Mesurem els mots?
L'anàlisi estadística de textos

Mónica Bécue-Bertaut
Departament d'Estadística i Investigació Operativa
Universitat Politècnica de Catalunya

Resumen

En las encuestas por cuestionario, es usual introducir preguntas abiertas. En los estudios de consumo, los expertos y/o consumidores evalúan los productos mediante una puntuación y, conjuntamente, un comentario libre. El análisis estadístico de este tipo de datos, llamados datos mixtos, requiere de métodos específicos. Proponemos aquí emplear una metodología que combina el análisis factorial múltiple clásico y el análisis factorial múltiple para tablas de contingencia. Esta extensión del AFM permite tratar globalmente tablas léxicas, creadas a partir de las respuestas textuales, y tablas cuantitativas y categóricas.

Presentamos la metodología apoyándonos en datos recogidos en la evaluación de dos conjuntos de vinos. El análisis sensorial constituye en efecto un área de aplicación privilegiada, dado que los expertos y consumidores desean frecuentemente complementar los clásicos perfiles sensoriales con descripciones libres que traducen más fielmente sus percepciones.

La metodología se puede emplear también en el tratamiento de datos de encuesta, cuando se aborda una problemática con una batería de preguntas cerradas y de preguntas abiertas.

1. Introducción

Conocer las opiniones de los usuarios y/o consumidores es un objetivo crucial en los estudios de mercados. En ciertos casos, dichas opiniones no se pueden trasmitir mediante una puntuación o la selección de un ítem entre los propuestos y es necesario introducir preguntas abiertas. Los entrevistados expresan así su opinión simple o compleja, lacónica o rica en matices, a veces contradictoria, de manera espontánea.

En la industria agroalimentaria, está establecido que la calidad de los alimentos no se puede medir únicamente a partir de análisis químicos o físicos. Es importante colectar las percepciones de los consumidores y/o expertos mediante puntuaciones, apreciaciones sensoriales cualitativas y, más recientemente, comentarios libres. Los datos recogidos suelen ser voluminosos y requieren de la aplicación de métodos estadísticos, algunos específicos de este campo.

Presentamos aquí una metodología para el tratamiento de comentarios libres en estudios sensométricos y/o respuestas abiertas en encuestas.

La estructura de la exposición es la siguiente. En la sección 2, se recuerdan los principios básicos de codificación de los textos. En la sección 3, se presenta una metodología para el análisis simultáneo de varias tablas de frecuencia (Bécue-Bertaut & Pagès, 2004) y su aplicación a comentarios de cata en un estudio efectuado sobre vinos del Priorat. La sección 4 expone una metodología para el análisis de tablas mixtas, con columnas-variables cuantitativas, categóricas y de tipo frecuencia y su aplicación al estudio de un conjunto de vinos, descritos mediante un comentario libre y una puntuación. La sección 5 describe brevemente una investigación en curso, destinada a comparar la apreciación de unos mismos vinos por dos paneles de catadores expertos, uno catalán y uno francés. La sección 6 concluye la exposición

2. Análisis estadístico de respuestas abiertas

2.1. Respuestas abiertas y trasmisión de opiniones complejas

Las respuestas abiertas aportan una información específica (Lebart et al., 2000) y permiten abordar temas difíciles y complejos que requieren respuestas espontáneas. Dichas respuestas no se pueden enmarcar en los estrechos límites de una corta serie de ítems, frecuentemente propuestos en las encuestas de de opinión.

La selección del vocabulario expresa una realidad subyacente compleja y difícil de aprehender por otro tipo de cuestionamiento. El análisis estadístico parte de las ocurrencias y co-ocurrencias de las distintas palabras y constituye una herramienta poderosa para acceder al contenido de los textos (Benzécri, 1981, Lebart, 2003; Murtagh et al., 2009).

2.2. Codificación de la información textual

Se considera un nuevo tipo de variable –llamada variable textual– codificada en una tabla individuos×palabras (o tabla léxica), construida mediante el recuento de las diferentes palabras en las respuestas individuales. Esta estructura más compleja que la estructura inducida por una variable cuantitativa (una columna) o una variable categórica (tantas columnas como categorías) resulta de la variabilidad del lenguaje que es también su riqueza.

Identificar las unidades léxicas requiere un cuidadoso pretratamiento (Labbé, 1990; Muller, 1977-1992; Salton & MacGill, 1983) que:

- Corrige las faltas de ortografía.
- Opera una lematización que convierte la forma gráfica de cada ocurrencia en su forma estandarizada llamada lema (o voz del diccionario: infinitivo para los verbos, singular para los sustantivos, masculino singular para los adjetivos, etc.) e identifica la categoría gramatical de cada ocurrencia. Seguimos aquí la norma lexicométrica introducido por Muller (1977-1992) y completada por Labbé (1990) que limita las unidades léxicas compuestas por varias ocurrencias a las locuciones fuertemente establecidas en la lengua. A pesar de que ciertos lemas puedan modificar su significado en función del contexto –en particular, los adjetivos pueden estar fuertemente influenciados por un adverbio de cantidad–, se conserva los distintos lemas como unidades separadas. En lo que sigue, empleamos el término genérico *palabra* como sinónimo de *lema*.
- Define la stoplist, o lista de palabras consideradas no útiles en el estudio en curso. En general, dicha lista contiene las preposiciones, artículos, pronombres, conjunciones y adjetivos posesivos y demostrativos.
- Selecciona un umbral de frecuencia, dado que la comparación entre respuestas sólo tiene significado si las palabras tienen una frecuencia mínima (Lebart et. al., 2000). Una regla pragmática consiste en eliminar las palabras escogidas por menos de 2% de los individuos.

Así, se adopta una norma para el recuento lexicométrico que es estable, comprensible y reproducible: no varía durante el pretratamiento de un corpus de textos, ni de un corpus a otro, es fácil de entender por los usuarios y de aplicar en cada estudio y/o por cada usuario.

El pretratamiento ofrece una codificación trasparente de las respuestas, operado por el analista mediante un proceso explícito e igualmente aplicado a todas las respuestas.

Obviamente, la significación de algunas palabras puede variar según el contexto. Como lo ha estudiado Lehrer (1975), la potencia del lenguaje proviene de su labilidad y el significado impreciso de las palabras no es sino una virtud. El *perfil léxico* se debe aprehender globalmente para tomar en cuenta el contexto de las palabras.

Las palabras conservadas constituyen las columnas de la tabla léxica, que es una tabla de frecuencias particular. La suma de una fila corresponde a la longitud conservada de la correspondiente respuesta mientras que la suma de una columna es la frecuencia total de la palabra que la encabeza.

3. Hablemos de vinos

3.1. Datos y problemática

En 2003, la agrupación de viticultores del Priorat, en colaboración con INCAVI, quiso experimentar la contribución de los comentarios libres como herramienta de evaluación de los vinos. Así, 31 catadores –ellos mismos viticultores, productores de vinos o enólogos de las bodegas– degustaron 43 vinos provenientes del Priorat. El protocolo de la sesión contemplaba sólo una descripción libre de los vinos, aunque en el marco de una ficha muy precisa (Figura 1). Así, para cada vino, los catadores describían cada uno de los aspectos

clásicos del vino (visual, olfativo y gustativo) mediante las palabras asociadas a cada uno de los atributos mencionados en la ficha.

FITXA	ASPECTE VISUAL		EXAMEN OLFACTIU						EXAMEN GUSTATIU						A QUIN POBLE PERTANY
NÚM CATADOR	COLOR	INTENSITAT	FRANQUESA	INTENSITAT	PERFIL AROMÀTIC	TIPUS D'AROMA	CARÀCTER	QUALITAT	ESTRUCTURA / TEXTURA	TANINS	VOLUM	RETRONASAL	POSTGUST	ARMONIA / EQUILIBRI	
1	Violeta	media	sí	media		?		bona	estruc Media-baja	tanino dulce	medio	vainilla	làctico	mal equilibrio	
												làctico		poco aroma	
2	vermellós	intens	sí	intens		fruits madurs negres	ben integrat	correcte	complex	elegants	ample	llarg	agradable	correcte	
						espècies				madurs					
3	morat	molt intens	sí	poc intens		tancat	complex moderat		potent	secants		fusta	mig secant	no	Gratallops
				2/5		fruita negra						secant		massa secant	
						anís, fusta									

Figura 1. Ficha de comentarios asociados a un vino.

En su momento, no se trató dicha información sino mediante una lectura clásica. Jordi Escayola Mansilla emprendió la tarea de analizar estos datos para su proyecto de fin de carrera de la Diplomatura de Estadística (UPC) (Escayola, 2008).

3.2. Estructura de la tabla múltiple construida

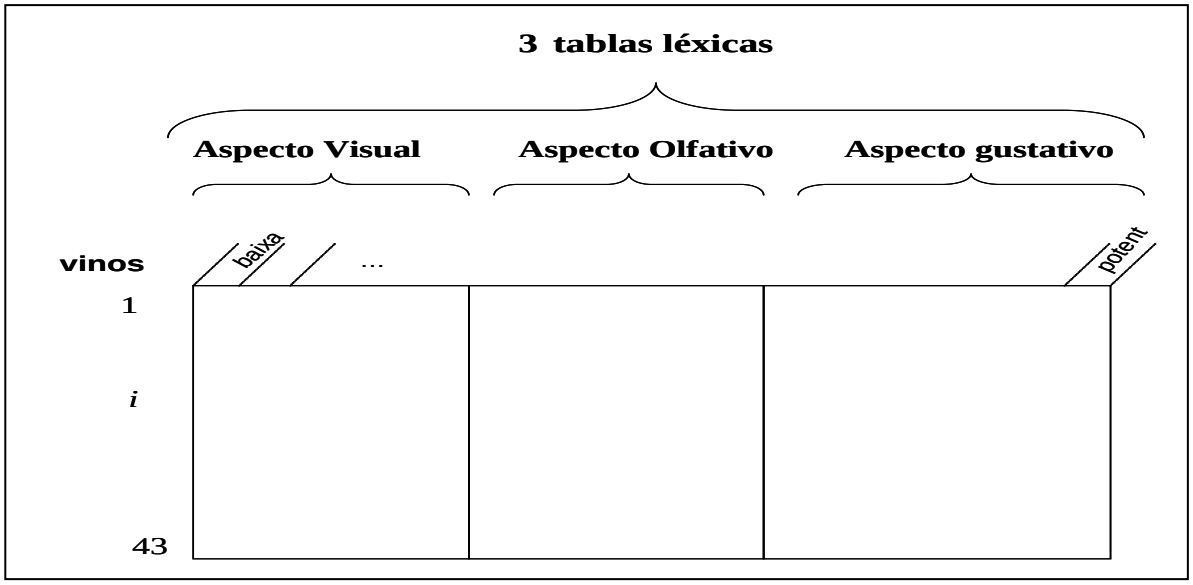

Figura 2. Tablas léxicas resultando de la codificación de las fichas

Se ha conservado sólo la división de las descripciones libres en los 3 aspectos visual, olfativo y gustativo, perdiendo la división más fina en atributos. Así, se consideran 3 variables textuales. A cada variable textual, corresponde un grupo de columnas-palabra.

Para cada uno de los 3 aspectos, se identifican las distintas palabras empleadas para describirlo y se recuenta su frecuencia. Además de lematizar, se reagrupan los sinónimos (por ejemplo, *vermell* y *vermellós* se reagrupan bajo el nombre genérico *vermell*). Después, se conservan las palabras empleadas al menos 4 veces para caracterizar un aspecto.

Así, se obtiene una tabla de frecuencias múltiple con, en fila, los 43 vinos y, en columna, 322 columnas-palabras divididas en 3 grupos (de, respectivamente, 39, 125 y 158 palabras) correspondientes a los 3 aspectos. Una misma palabra puede encabezar varias columnas si se ha empleado para describir diferentes aspectos. Para esto, la etiqueta de una palabra viene precedido de V, O o G según el aspecto que caracteriza.

3.3. Análisis de la tabla léxica múltiple

El análisis de correspondencias (AC) es el método en ejes principales de referencia para el análisis de una tabla de contingencia o de frecuencia (Benzécri 1973, Escofier, 2003; Lebart et al. 2006; Escofier and Pagès, 1988-1998). Su aplicación al análisis textual es frecuente (Benzécri, 1981; Lebart et al. 2000; Murtagh, 2005; Murtagh et al. 2009).

Se pueden aplicar el AC, por separado, a las 3 tablas (Visión, Olfato, Gusto) y comparar las estructuras inducidas sobre los vinos por cada grupo de columnas, pero la comparación es un trabajo arduo y la síntesis de los resultados es compleja. Además, no se dispondría de la estructura "global", es decir inducida globalmente por los 3 aspectos.

El análisis factorial múltiple para tablas de contingencia (AFMTC) (Bécue-Bertaut & Pagès, 1999, 2004) es una herramienta idónea para analizar la tabla léxica múltiple. Dicho método es una extensión del análisis factorial múltiple (Escofier & Pagès, 1988-1998) cuyos principios recordamos primero.

3.3.1. Breve presentación del AFM

El análisis factorial múltiple analiza una tabla múltiple en la cual un conjunto de individuos está descrito por *J* grupos de variables, cuantitativas o cualitativas. El AFM realiza un análisis en componentes principales (ACP) no estandarizado de la tabla yuxtapuesta, pero ponderando las variables de manera a equilibrar los distintos grupos. Por esto, el peso de las columnas-variables del grupo *j* se divide por λ_1^j, primer valor propio obtenido en los análisis separados –ACP o análisis de correspondencias múltiples (ACM) según la naturaleza de las variables– de la subtabla *j*. Dicho método ofrece resultados :

- Análogos a los del ACP o del ACM; principalmente una representación global de las filas (individuos) y de las columnas (variables o categorías) ;
- Específicos de las tablas múltiples : principalmente la representación superpuesta de las estructuras inducidas por cada una de las subtablas sobre el conjunto de los individuos –llamadas estructuras parciales– y la representación de los factores obtenidos en los análisis separados.

3.3.2. AFM para tablas de contingencia múltiples (AFMTC)

Bécue & Pagès (1999, 2004) han propuesto una metodología para el análisis simultáneo de un conjunto de tablas de contingencia. Dicho método parte de los principios del análisis de correspondencias binarias intra-tablas (Benzécri, 1983 ; Escofier & Drouet 1983), generalizada por Cazes & Moreau (1991, 2000) con el nombre de análisis de correspondencias interno (ACI), y , así, toma en cuenta las diferencias entre los márgenes de las filas. Además, se adopta el enfoque del AFM para equilibrar la influencia de las diferentes tablas y para proporcionar gráficos específicos de la estructura en grupos de las columnas. La figura 3 muestra la tabla a analizar y precisa la notación empleada.

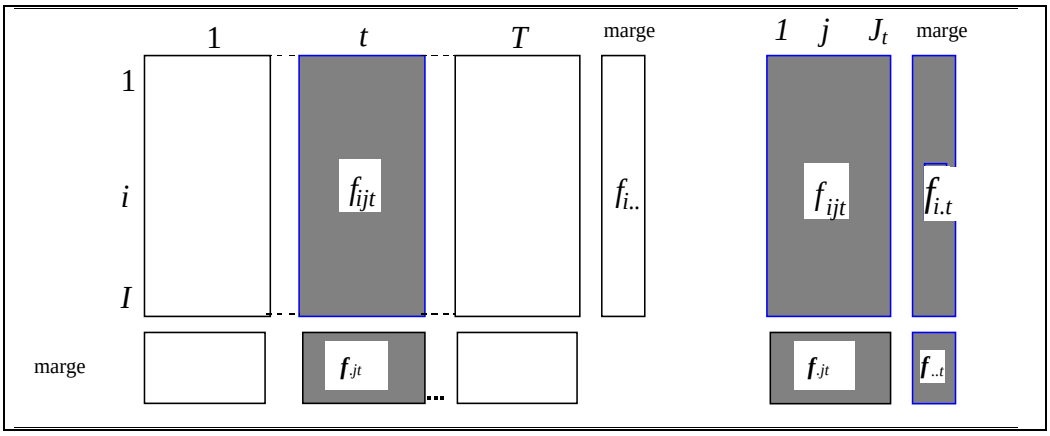

Figura 3. Tabla de contingencia múltiple y márgenes: notación.
En esta tabla, f_{ijt} : frecuencia relativa asociada a la fila i (i=1,…,I) y columna j (j=1,…,J) de la tabla t (t=1,…,T); un índice sustituido por un punto indica la suma sobre este índice.

Cada tabla léxica constituye una tabla de frecuencia, que se puede ver como una tabla de contingencia particular. Para introducirlos en un AFM, se sustituye cada uno por una subtabla de término general y_{ijt} (fila i, columna j, subtabla t) :

$$y_{ijt} = \frac{f_{ijt} - \left(\frac{f_{i.t}}{f_{..t}}\right) \cdot f_{.jt}}{f_{i..} f_{.jt}} = \frac{1}{f_{i..}} \left[\frac{f_{ijt}}{f_{.jt}} - \frac{f_{i.t}}{f_{..t}} \right]$$

El ACP de dicha tabla conduce a los resultados del ACI. Entre los corchetes, se tiene la desviación entre un perfil-columna y el perfil marginal de la tabla a la cual pertenece la columna.

En AFM, el equilibrio entre subtablas es obtenido par la surponderación de cada columna j por $1/\lambda_1^j$, inversa del primer valor propio del análisis separado de la subtabla j.

En el AFMTC, los análisis separados no son los AC usuales pero análisis de correspondencias con un peso impuesto, igual al margen en fila de la tabla yuxtapuesta (análisis llamados seudo-separados). Así, las filas conservan el mismo peso en los distintos análisis. La deformación inherente a las sustitución de los AC separados por los AC seudo-separados, mínima si los márgenes en fila difieren poco de una subtabla a otra, resulta del compromiso necesario a la comparación de tablas de contingencia con diferentes márgenes.

Después, se efectúa el análisis global que consiste en un ACP no estandarizado de la tabla yuxtaponiendo las subtablas de término general y_{ijt} definido antes con:

- Pesos de las filas (y métrica en el espacio de las columnas): $\{f_{i..}$; i = 1,…, $I\}$, $f_{i..}$es el peso relativo medio de los individuos calculado sobre la tabla yuxtaponiendo las tablas de contingencia;
- Pesos de las columnas del grupo t (y métrica en el espacio de las filas): $\{f_{.jt}/\lambda_1^t$; j= 1, …, J_t ; t= 1, …, $T\}$.

Este análisis ofrece resultados:

- Similares a los del AC aplicado a las tablas yuxtapuestas (principalmente, una representación global de los individuos-fila y de las palabras-columna)

- específicos de las tablas múltiples, principalmente, la representación superpuesta de las estructuras inducidas sobre los individuos por cada grupo de columnas -estructuras parciales- y la representación de los factores derivados de los análisis separados.

La lectura de los resultados viene facilitada por numerosas ayudas a la interpretación del AFM.

3.4. Resultados

La tabla yuxtaponiendo las 3 tablas léxicas, correspondiente a los 3 aspectos descritos en la degustación, se ha analizado mediante el AFMTC descrito antes.

3.4.1. Visualización de los vinos-fila y de las palabras-columna sobre el primer plano factorial

Las figuras 4 y 5 muestran la representación de los vinos y de las palabras sobre el primer plano factorial. Debido al enfoque empleado, las palabras son representadas como lo son usualmente las variables en ACP. Otra opción consistiría en representarlas como en AC, en el centro de gravedad de los vinos en la descripción de los cuales se han empleado.

Los 3 grupos de palabras contribuyen de manera equilibrada a la inercia del primer eje. Dicha dirección de dispersión es común a los 3 grupos y no difiere mucho de los primeros ejes factoriales obtenidos en los análisis separados: el primer valor propio del análisis global (2.7) es próximo a su máximo (igual a3 en este caso), que se obtiene cuando los ejes de las análisis separados son idénticos (Escofier & Pagès, 1998, p. 161).

Dicho eje contrasta las palabras indicando defectos de los vinos (*V-baixa, V-ataronjat, O-dolenta, G-animal,* etc.), a la izquierda del eje, con palabras asociadas con los mejores vinos, sobre todo relativas a las característicos olfativas y gustativas (*O-excelent, O-complexe, O-elegant, G-bona*, etc.), a la derecha del eje. Se puede interpretar como un eje de "calidad".

El segundo eje no tiene una interpretación tan clara, aunque se puede ver que, también, refleja la calidad de los vinos.

Nos enfrentamos aquí a una dificultad: en el análisis textual la interpretación se apoya sobre la información suplementaria conocida, aquí ausente.

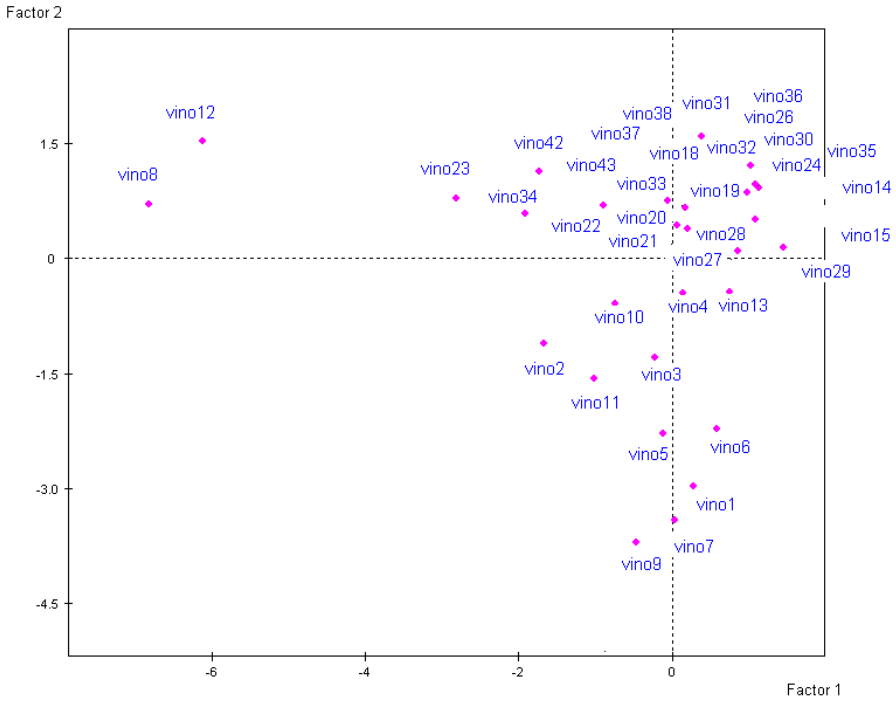

Figura 4. Los vinos sobre el primer plano factorial
Se han eliminado de esta representación algunos de los vinos en la zona central

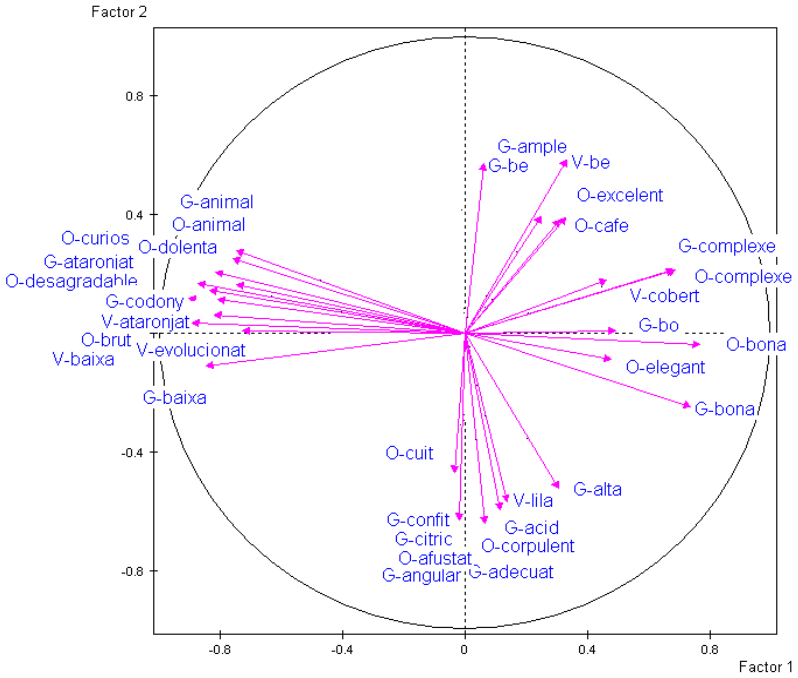

Figura 5. Representación de las palabras más contributivas sobre el primer plano factorial

3.4.2. Superposición de las representaciones parciales y globales

Es posible que un vino no reciba la misma apreciación según el aspecto considerado. Un vino juzgado excelente por sus características olfativas puede, por ejemplo, revelarse decepcionante en boca.

El AFMTC permite representar a cada individuo según el punto de vista global (punto global, como en la figura 4) pero también según los puntos de vista que corresponden a los 3 grupos de columna (puntos parciales) (Escofier & Pagès, 1988-1998; Bécue-Bertaut & Pagès, 2008a).

La desviación entre los puntos parciales correspondientes a un mismo vino se interpretan según el significado dado a los ejes. Así, por ejemplo, el vino 8, de baja apreciación global, lo es sobre todo por sus características visuales mientras que el vino 12 tiene una apreciación más negativa desde el punto de vista olfativo y gustativo que desde sus características visuales. Los vinos destacados en la parte negativa del eje 2, ofrecen ciertas contradicciones entre los aspectos olfativo y gustativo.

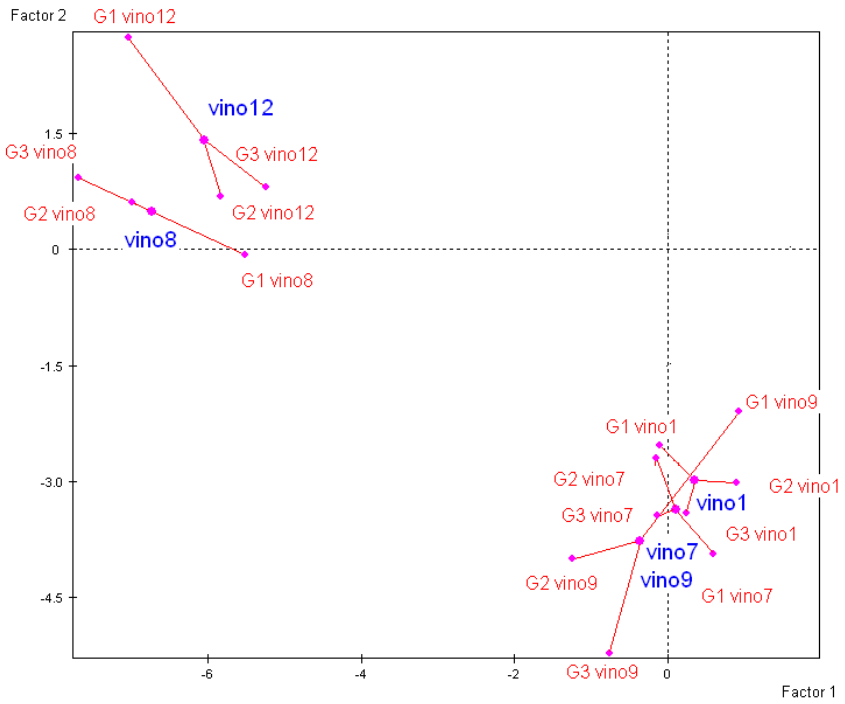

Figura 6. Representación global y parcial de los vinos 8, 12, 9, 7, 1 sobre el primer plano factorial

4. Utilización de información cerrada

En la sección anterior, hemos encontramos dificultades en interpretar los resultados debido a la ausencia de información suplementaria. Otro ejemplo, escogido también en el mundo del vino, permitirá mostrar la aportación de la dicha información

4.1. Ejemplo: Una guía de vinos

El ejemplo trata de una guía dedicada a los vinos de Castilla y León (El Mundo, 2005) que presenta 522 vinos, entre los cuales 364 tintos. Cada vino está evaluado por comentarios de cata y una puntuación (sobre 100) que, para los vinos presentados en esta edición, varía entre 70 et 97. Además, está indicado el precio, la cepa y la añada. Después del pretratamiento de los comentarios, se conservan 222 palabras. La tabla léxica cruza los 364 tintos (filas) y las 222 palabras (columnas).

La tabla 1 muestra algunos comentarios. Se pone de manifiesto, y esto se verifica en el conjunto de la guía, que los vinos mejor puntuados reciben comentarios más extensos. Es un fenómeno clásico en el análisis textual: en una encuesta por cuestionario, más interesado se muestra el entrevistado, más larga es su respuesta; si deseamos saber qué tema interesa más a un político, busquemos las frases más largas en sus intervenciones.

¿Cómo utiliza la información suplementaria, en particular la puntuación? Es el objeto de esta sección.

---- ***Vino 53*** *(*80*)* ***Mesoneros de Castilla-2003***

Tinto tempranillo joven limpio y típico, con fruta de hueso en la nariz. en boca los taninos resultan algo rasos.

---- ***Vino 212*** *(*85*)* ***Legaris-2001***

Tuestes, gominolas y buenos balsámicos marcan la intensidad media frutal de este crianza. En boca aparece muy lineal, con consistencia media; el retrogusto frutal todavía tapado por una madera algo rústica.

---- ***Vino 30*** *(*91*)* ***Tares P3-2001 premium***

Mucho terruño se detecta en el bouquet de este gran tinto; pólvora, sílex, pizarra, cascajo caliente con el contraste de tierra húmeda y mucha fruta madura de hueso. Concentrado, tacto graso sobre el paladar; impresionante viscosidad en la lengua, otra vez impresiones de tierra húmeda y pólvora en el largo final.

---- ***Vino 314*** *(*97*)* ***Vega Sicilia 'Único-1994***

Hay que realizar un ejercicio de disciplina gustativa de primer rango para describir este gran vino. el bouquet es fresco, bien armado de fruta roja que se ve potenciada por tintes de chocolates, tabacos, notas de sotobosque y una madera que se manifiesta pero que resulta difícil de localizar y menos de concretar. Tenemos el caso raro de un tinto que sale ileso del paso del tiempo sin lucir su armadura, que es la barrica. En boca joven, aunque ya tiene su cuerpo vigoroso y enérgico bastante ensamblado, con la excepción de algunos taninos saltamontes que quedan para domesticar. Largo y vibrante final que mezcla madurez con una notable finura fresca.

Tabla 1. Extracto de la guía de vinos

4.2. Introducción de las variables suplementarias en el análisis

4.2.1. Tabla múltiple y notación

Columnas		Tabla léxica		$\sum$	Variable cuantitativa
Filas	*1*	*j*	*J*		
	1	*Proporciones*			
Vinos	*i*	$f_{ij} = \frac{n_{ij}}{n}$		$f_{i.}$	x_i
	I				
$\sum$		$f_{.j}$		*1*	

Figure 7. Tabla múltiple a analizar
La variable cuantitativa está estandarizada para los pesos inducidos por el margen de la tabla léxica ($f_{i.}$)

- n_{ij}: frecuencia de la palabra *j* en el comentario correspondiente al vino *i*;
- *n*: longitud del corpus, es decir, número total de ocurrencias, sumado sobre el conjunto de los comentarios;
- $f_{ij} = \frac{n_{ij}}{n}$: proporción de ocurrencias, sobre la longitud del corpus, de la palabra *j* en el comentario sobre el vino *i*;
- $f_{i.} = \sum_j f_{ij}$: proporción de ocurrencias correspondientes al vino *i*;
- $f_{.j} = \sum_i f_{ij}$: proporción de ocurrencias de la palabra *j*;
- x_i : puntuación dada al vino (*estandarizada);*
- λ_1^J : primer valor propio del AC aplicado a la tabla de frecuencias.

4.2.2. Posicionar las palabras sobre el eje de la nota

Una primera manera de estudiar la relación entre la puntuación (variable cuantitativa) y las palabras (columnas de la tabla léxica) consiste en emplear la "media ponderada" (weighted averaging, Orlaci, 1975). Cada palabra se posiciona mediante la media ponderada de las puntuaciones recibidas por los vinos cuyos comentarios contienen dicha palabra. La ponderación corresponde a la importancia relativa de la palabra en el comentario del correspondiente vino. Así, a la palabra *j* corresponde el valor $\bar{x}_j$, obtenido mediante el siguiente cálculo:

$$\bar{x}_j = \sum_{i=1}^{I} \frac{f_{ij}}{f_{.j}} x_i$$

La figura 8 muestra la gráfica resultante. Además, se han posicionado los vinos con mejor puntuación. Esta figura ordena las palabras desde la más negativa, (en el contexto de una guía que sólo menciona vinos con una calidad aceptable) a la más positiva, desde *corto, correcto y herbaceo* a *enérgico, gran y denso*.

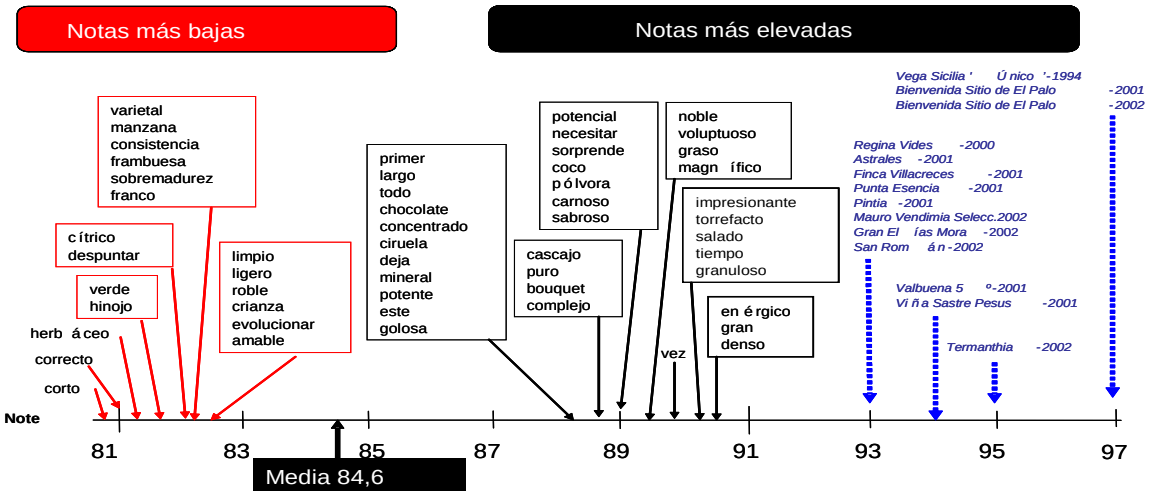

Figura 8. Posicionamiento de las palabras sobre el eje de las notas

4.2.3. *Posicionar las palabras sobre el primero plano factorial proporcionado por el AC de la tabla léxica*

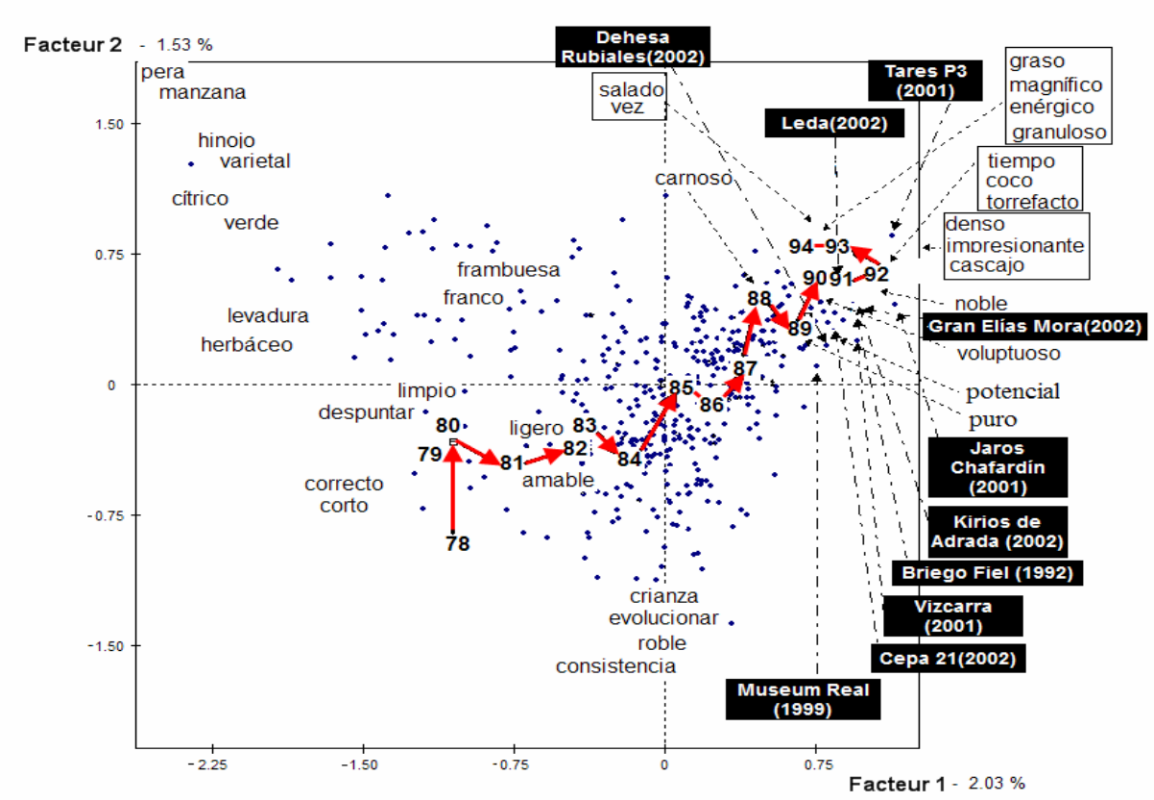

Figura 9. Primer plano factorial proporcionado por el AC de la tabla Vinos ×Palabras Posicionamiento de los diferentes valores de la nota considerados como categorías suplementarias

Otra manera de relacionar la tabla léxica y la puntuación consiste en efectuar el AC de la tabla léxica para después posicionar la puntuación como suplementaria. La posición de dicha puntuación sobre un eje corresponde a su correlación con el eje. Se observa una alta correlación entre la nota y el primer eje de AC (.70), lo que significa dicho eje expresa la nota en gran parte. El primer eje como un eje de calidad, aunque pueda existir una dirección de dispersión en la nube de vinos inducidas por tabla léxica más relacionada con la puntuación. La siguiente sección proporciona una herramienta para identificar dicha dirección.

4.2.4. Dar un papel simétrico activo a las palabras y a la puntuación, considerados como dos maneras de evaluar los vinos

Se construye la tabla múltiple, yuxtaponiendo la tabla léxica (con J columnas-palabra) y una columna con la puntuación dada a los vinos. Se analiza dicha tabla mediante el AFM, tratando la tabla léxica como en AFMTC. Este AFM extendido es equivalente a realizar un ACP ponderado:

- Sobre la tabla yuxtaponiendo la tabla, construida a partir de la tabla léxica, de término general $\left(f_{ij} - f_{i.} \cdot f_{.j}\right) / \left(f_{i.} \cdot f_{.j}\right)$ y la columna cuantitativa (centrada o centrada y estandarizada.
- Dando a las filas el peso inducido por el AC (es decir, $\{f_{i.}\ ;\ i = 1,\ldots, I\}$);
- Dando a las columnas-palabra los pesos inducidos por el AC (es decir, $\{f_{.j}\ ;\ j = 1, \ldots, J\}$) divididos por el primer valor propio obtenido en el AC separado de la tabla léxica – notado $\lambda_1^{\ J}$ – y a la columna cuantitativa un peso igual a 1.

Este AFM extendido proporciona los resultados clásicos del ACP, así como herramientas para comparar las diferentes estructuras inducidas sobre los vinos por los comentarios, por una parte, y la puntuación, por otra parte.

La figura 10 presenta el primer plano factorial inducido por este AFM. Se posicionan los diferentes valores de la puntuación como categorías suplementarias, es decir, en el centro de gravedad de los vinos que han puntuados con este valor.

El primer eje está muy fuertemente correlacionado con la nota (0.95). Como era de esperar, pocas palabras son francamente negativas: una guía de vinos sólo menciona vinos recomendables. Los aspectos hedónicos del vino se subrayan (*impresionante, esplendido*, a la derecha del eje, opuestos a *agradable, amable, medio*, a la izquierda). Lo que se suele llamar "la potencia del vino" parece ser un criterio dominante en la evaluación: los adjetivos *denso, graso, concentrado* y *largo* (*en boca*), a la derecha del eje, se oponen a *ligero* o *corto* a la izquierda. Algunos términos conciernen defectos (*sequedad, sobremaduro, evolucionar);*, se encuentran en la parte izquierda el primer eje, indicando que cualifican los vinos peor puntuados.

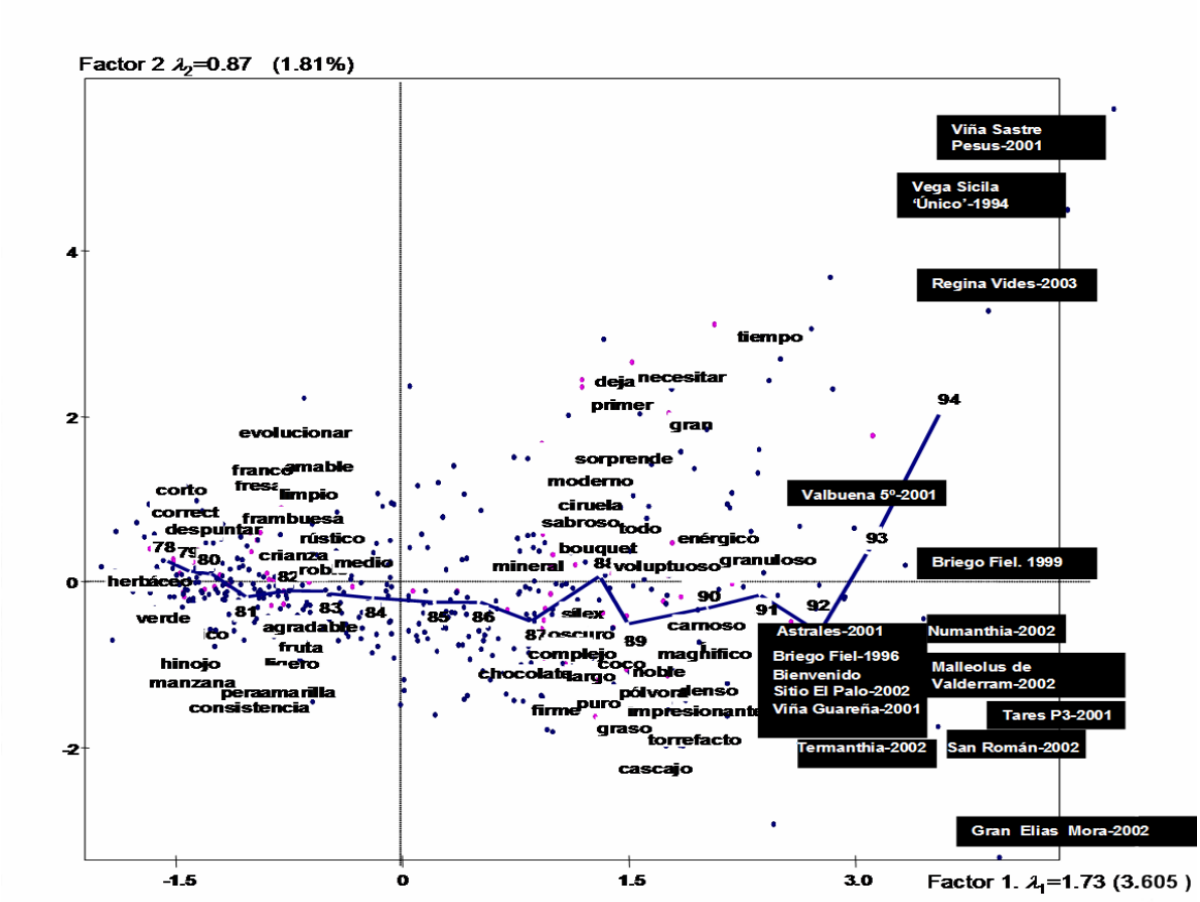

Figura 10. Primer plano factorial proporcionado por el AFM aplicada a la tabla yuxtaponiendo la tabla léxica y la columna con la puntuación

Algunas palabras – como *fácil, agradable* y *franco*–, consideradas palabras positivas en el lenguaje usual, no lo son tanto aquí, como lo indica su posición en la parte izquierda del eje. No es que los autores de la guía piensen que un vino excelente debe ser *desagradable, difícil* y *no franco*, sino que oponen dichos términos a *complejo* o *sorprendente*. Apreciar el vino es también disfrutar de un punto de vista intelectual.

La trayectoria de la puntuación es bastante regular a lo largo del primer eje. El ordenamiento de las categorías no está alterado: el vocabulario presenta realmente una dimensión valorativa, bien expresada por este eje. Una validación de tipo bootstrap, no presentada aquí, ha permitido validar este hecho.

Segundo eje

El segundo y siguientes ejes presentan correlaciones bajas o nulas con la puntuación de los vinos. Además, no están correlacionados, o muy poco, con las otras características como la añada o el precio, estas ultimas variables utilizadas como suplementarias en el análisis. Expresan la parte de la variabilidad del vocabulario no relacionado con dichas variables.

Los vinos con coordenadas positivas (respectivamente, negativas) sobre el segundo eje tienden a presentar comentarios más largos (respectivamente, más cortos) que lo esperado debido a su puntuación.

En el caso de los vinos mejor puntuados (parte positiva del primer eje), el segundo eje opone los vinos cualificados como *complejos, voluptuosos, sorprendentes*, etc. –que sugieren comentarios más largos– a los vinos *potentes, impresionantes, torrefactos, densos*, etc., más clásicos.

En el caso de los vinos con coordenadas negativas sobre el primer eje – por tanto, con puntuaciones bajas – el segundo eje destaca tres vinos considerados muy particulares (*Mesoneros de Castilla*-2003, *Valdelosfriales*-2003 y *Fuentenarro*-2002). Dichos vinos son *jóvenes, típicos* y con pobres características *en boca*.

Discordancias entre los comentarios libres y la puntuación del vino

Aunque exista una muy fuerte correlación entre los comentarios libres y la puntuación, algunos vinos muestran comentarios más laudatorios (respectivamente, menos laudatorios) que su puntuación dejaba esperar. Esta discordancia está puesta de relieve mediante la representación superpuesta –no reproducida aquí – de las dos estructuras parciales. Por ejemplo, el punto parcial "comentarios libres" de *Tares P3-2001 premium* obtiene una coordenada mucho más positiva sobre el primer eje que el punto parcial "puntuación" correspondiente al mismo vino. De hecho, los comentarios de este vino contienen 8 de las 20 palabras de mayor coordenada sobre el primer (*impresionante, grande, a la vez, graso, cascajo, pólvora, largo y impresión*), una de ellas utilizada dos veces (*pólvora*).

Los lectores interesados encontrarán más información en Bécue-Bertaut & Pagès (2008b).

5. Un experimento en curso: *parlem de vins en català i en francès*!

La interpretación de los datos de cata presentados en la sección 3 sufría de la ausencia de información suplementaria. No obstante, el estudio mostró que la descripción libre de los vinos comportaba una dimensión valorativa importante que conviene aprovechar.

Por esta razón, se decidió:

- poner a prueba la contribución de las descripciones libres en sensometría, es decir, en la recogida y análisis de datos sensoriales mediante métodos estadísticos.
- Comparar las evaluaciones dadas por degustadores catalanes y franceses a un mismo conjunto de vinos.

Esta experiencia se realizó con éxito en la *Escola d'Enologia* (Espiells) y en el *Conseil Interprofessionel des Vins du Rousssillon*-CIVR (Perpiñan)

Cada sesión comportaba 2 ejercicios de cata distintos sobre los mismos 8 vinos catalanes:

- En la primera parte, cada degustador debía formar "clases de vinos" según la similitud percibida entre ellas. Los criterios de similitud eran los propios, sin que puedan existir buenas o malas respuestas. Después, las clases se apuntaban en la hoja proporcionada al degustador. Además, se pedía a cada uno que asocie a cada grupo las palabras que caracterizaban bien al conjunto del grupo.

- En la segunda parte, los mismos degustadores rellenaban una ficha de cata, en la cual debían describir los aspectos visual, olfativo y gustativo de cada vino; así rellenaban una ficha simplificada en comparación con la ficha empleada en la cata de 2003.

Además, se dispone de las características químicas de los vinos.

Los datos están capturados y en fase de depuración. La metodología presentada en la sección 4 permitirá analizar estos datos complejos, inclusión las descripciones libres escritas o en catalán o en francés sin proceder a ninguna traducción. Las proximidades entre palabras inducidas por el análisis podrán sugerir equivalencias entre palabras.

6. Conclusiones

La codificación de los textos respeta tres puntos esenciales: estabilidad, legibilidad y reproductibilidad. Dicha codificación, relativamente fácil de implementar gracias a las herramientas elaboradas en la industria del lenguaje natural, asegura una codificación independiente del codificador y facilita la comparación entre estudios.

La metodología presentada permite analizar simultáneamente comentarios libres y variables cuantitativas y/o categóricas; así constituye una herramienta útil de numerosas áreas. En general:

- Las preguntas cerradas/ las puntuaciones aportan un marco sólido, pero tienden a reproducir las asunciones que han guiado la concepción del cuestionario cuando se analizan solas;
- Las preguntas abiertas/ los comentarios libres proporcionan información rica pero que difícilmente se puede analizar por sí sola.

Tonar en cuenta los dos tipos de cuestionamiento proporciona una herramienta que adiciona sus respectivas ventajas a la vez que limita los defectos propios de cada uno. Es obvio que el procedimiento presentado no permite capturar toda la información contenida en los comentarios o respuestas libres que constituyen un material muy complejo. No obstante, esta pérdida de información aporta a cambio una ganancia en comprensión, lo que es un principio básico de la estadística.

Acknowledgments

This work has been partially supported by the Spanish Ministry of Education and Science, FEDER (grant SEJ2005-00741/ECON) as well as the Catalan Commission for Universities DURSI (grant SGR 00004/2005).

Agradecemos a Joan Miquel Canal, director del Parc Científic de les Industries Enològiques de Falset habernos facilitado los datos de la cata del Priorat de 2003.

7. References

Bécue-Bertaut, M., Pagès, J., (1999). Intra-sets multiple factor analysis. Application to textual data. In: Bacelar-Nicolau, H., Costa Nicolau, F., Janssen, J. (eds.), *Applied Stochastic Models and Data Analysis*. INE, Lisbon, 72-79.

Bécue-Bertaut, M., Pagès, J., (2004). A principal axes method for comparing contingency tables: MFACT. *Comput. Statist. Data Anal.* 45(3), 481-503.

Bécue-Bertaut, M., Pagès, J., (2008a). Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data. *Comput. Statist. Data Anal* 52(6) *3255-3268.*

Bécue-Bertaut, M., Álvarez Esteban R., Pagès (2008b) Rating of products through scores and free-text assertions. Comparing and combining both. *Food Quality and Preference* 19(1) 122-134.

Benzécri, J.P. (1973). *Analyse des Données, Vol.* 2: *Analyse des correspondances*. Paris: Dunod.

Benzécri, J.P. (1981). *Pratique de l'analyse des données, Vol.* 3, *Linguistique & Lexicologie*. Paris: Dunod.

Benzécri, J.P., (1983). Analyse de l'inertie intraclasse par l'analyse d'un tableau de contingence. Les Cahiers de l'Analyse des Données 8(3), 351-358.

El Mundo (2005). *Guía de catas 2005. Vinos de Castilla y León*. Valladolid: Biblioteca La Posada.

Cazes, P., Moreau, J., (1991). Analysis of a contingency table in which the rows and the columns have a graph structure. In: Diday, E., Lechevallier, Y. (Eds), *Symbolic-numeric data analysis and learning*, Nova Science Publishers, New York, 271-280.

Cazes P., Moreau, J., (2000). Analyse des correspondances d'un tableau de contingence dont les lignes et les colonnes sont munies d'une structure de graphe bistochastique. In: Moreau, J., Doudin, P.A., Cazes, P. (Eds), *L'analyse des correspondances et les techniques connexes. Approches nouvelles pour l'analyse statistique des données,* Springer, Berlin-Heidelberg, 87-103.

El Mundo (2005). *Guía de catas 2005. Vinos de Castilla y León*. Valladolid: Biblioteca La Posada.

Escayola J. (2008). Sensometría. Aportación a las descripciones libres. Projecte de fi de carrera de la Diplomatura de Estadística de la UPC.

Escofier, B. (1983). Analyse de la différence entre deux mesures sur le produit de deux mêmes ensembles. *Les Cahiers de l'Analyse des Données* 8, 325-329.

Escofier, B. (2003). *Analyse des correspondances*. Rennes: Presses Universitaires de Rennes.

Escofier, B., Drouet, D. (1983) Analyse des différences entre plusieurs tableaux de fréquence. *Les Cahiers de l'Analyse des Données* 8(4), 491-499.

Escofier, B., & Pagès, J. (1988-1998) *Analyses factorielles simples et multiples*. Paris: Dunod.

Labbé, D. (1990) *Normes de saisie et de dépouillement des textes politiques*. CERAT Cahier nº 7. http://web.upmf-grenoble.fr/cerat/Recherche/PagesPerso/LabbeNormes.pdf

Lebart, L. (2003) Analyse des données textuelles. In G. Govaert, *Analyse des données* (pp. 151-168). Paris: Lavoisier.

Lebart, L., Salem, A., Bécue M. (2000). *Análisis estadístico de textos*. Milenio.

Lebart, L., Piron, M., & Morineau, A. (2006) *Statistique exploratoire multidimensionnelle. Visualisation et inférence en fouille de données*. Paris: Dunod.

Lehrer, A. (1975) Talking about Vino. *Language, 51*(4), 901-923.

Muller, Ch. (1977-1992) *Principes et méthodes de statistique lexicale*, Paris: Larousse; reprint Genève: Champion-Slatkine.

Murtagh, F. (2005). *Correspondence Analysis and Data Coding with R and Java*. Chapman & Hall, CRC Press, London

Murtagh, F., Ganz, A., McKie, S. (2009) The structure of narrative: The case of film scripts. *Pattern Recognition* 42(2), 302-312.

Orlaci (1975). *Multivariate analysis in vegetation research*. Junk. La Hague.

Salton, G., & McGill, M.J. (1983). *Introduction to modern information retrieval*. McGraw-New York: Hill Book Company.

ten Kleij, F., & Musters, P.A.D. (2003) Text analysis of open-ended survey responses: a complementary method to preference mapping. *Food quality and preference, 14*, 43-52.