elementos activos ha proporcionado resultados consistentes, incluso al utilizar umbrales de palabras distintos. Esta consistencia se ha manifestado comparando los resultados de datos textuales con los obtenidos a partir de un cluster con los datos no textuales de la primera cata.

In this paper, textual data are applied as active data with the aim of building product stable configurations, in spite of the variations among three types of experts.

There was two successive wine tastings. Napping method of gathering information and ultra-flash profile were used in the first one, building a vocabulary and obtaining a list of descriptors. Experts agreed a new list from the results of the first tasting vocabulary.

Before applying the CA with the words of the second vocabulary it is necessary to fix a threshold for the selection of words. Different thresholds lead to different coordinates configurations. These coordinates from CA are selected to build an average configuration using MFA. *RV* coefficient indicated that average MFA configuration was similar to those obtained for various thresholds.

MFA and CA results can be compared but it can be necessary using orthogonal rotations. Methodology applied is useful to compare global configurations, individual factors and subspaces.

The use of textual data as active elements has led to similar wines configurations, even using different frequency thresholds of words.

This consistency is shown by comparing the textual data results with those obtained from a cluster with non-textual data from the first tasting.

BIBLIOGRAFÍA/REFERENCES

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

Berglund, B., Berglund, U., Engen, T. y Ekman, G. (1973). Multidimensional analysis of twenty-one odors. *Scandinavian Journal of Psychology*, *14*, 131-137.

Brochet, F. y Dubourdieu, D. (2001). Wine Descriptive Language Supports Cognitive Specificity of Chemical Senses. *Brain and Language*, *77*(2), 187-196.

Campo, E., Ballester, J., Langlois, J., Dacremont, C. y Valentin, D. (2010). Comparison of conventional descriptive analysis and a citation frequency-based descriptive method for odor profiling: An application to Burgundy Pinot noir wines. *Food Quality and Preference*, *21*, 44-55.

Delarue, J. y Sieffermann, J-M. (2004). Sensory mapping using Flash profile. Comparison with a conventional descriptive method for the evaluation of the flavour of fruit dairy products. *Food Quality and Preference*, *15*, 383-392.

Escofier, B. y Pagès, J. (1990). *Analyses factorielles simples et multiples: Objectifs, méthodes, interprétations*. Paris: Dunod.

Escofier, B. y Pagès, J. (1994). Multiple factor analysis (AFMULT package). *Computational Statistics and Data Analysis*, *18*, 121-140.

Gower, J.C. y Dijksterhuis, G.B. (2004). *Procrustes problems*. Oxford University Press.

Husson, F., Lê, S. y Mazet, J. (2007). FactoMineR: Factor Analysis and Data Mining with R.R package version 2.4.0, URL http://factominer.free.fr/ (accessed September 2010).

Josse, J., Pagès, J. y Husson, F. (2008). Testing the significance of the RV coefficient. *Computational Statistics and Data Analysis*, *53*, 82-91.

Krzanowski, W.J. (1990). Principles of multivariate analysis, a user's perspective. Oxford Statistical Science Series.

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

Labbé, D. (1990). Normes de dépouillement et procédures d'analyse des textes politiques, CERAT.

Labbé, D., Rytz, A. y Hugi, A. (2004). Training is a critical step to obtain reliable product profiles in a real food industry context. *Food Quality and Preference*, *15*, 341-348.

Lebart, L., Salem, A. y Berry, L. (1998). Exploring textual data. Dordrecht, Boston: Kluwer Academic Publisher.

Lelièvre, M., Chollet, S., Abdi, H. y Valentin, D. (2008). What is the validity of the sorting task for describing beers? A study using trained and untrained assessors. *Food Quality and Preference, 19*, 697-703.

Pagès, J. (2003). Recueil direct de distances sensorielles: Application à l'évaluation de dix vins blancs du Val-de-Loire. *Sciences des Aliments*, *23*, 679-688.

Pagès, J. (2005). Collection and analysis of perceived product interdistances using multiple factor analysis: application to the study of 10 white wines from the Loire Valley. *Food Quality and Preference*, *16*(7), 642-649.

Perrin, L. y Pagès, J. (2009). Construction of a product space from the ultra-flash profiling method: application to 10 red wines from the Loire Valley. *Journal of Sensory Studies*, *24*(3), 372-395.

Sauvageot, F., Urdapilleta, I. y Peyron, D. (2006). Within and between variations of texts elicited from nine wine experts. *Food Quality and Preference*, *17*(6), 429-444.

Takane, Y. (1980). Analysis of categorizing behavior by a quantification method. *Behaviormetrika*, *8*, 75-86.

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

Williams, E.J. (1949). Experimental designs balanced for the estimation of residual effects of treatments. *Australian Journal of Scientific Research*, Ser. A 2, 149-168.