*27 Congreso Nacional de Estadística e Investigación Operativa*
*Lleida, 8–11 de abril de 2003*

# CLASIFICACIÓN NO PARAMÉTRICA DE DATOS COMPOSICIONALES

J. A. Martín-Fernández[1], V. Pawlowsky-Glahn[1], C. Barceló-Vidal[1].

[1]Departament de Matemàtica i Informàtica Aplicada
Universitat de Girona, E-17071 Girona, Spain
E-mail: josepantoni.martin@udg.es; vera.pawlowsky@udg.es; carles.barcelo@udg.es

**RESUMEN**

Se ha constatado la inexistencia de un cuerpo teórico y metodológico apropiado que permita desarrollar pautas y recomendaciones a seguir en el momento de realizar una clasificación no paramétrica de datos composicionales. Presentamos la metodología a aplicar en la realización de una clasificación automática de datos composicionales contemplando medidas de tendencia central, de dispersión y de disimilitud apropiadas para la tipología específica de estos datos. Mostramos sobre un conjunto de datos real los resultados que se obtienen al aplicar esta metodología.

**Palabras y frases clave**: cluster, diagrama ternario, distancia de Aitchison, símplex.
**Clasificación AMS**: 62H30; 62Pxx.

# 1 Introducción

El interés de una clasificación radica fundamentalmente en descubrir, analizar e interpretar la estructura de los datos. Aplicando esta técnica puede obtenerse una reducción del número de datos de la muestra asimilando cada individuo al representante de cada grupo, habitualmente el centroide y, además, la clasificación, puede dar lugar a un análisis estadístico e interpretación de las características de cada grupo por separado.

En esta comunicación, por motivos de brevedad, analizamos únicamente el caso de las técnicas de clasificación jerárquicas. Somos conscientes que realizando unas sencillas adaptaciones se consigue extender nuestras conclusiones al caso de las técnicas jerárquicas descendentes u otras técnicas no paramétricas de clasificación.

El proceso de la mayor parte de los diferentes tipos de clasificación no paramétrica puede plasmarse en un esquema como el siguiente:

INDIVIDUOS $\Longrightarrow$ ELECCIÓN de la MEDIDA DE DIFERENCIA $\Longrightarrow$
ELECCIÓN del MÉTODO DE CLASIFICACIÓN $\Longrightarrow$ GRUPOS

Este planteamiento comporta, como paso previo a la aplicación de las técnicas de clasificación automática no paramétrica, la necesidad de establecer una o varias de las siguientes medidas:

1. Una medida de diferencia entre dos datos.
2. Una medida de tendencia central de un conjunto de datos.
3. Una medida de dispersión de un conjunto de datos.

La medida de diferencia entre dos datos nos ha de permitir asignar individuos similares o cercanos a un mismo grupo, e individuos diferentes o alejados a grupos diferentes. Entre las técnicas de clasificación jerárquica ascendente se encuentran algunas técnicas que sólo requieren tener definida una medida de diferencia. Este es el caso de los métodos del *máximo*, del *mínimo* y de la *media*. Otras técnicas, como el método del *centroide*, requieren tener establecida, además, una medida de tendencia central. Finalmente, existen otras técnicas que requieren adicionalmente una medida de dispersión. Entre éstas se encuentra el método de Ward. En todo caso, un hecho relevante es que todas estas medidas deben ser establecidas teniendo en cuenta las características matemáticas del soporte de los datos a clasificar.
El espacio soporte asociado a los datos composicionales es el símplex:

$$\mathcal{S}^D = \{[x_1, \dots, x_D] : x_i > 0 \ (i = 1, \dots, D);\ x_1 + \cdots + x_D = 1\}. \tag{1}$$

Las operaciones básicas definidas en el símplex son la *perturbación* y la *potenciación* simbolizadas, respectivamente, por $\oplus$ y $\otimes$, y definidas mediante las expresiones

$$\mathbf{x} \oplus \mathbf{y} = \left[\frac{x_1 y_1}{\sum x_k y_k}, \dots, \frac{x_D y_D}{\sum x_k y_k}\right], \tag{2}$$

y

$$a \otimes \mathbf{x} = \left[\frac{x_1^a}{\sum x_k^a}, \dots, \frac{x_D^a}{\sum x_k^a}\right], \tag{3}$$

donde $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$, y $a$ es un número real. Es trivial establecer que la operación interna $\oplus$ y la operación externa $\otimes$ dotan al símplex $\mathcal{S}^D$ de una estructura de espacio vectorial. En consecuencia, las medidas que se deben establecer deberán tener en cuenta la existencia de esta estructura si se desea que las medidas sean compatibles con la tipología de los datos. Por lo que se refiere a la medida de diferencia, será deseable que la disimilitud entre dos datos se conserve cuando a éstos se les aplica la misma perturbación. En relación a la medida de tendencia central, será deseable que

el centro de un conjunto despúes de haberle aplicado una perturbación coincida con el resultado que se obtenga al aplicar la misma perturbación al centro del conjunto sin perturbar. Respecto a la medida de variabilidad será deseable que esta medida sea invariante por la operación perturbación. Las características matemáticas del soporte de los datos composicionales y sus medidas adecuadas han sido analizadas en profundidad en muchos trabajos, entre los que destacamos los de Barceló-Vidal et al. (2001, 2003), Aitchison et al. (2000), y Martín-Fernández et al. (1998a, 1998b).

En las siguientes secciones se presenta la metodología a aplicar para realizar una clasificación jerárquica aglomerativa de datos composicionales indicando las medidas de tendencia central, de dispersión y de disimilitud que son compatibles con la tipología específica de este tipo de datos. En la tercera sección aplicamos esta metodología al análisis de un conjunto real de datos sobre la población ocupada de las diferentes comarcas catalanas. Este mismo conjunto de datos fue motivo de estudio en Vives y Villarroya (1996) donde la clasificación se realizó utilizando la disimilitud de Bhattacharyya (arccos) y el método de agrupación jerárquico aglomerativo de la media. Para finalizar nuestro trabajo, comparamos nuestros resultados con los resultados presentados en Vives y Villarroya (1996).

# 2 Metodología propuesta

En una clasificación automática las fases de elección de la medida de disimilitud y de elección del método de clasificación desempeñan un papel crucial. Por lo que se refiere a la elección de la medida de disimilitud, la idea clave a tener en cuenta es que una disimilitud puede ser adecuada o no dependiendo de la tipología de los datos a clasificar. No existe una disimilitud adecuada para todos los tipos de datos, y, en general, para cualquier tipo de datos puede encontrarse más de una medida de disimilitud que sea adecuada. En la fase de elección del método de clasificación debe decidirse en primer lugar qué tipo de técnica no paramétrica se va a utilizar: jerárquica o no jerárquica. La decisión se toma fundamentalmente en base a si se conoce o no el número de grupos a construir. De nuestra experiencia en la realización de clasificaciones, se desprende que en la gran mayoría de los estudios se desconoce a priori el número de grupos a considerar, y en consecuencia, las técnicas más utilizadas son las jerárquicas.

La Figura 1 muestra de manera esquemática las fases a seguir en la realización de una clasificación automática no paramétrica de datos composicionales mediante un método jerárquico. Si el método de clasificación no fuese jerárquico el esquema seguiría siendo válido suprimiendo la etapa intermedia de elección del número de grupos a considerar. Como puede apreciarse, este esquema no es únicamente válido para datos de tipo composicional. Las particularidades a tener en cuenta para el caso de datos composicionales las exponemos en las secciones siguientes. En la Figura 1 se observa que la realización de una clasificación se basa en un proceso de naturaleza inductiva-deductiva. La naturaleza de este proceso es común a la gran mayoría de

técnicas estadísticas y está en el fundamento del propio método estadístico. En la realización de una clasificación, la etapa de diagnosis o crítica de resultados consiste en analizar si la agrupación obtenida puede considerarse razonable. En este contexto, entendemos que una clasificación razonable es aquélla en la que observaciones que pertenezcan a grupos diferentes muestren un patrón claramente diferenciado en el valor que toman en las diferentes variables. Este patrón diferenciador de los grupos obtenidos deberá ser interpretable en relación al contexto o población de la que haya sido extraído el conjunto de los datos. Si la clasificación no se considera razonable el proceso iterativo-deductivo contempla la posibilidad de modificar la elección de la medida de disimilitud, la elección del método de clasificación o, en su caso, la elección del número de grupos a considerar.

## 2.1 Medida de disimilitud

Somos conscientes que el hecho de utilizar una medida coherente con la tipología de los datos no es garantía *sine qua non* para obtener una clasificación razonable para cualquier conjunto de datos. Por otro lado, también somos conscientes que, en ciertos casos, es posible obtener una clasificación razonable incluso utilizando una medida de disimilitud incoherente con la tipología de los datos. Sin embargo, pensamos que estos casos son inusuales, y que, con mayor probabilidad, el hecho de realizar una agrupación usando una medida inadecuada nos llevará a obtener resultados erróneos y clasificaciones poco verosímiles. En unos primeros trabajos iniciales se mostró (Martín-Fernández et al., 1998a, 1998b) que las medidas de disimilitud más usuales no son coherentes con la naturaleza composicional de los datos. En estos mismos trabajos se presentó la distancia de Aitchison $d_a$ cuya expresión al cuadrado es

$$d_a^2(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{D} \left\{ \ln \frac{x_i}{g(\mathbf{x})} - \ln \frac{y_i}{g(\mathbf{y})} \right\}^2, \tag{4}$$

como una medida de disimilitud adecuada por ser compatible con las operaciones básicas –véanse las expresiones (2) y (3)– definidas en el símplex. El término $g(\mathbf{x})$ que aparece en (4) representa la media geométrica del vector composicional $\mathbf{x}$. En Martín-Fernández et al. (1998c) se propuso otra medida de disimilitud coherente con la operación interna perturbación (2): la medida de Kullback-Leibler composicional $d_{\mathcal{KL}}$ (al cuadrado)

$$d_{\mathcal{KL}}^2(\mathbf{x},\mathbf{y}) = \frac{D}{2} \ln \left( A(\frac{\mathbf{x}}{\mathbf{y}}) \cdot A(\frac{\mathbf{y}}{\mathbf{x}}) \right), \tag{5}$$

donde $A(\mathbf{x}/\mathbf{y})$ representa la media aritmética del vector de ratios $\mathbf{x}/\mathbf{y}$.

En consecuencia, en esta fase de la clasificación, la elección de la medida de disimilitud consistirá en decidir si se utiliza la distancia $d_a$ o la disimilitud $d_{\mathcal{KL}}$. Por

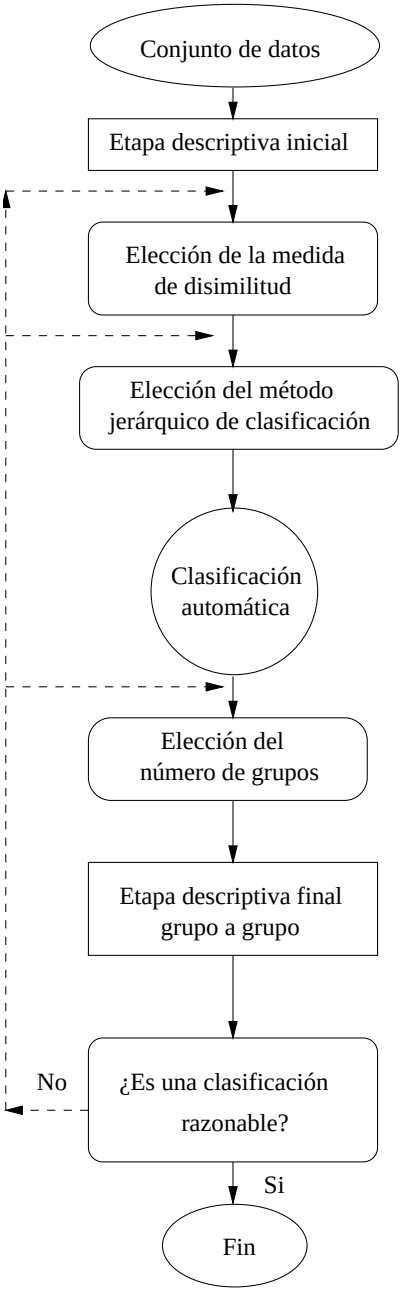



Figura 1: Esquema de las fases a seguir en la realización de una clasificación automática no paramétrica de datos composicionales.

motivos de brevedad, en el caso práctico que presentamos en este trabajo escogeremos la distancia de Aitchison (4) por ser la más utilizada. Es necesario remarcar que se demostró que las dos medidas, $d_a$ y $d_{\mathcal{KL}}$ están relacionadas a partir de una función monótona.

## 2.2 Medidas de tendencia central y de dispersión

Las medidas de tendencia central constituyen, junto con las medidas de diferencia y de dispersión, el elemento diferenciador clave entre las técnicas habituales de clasificación automática no paramétrica. Es bien sabido que la medida de tendencia central más utilizada para conjuntos de datos en el espacio real es la media aritmética o centroide del conjunto. También es conocido que esta medida, tan usualmente aplicada, no es coherente con el carácter composicional de los datos (Martín-Fernández et al., 1998b). La estructura algebraica del espacio soporte de los datos composicionales conduce a proponer el uso de la *media geométrica composicional* como medida representativa del centro de un conjunto de datos composicionales puesto que es compatible (Martín-Fernández et al., 1998b) con las operaciones (2) y (3), y con la distancia de Aitchison (4):
Si $\mathbf{X} = [\mathbf{x}_1; \mathbf{x}_2; \dots ; \mathbf{x}_N]$ es un conjunto de datos composicionales, entonces la *media geométrica composicional* $\hat{\xi}$ del conjunto $\mathbf{X}$ es

$$\hat{\xi} = \left[ \frac{g_1}{\sum g_k}, \frac{g_2}{\sum g_k}, \dots , \frac{g_D}{\sum g_k} \right], \tag{6}$$

donde $g_k = \left( \prod_{i=1}^{N} x_{ik} \right)^{1/N}$ representa la media geométrica de la $k$-ésima parte de los datos.

Es bien sabido que una de las medidas de dispersión más utilizada para conjuntos de datos en el espacio real es la traza de la matriz de covarianzas asociada al conjunto. Esta medida de variabilidad, que es compatible con la distancia euclídea y con el centroide del conjunto, no es adecuada para el caso de los datos composicionales. La medida de variabilidad que posee buenas propiedades (Martín-Fernández et al., 1998b) en relación a las características matemáticas del espacio soporte es la *medida de variabilidad total*, o *totvar* :
Si $\mathbf{X} = [\mathbf{x}_1; \mathbf{x}_2; \dots ; \mathbf{x}_N]$ es un conjunto de datos composicionales, entonces la *variabilidad total* o $totvar(\mathbf{X})$ del conjunto $\mathbf{X}$ es

$$totvar(\mathbf{X}) = \frac{1}{N-1} \sum_{i=1}^{N} d_a^2(\mathbf{x}_i, \hat{\xi}) = \frac{1}{N(N-1)} \sum_{i<j}^{N} d_a^2(\mathbf{x}_i, \mathbf{x}_j), \tag{7}$$

donde $\hat{\xi}$ es la media geométrica composicional del conjunto $\mathbf{X}$.

# 3 Un caso práctico: *Población ocupada por grupos profesionales*

El conjunto de datos *Población ocupada por grupos profesionales* ha sido motivo de un estudio detallado en el trabajo de Vives y Villarroya (1996). Este conjunto de datos, que simbolizamos por $\mathbf{X}$, está formado por la observación sobre 41 unidades muestrales. Cada unidad corresponde a una de las 41 comarcas en que se encuentra dividida Catalunya en el censo del año 1991 –véase la Figura 2.

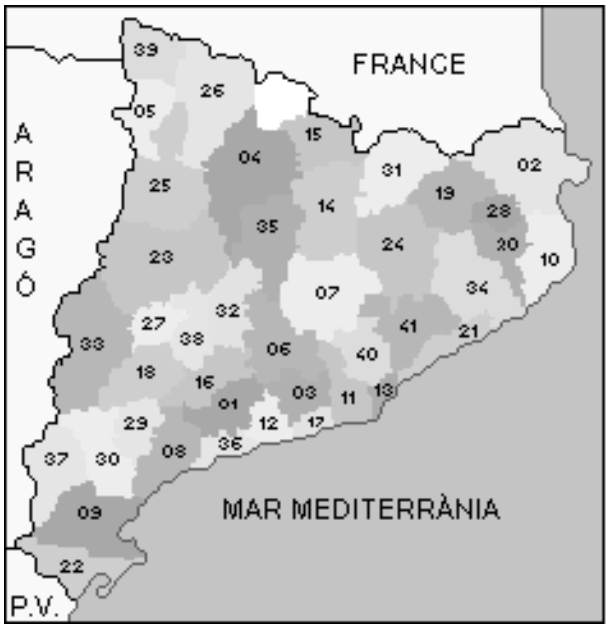



Figura 2: Mapa de las comarcas de Catalunya. (Fuente:*Generalitat de Catalunya*)

De cada una de las comarcas se observó la proporción de la población activa en cada uno de los 8 grupos profesionales siguientes:

| | |
|---|---|
| $\mathbf{X}_1$: *Profesionales y técnicos*; | $\mathbf{X}_2$: *Personal directivo*; |
| $\mathbf{X}_3$: *Servicios administrativos*; | $\mathbf{X}_4$: *Comerciantes y vendedores*; |
| $\mathbf{X}_5$: *Hostelería y otros*; | $\mathbf{X}_6$: *Agricultura y pesca*; |
| $\mathbf{X}_7$: *Industria*; | $\mathbf{X}_8$: *Fuerzas armadas.* |

Cada uno de estos grupos profesionales es una variable o columna del conjunto de datos. Por lo tanto, el conjunto $\mathbf{X}$ está formado por 41 observaciones, $\mathbf{x}_i$, $i = 1, 2, \ldots, 41$, en el espacio $\Re^8_+$.
El objetivo del estudio consiste en realizar una clasificación automática no paramétrica de las 41 comarcas catalanas que permita analizar la existencia de grupos de comarcas que sean similares en relación a la distribución de su población activa.

## 3.1 Aplicación de la metodología propuesta

Este mismo conjunto de datos fue motivo de estudio en Vives y Villarroya (1996) donde la clasificación se realizó utilizando la disimilitud de Bhattacharyya (arccos) y el método de agrupación jerárquico aglomerativo de la media. En los resultados presentados los autores consideraban razonable una clasificación de las 41 comarcas catalanas en 4 grupos diferentes: el grupo *Agrícola*, compuesto por 19 comarcas; el grupo *Turístico*, formado únicamente por la comarca de la Val d'Aran; el grupo *Administrativo*, formado únicamente por la comarca del Barcelonès; y el grupo *Industrial*, compuesto por 20 comarcas. Los autores Vives y Villarroya (1996) destacan en su trabajo que han querido dar el mismo peso a todas las comarcas. Esta decisión responde a su interés en estudiar las relaciones y las características de las comarcas, independientemente del total de población activa de cada comarca. En este sentido los autores resaltan que una alternativa habría sido la aplicación de las técnicas del Análisis de Correspondencias (AC). Sin embargo, es bien conocido que los resultados obtenidos mediante el AC se ven afectados por los tamaños muestrales como consecuencia de que las técnicas del AC se basan en la distancia $\chi^2$. Este planteamiento fue el que nos indujo a realizar el estudio del conjunto de datos *Población ocupada por grupos profesionales* utilizando la metodología propuesta para los datos composicionales. Entre los diferentes métodos de clasificación automática no paramétrica escogemos para nuestro estudio los jerárquicos aglomerativos. Esta elección se basa en la intención de usar métodos de la misma familia que el método jerárquico de la media, utilizado en el trabajo de Vives y Villarroya (1996).

En el Cuadro 1 se muestran los valores de tres coeficientes habitualmente utilizados en este tipo de clasificaciones: el coeficiente de correlación cofenética, el índice de Mojena, y el índice de Calinski. Estos tres índices han sido calculados para cada uno de los cinco métodos aglomerativos de clasificación. Recordemos que el coeficiente de correlación cofenética mide el grado de relación entre el índice de jerarquía resultado de la estructura jerárquica y la medida de diferencia. Observamos en el Cuadro 1 que los valores más altos en este índice se manifiestan para los métodos del centroide y de la media. El índice de Mojena informa sobre el número de grupos que refleja la estructura jerárquica resultado de la clasificación. Recordemos que este índice se calcula en base a la búsqueda de "saltos grandes" en los niveles de fusión del dendrograma. Los valores del índice de Mojena sugieren para todos los métodos, excepto para el método del centroide, la existencia de 5 grupos en el conjunto de datos. En el índice de Calinski o índice $C$ se aprecia una mayor divergencia entre los resultados para los diferentes métodos de clasificación. Recordemos que este índice se basa en la comparación entre la variabilidad dentro de los grupos y la variabilidad entre los grupos. El índice calcula el número de grupos en que debe dividirse el conjunto a clasificar de manera que los grupos resulten ser lo más homogéneos dentro de si y lo más heterogéneos entre ellos. El Cuadro 1 muestra que el único método que sigue indicando la existencia de 5 grupos es el método de la media.

| Índice | Ward | Centroide | Mínimo | Máximo | Media |
|---|---|---|---|---|---|
| Correlación cofenética | 0.59 | 0.74 | 0.55 | 0.52 | 0.74 |
| Mojena | 5 | 6 | 5 | 5 | 5 |
| Calinski | 2 | 6 | 3 | 6 | 5 |

Tabla 1: Índices de correlación cofenética, de Mojena, y de Calinski para las clasificaciones obtenidas al aplicar los diferentes métodos de agrupación.

A la vista de los valores del Cuadro 1 y con el objetivo de realizar una comparación de resultados con la clasificación del trabajo de Vives y Villarroya (1996), decidimos analizar únicamente la clasificación obtenida con el método de la media.

En la Figura 3 se ha representado el dendrograma obtenido al aplicar el método de la media usando la distancia de Aitchison (4). En esta misma figura se muestra un nivel de corte del árbol que da lugar a 5 grupos. Es importante resaltar que, si bien en una primera opción se ha analizado la clasificación resultante de considerar los 5 grupos que sugieren los índices de Mojena y de Calinski, se han analizado también los grupos resultantes al considerar un menor o un mayor número de grupos. Sin embargo, a la vista de los resultados obtenidos, se ha decidido que la clasificación en 5 grupos es la agrupación más razonable puesto que esta clasificación es la que manifiesta un patrón diferenciador entre grupos más acusado.

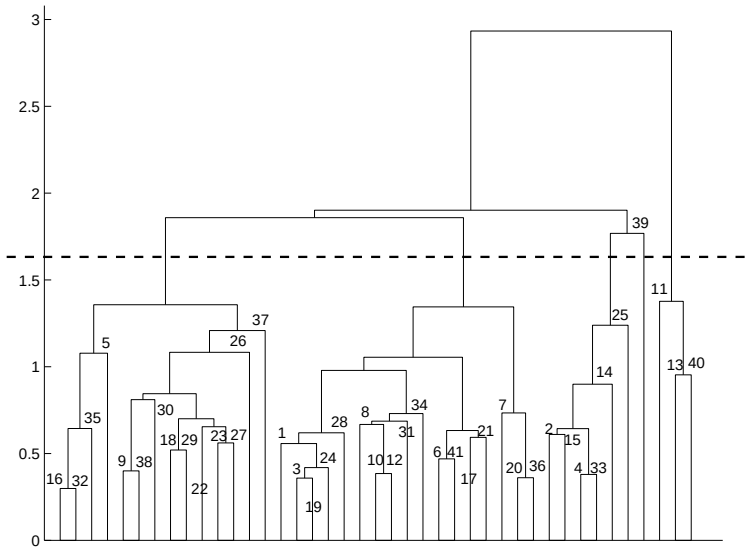



Figura 3: Dendrograma utilizando la distancia de Aitchison y el método de agrupación *de la media*. La línea discontinua indica el nivel de corte del árbol para obtener 5 grupos.

En el Cuadro 2 se muestran los cinco grupos de comarcas resultantes de aplicar el método de media usando la distancia Aitchison (4). En el Cuadro 2, los grupos están

ordenados, en sentido decreciente, según el valor en la parte *Agricultura y pesca* de la media geométrica composicional del grupo o centro del grupo.

| | |
|---|---|
| **Grupo 1** o *Agrícola* | **Grupo 3** o *Turístico* |
| Alta Ribagorça (5) | Val d'Aran (39) |
| Baix Ebre (9) | **Grupo 4** o *Industrial-Medio* |
| Conca de Barberà (16) | Alt Camp (1) |
| Garrigues (18) | Alt Penedès (3) |
| Montsià (22) | Anoia (6) |
| Noguera (23) | Bages (7) |
| Pallars Sobirà (26) | Baix Camp (8) |
| Pla d'Urgell (27) | Baix Empordà (10) |
| Priorat (29) | Baix Penedès (12) |
| Ribera d'Ebre (30) | Garraf (17) |
| Segarra (32) | Garrotxa (19) |
| Solsonès (35) | Gironès (20) |
| Terra Alta (37) | Maresme (21) |
| Urgell (38) | Osona (24) |
| **Grupo 2** o *Militar* | Pla de l'Estany (28) |
| Alt Empordà (2) | Ripollès (31) |
| Alt Urgell (4) | Selva (34) |
| Berguedà (14) | Tarragonès (36) |
| Cerdanya (15) | Vallés Oriental (41) |
| Pallars Jussà (25) | **Grupo 5** o *Industria-Servicios* |
| Segrià (33) | Baix Llobregat (11) |
| | Barcelonès (13) |
| | Vallès Occidental (40) |

Tabla 2: Las 41 comarcas de Catalunya clasificadas en 5 grupos resultado de aplicar el método de la media con la distancia de Aitchison. Entre paréntesis aparece el número identificador de la comarca.

Con el objetivo de ilustrar las diferencias entre los 5 grupos resultado de la clasificiación, en la Figura 4 representamos las medias geométricas composicionales de cada uno de los 5 grupos mediante un diagrama de barras. En los diagramas de barras de esta figura, la altura de cada barra representa el valor de la correspondiente parte de la media geométrica composicional. En el diagrama ternario de la Figura 5(a) representamos la subcomposición formada por las partes $\mathbf{X}_6$, $\mathbf{X}_7$, y $\mathbf{X}_8$. En la Figura 5(b) se muestra la misma subcomposición después de centrar los datos siguiendo la metodología propuesta en Martín-Fernández et al. (1999). En la Figura 6, siguiendo el trabajo de Aitchison y Greenacre (2002), se representa el diagrama *biplot* del conjunto de datos $\mathbf{X}$.
De los 5 grupos obtenidos se destacan las siguientes características:

- Grupo 1 o *Agrícola*. Las 14 comarcas que pertenecen a este grupo tienen como característica principal la de tomar valores altos en la parte *Agricultura y pesca*. En la Figura 4 se observa que es el grupo con mayor valor en la parte

*Agricultura y pesca* de la media geométrica composicional. En las subcomposiciones representadas en la Figura 5 se observa que el Grupo 1 aparece situado como el grupo más cercano al vértice de la sexta parte $\mathbf{X}_6$ que corresponde a la actividad *Agricultura y pesca*. En el diagrama *biplot* de la Figura 6 el Grupo 1 se encuentra situado alrededor del eje de la sexta parte, correspondiente a *Agricultura y pesca*, y en posiciones alejadas del centro del diagrama, poniendo de manifiesto el hecho que las comarcas pertenecientes a este grupo poseen una elevada proporción de población activa dedicada a la *Agricultura y pesca*.

- Grupo 2 o *Militar*. Este grupo está formado por 6 comarcas que tienen un elevado porcentaje de personal perteneciente a las actividades correspondientes a las *Fuerzas armadas*. Son comarcas fronterizas o con un número elevado de instalaciones militares que provoca que sean observaciones que toman valores relativamente altos en la parte $\mathbf{X}_8$. Observando la Figura 4 puede apreciarse que, exceptuando el Grupo 3, el Grupo 2 es el grupo con mayor valor en la parte $\mathbf{X}_8$ (*Fuerzas armadas*) de la media geométrica composicional. Por lo que se refiere a las partes correspondientes a los otros grupos profesionales, el Grupo 2 no destaca por tomar valores alejados de la media del total de comarcas de Catalunya. En los diagramas que muestran las Figuras 5(a) y 5(b) las comarcas del Grupo 2 aparecen entre las comarcas más cercanas al vértice de la parte $\mathbf{X}_8$. En el diagrama *biplot* de la Figura 6 se aprecia que las comarcas del Grupo 2 aparecen situadas muy cercanas al eje de la variable correspodiente a *Fuerzas Armadas* y alejadas del origen de coordenadas. Este gráfico pone de manifiesto que el Grupo 2 está formado por comarcas que toman valores relativamente altos en la parte $\mathbf{X}_8$ y toman valores medios en las otras partes.

- Grupo 3 o *Turístico*. Este grupo está formado únicamente por la comarca de la Val d'Aran. Esta comarca se distingue por tomar valores altos conjuntamente en las partes $\mathbf{X}_2$ (*Personal Directivo*), $\mathbf{X}_5$ (*Hostelería y otros*) y $\mathbf{X}_8$ (*Fuerzas Armadas*). Estas características ponen de manifiesto la existencia de una fuerte industria turística de ámbito local. En los diagramas de barras de la Figura 4 se aprecia que la barra de las partes $\mathbf{X}_2$ y $\mathbf{X}_5$ de la media geométrica composicional es más alta en este grupo que en el resto de grupos. Nótese que la barra de la parte $\mathbf{X}_8$ también es alta, acorde con la existencia de numerosas instalaciones militares en la comarca. En los diagramas ternarios de la Figura 5 la comarca de la Val d'Aran se ha representado mediante un triángulo invertido. En la Figuras 5(a) y 5(b), en las que no interviene la parte $\mathbf{X}_5$, la comarca de la Val d'Aran aparece entre las comarcas más cercanas al vértice de la parte $\mathbf{X}_8$. En el diagrama *biplot* de la Figura 6 se aprecia que la comarca de la Val d'Aran aparece situada en el semiplano determinado por las variables correspondientes a *Hostelería y otros* y a *Fuerzas Armadas*, y muy alejada del origen de coordenadas. Nótese que el símbolo de esta comarca aparece dentro de un círculo indicando que la comarca de la Val d'Aran puede ser catalogada

como una observación atípica.

- Grupo 4 o *Industrial-Medio*. Las 17 comarcas pertenecientes a este grupo se caracterizan por tomar valores medios en las partes $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_3$, $\mathbf{X}_4$, $\mathbf{X}_5$, y $\mathbf{X}_8$. Son comarcas cuya distribución de la población activa en estas variables se asemeja a la distribución de la población activa en toda Catalunya. Sin embargo, en contraposición a las comarcas pertenecientes al Grupo 1, las comarcas de este Grupo 4 toman valores más altos en la parte $\mathbf{X}_7$ (*Industria*) y más bajos en la parte agrícola $\mathbf{X}_6$. Estas características de la tendencia central del Grupo 4 pueden apreciarse de manera gráfica en el diagrama de barras de la Figura 4. En los diagramas ternarios de la Figura 5 se observa que las comarcas del Grupo 4 aparecen en la zona central de la nube de puntos pero siempre situadas más cercanas al vértice de la parte industrial $\mathbf{X}_7$ que del vértice de la parte agrícola $\mathbf{X}_6$. En el diagrama *biplot* de la Figura 6 puede apreciarse en su zona central las comarcas de esta Grupo 4. Nótese que la mayoría de las comarcas del grupo aparecen en la zona del semieje negativo de la variable correspondiente a la *Agricultura y Pesca* con lo que se manifiesta que las comarcas pertenecientes al Grupo 4 toman valores bajos en la parte agrícola.

- Grupo 5 o *Industria-Servicios*. Este grupo está formado por tres comarcas cuya característica principal es la de tomar valores relativamente altos conjuntamente en las cinco primeras partes y en la parte $\mathbf{X}_7$ de actividades industriales. En este grupo destacan el valor alto en la parte $\mathbf{X}_3$ de *Servicios Administrativos*, aportado por la comarca del Barcelonès y el valor alto en la parte industrial $\mathbf{X}_7$, aportado por las comarcas del Baix LLobregat y la comarca del Vallès Occidental. En el diagrama de barras de la Figura 4 puede apreciarse que las comarcas del Grupo 5 toman valores altos en todas las partes excepto en la parte agrícola y en la parte militar $\mathbf{X}_8$. En los diagramas ternarios de la Figura 5 en los que interviene la parte $\mathbf{X}_6$ las tres observaciones perteneciente a este Grupo 5 aparecen como las más alejadas del vértice de la parte $\mathbf{X}_6$. En los diagramas, donde se representan subcomposiciones en las que interviene la parte $\mathbf{X}_7$, las tres observaciones del Grupo 5 aparecen entre las más cercanas a su vértice. Como puede apreciarse en el diagrama *biplot* de la Figura 6 la parte agrícola de estas comarcas es muy pequeña. Nótese que dos de estas tres comarcas, el Baix LLobregat y el Barcelonès, pueden ser catalogadas como observaciones atípicas.

Las características que acabamos de exponer de cada uno de los 5 grupos resultantes de la clasificación ponen de manifiesto que las observaciones que pertenecen a grupos

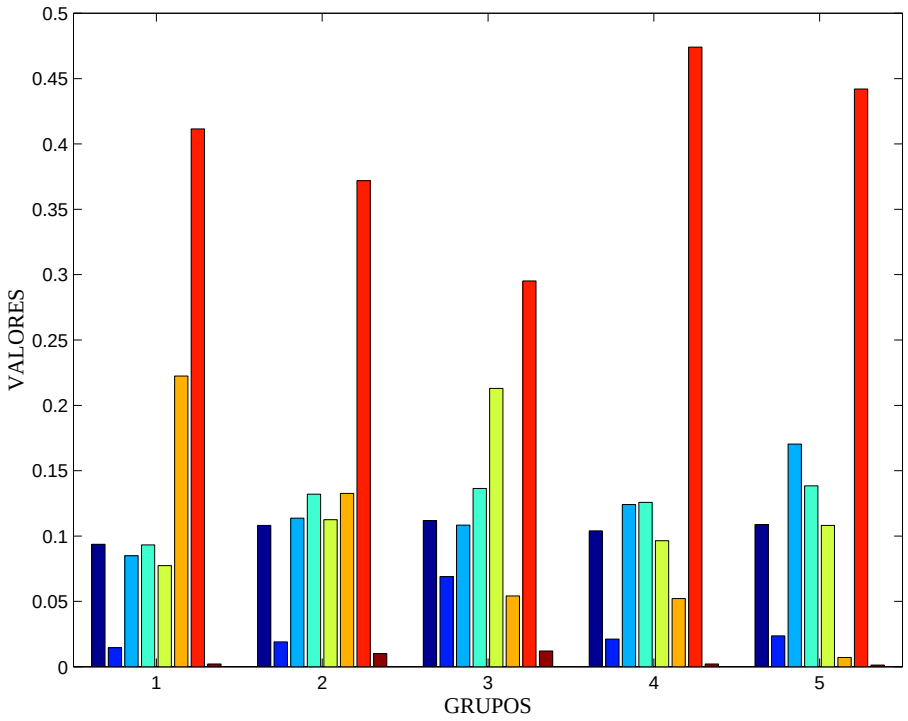
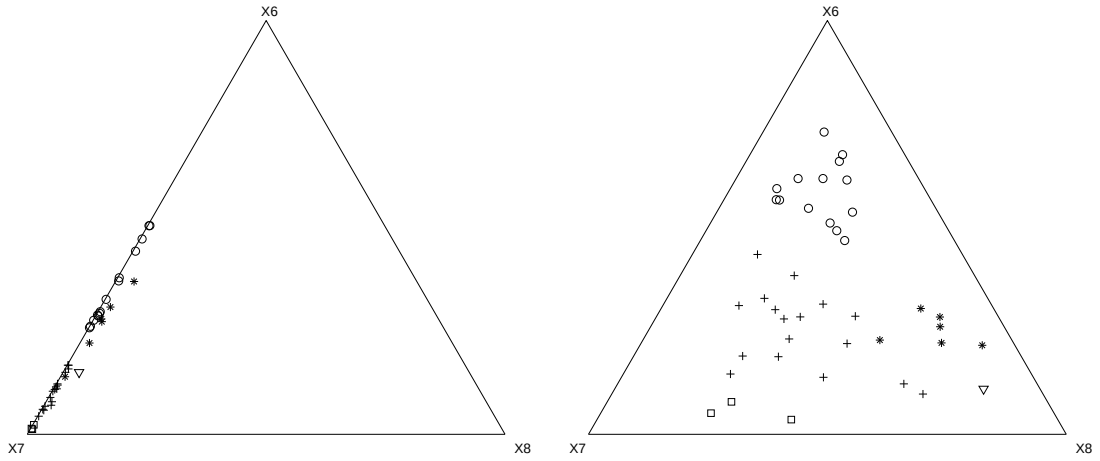



Figura 4: Diagramas de barras de la media geométrica composicional de cada uno de los 5 grupos de comarcas resultado de aplicar el método de la media y la distancia de Aitchison. Cada barra representa el valor de la correspondiente parte de la media geométrica composicional.




(a) (b)

Figura 5: Subcomposiciones del conjunto en las partes $X_6$, $X_7$, y $X_8$: (a) datos sin centrar; (b) con los datos centrados. Se muestran los 5 grupos resultado de aplicar el método de la media y la distancia de Aitchison. (Grupo 1:'$\circ$'; Grupo 2: '$*$'; Grupo 3: '$\nabla$'; Grupo 4:'$+$'; Grupo 5:'$\square$').

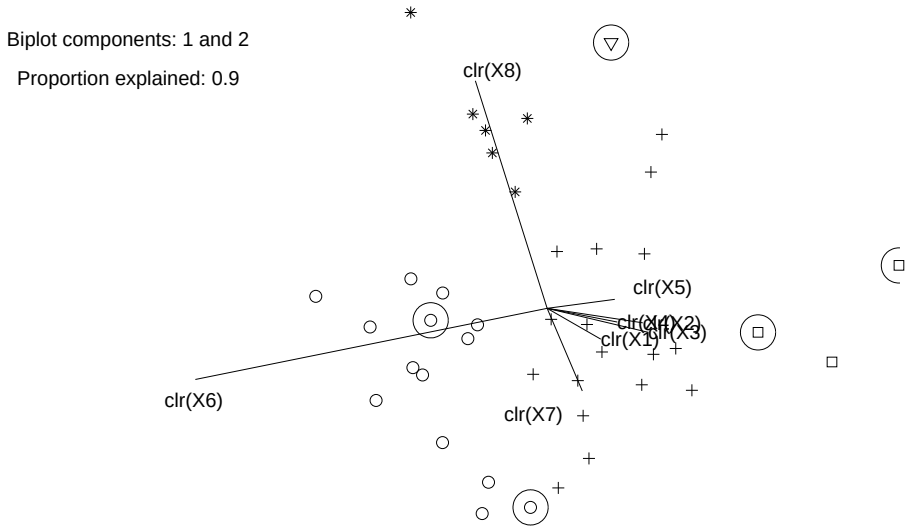



Figura 6: Diagrama biplot. Se muestran los 5 grupos determinados por el método de la media. (Grupo 1:'$\circ$'; Grupo 2: '$*$'; Grupo 3: '$\nabla$'; Grupo 4:'$+$'; Grupo 5:'$\square$'). Las observaciones dentro de un círculo pueden ser catalogables como atípicas.

diferentes muestran un patrón claramente diferenciado en el valor que toman en las diferentes variables. En consecuencia, consideramos que la agrupación obtenida es una clasificación razonable. Somos conscientes que al habernos limitado a estudiar los resultados que proporciona el método de la media no hemos completado el estudio del conjunto *Población ocupada por grupos profesionales*. Pueden obtenerse otras clasificaciones razonables mediante la aplicación de otros métodos de clasificación automática no paramétrica. Sin embargo, con el ánimo de no extender en demasía el estudio del caso práctico que nos ocupa y recordando que uno de los centros de interés del estudio es la comparación de resultados con la clasificación presentada en Vives y Villarroya (1996), decidimos no desarrollar otras clasificaciones resultantes de aplicar métodos diferentes al de la media.

## 3.2 Comparación de resultados

En una primera lectura puede observarse que existe una notable similitud entre los resultados de las clasificaciones del trabajo de Vives y Villarroya (1996) y los resultados obtenidos utilizando la metodología propuesta para datos composicionales. De esta similitud destacan los dos aspectos siguientes:

- Gran coincidencia en las comarcas que son calificadas como agrícolas y como industriales en las dos clasificaciones.
- En las dos agrupaciones la comarca de la Val d'Aran constituye, por si sola, un grupo cuya característica diferenciadora es una alta proporción en la parte

turística.

Sin embargo, una comparación más pausada nos lleva a detectar la existencia de diferencias entre las dos clasificaciones. De las diferencias detectadas destacamos los dos aspectos siguientes:

- En nuestra clasificación se consideran 5 grupos de comarcas; un grupo más que en la clasificación de Vives y Villarroya (1996) que contemplaba 4 bloques diferentes. Las comarcas que pertenecen a este nuevo grupo –véase el Grupo 2 del Cuadro 2– se distinguen por su alta proporción relativa de población activa dedicada a actividades englobadas bajo el nombre de *Fuerzas armadas*. De las 6 comarcas pertenecientes a este grupo la comarca del Berguedà era asignada en el trabajo de Vives y Villarroya (1996) al Bloque Industrial. Las otras 5 comarcas pertenecientes al Grupo 2 de nuestra clasificación formaban parte del grupo denominado *Bloque Agrícola* en la clasificación de Vives y Villarroya (1996).

- En nuestra agrupación la comarca del Barcelonès no constituye un grupo por si sola. El Grupo 5 o *Industria-Servicios* de nuestra agrupación está formado, además de la comarca del Barcelonès, por las comarcas del Baix Llobregat y del Vallès Occidental. Este grupo pone de manifiesto la existencia en Cataluña de un área geográfica, que engloba la ciudad de Barcelona y sus alrededores, donde la proporción de servicios administrativos y de tejido industrial es muy elevada. En la clasificación de Vives y Villarroya (1996), la comarca del Barcelonès constituía por si sola el grupo denominado *Administrativo* y las otras dos comarcas formaban parte del grupo denominado *Industrial*.

Sin estar en nuestro ánimo el calificar como mejor o peor una de los dos clasificaciones, creemos importante destacar que la distancia de Aitchison es mucho más sensible a las variaciones relativas en las proporciones que la disimilitud utilizada en el trabajo de Vives y Villarroya (1996). Esta característica se pone especialmente de manifiesto en el hecho que nuestra clasificación contempla la existencia de un grupo de comarcas cuya distinción principal se basa en la alta proporción relativa de la parte *Fuerzas armadas* cuyo rango de variación en toda Catalunya abarca desde un mínimo del 0.1% hasta un máximo del 2%. Consideramos que esta mayor sensibilidad de la distancia de Aitchison ante valores cercanos a cero es una virtud que convierte a esta distancia en una medida muy útil en el estudio de conjuntos de datos que contengan partes con valores casi nulos pero conceptualmente significativos. En el estudio de conjuntos de datos sin partes con valores cercanos a cero las divergencias de resultados serían menores.

# 4 Agradecimientos

Este trabajo de investigación ha sido parcialmente subvencionado por la DGESIC (Ref.: BFM2000-0540) y por el Dept. de Informàtica i Matemàtica Aplicada de la

Universitat de Girona (UdG).

# Referencias

Aitchison, J. y Greenacre, M. (2002): Biplots of Compositional Data. *Applied Statistics*. 51, Part 4, 375–392.

Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J.A. and Pawlowsky-Glahn, V. (2000): Logratio analysis and compositional distance. *Mathematical Geology*. 32(3), 271–275.

Barceló-Vidal, C., Martín-Fernández,J.A. and Pawlowsky-Glahn, V. (2001): Mathematical Foundations of Compositional Data Analysis. In *Proceedings of the Annual Conference of the International Association for Mathematical Geology*. Cancún (México), CD-ROM, 20p.

Barceló-Vidal, C., Martín-Fernández, J. A. and Pawlowsky-Glahn, V. (2003): Fundamentos matemáticos de los datos composicionales. En *Libro de actas del XXVII Congreso Nacional de Estadística e Investigación Operativa*. Lleida (E). (*en esta publicación*).

Martín-Fernández, J.A., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (1998a): Measures of Difference for Compositional Data and Hierarchical Clustering Methods. In *Proceedings of the Fourth Annual Conference of the International Association for Mathematical Geology*. Ed. A. Buccianti, G. Nard, and R. Potenza. Nápoles (I), Part 2, 526–531.

Martín-Fernández, J.A., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (1998b): A critical approach to non-parametric classification of compositional data. In *Proceedings of the 6th Conference of the International Federation of Classification Societies. Università La Sapienza, Roma*. Ed. A. Rizzi, M. Vichi, and H.H. Bock. Springer-Verlag, Berlín (D), 49–56.

Martín-Fernández, J.A., Barceló-Vidal, C., y Pawlowsky-Glahn, V. (1998c): Medida de diferencia de Kullback-Leibler entre datos composicionales. En *Libro de Actas del XXIV Congreso Nacional de Estadística e Investigación Operativa*. Ed. Sociedad Espanõla de Estadística e Investigación Operativa, Almería (E), 291–292.

Martín-Fernández, J. A., Bren, M., Barceló-Vidal, C., y Pawlowsky-Glahn, V. (1999): A measure of difference for compositional data based on measures of divergence. En *Proceedings of IAMG'99. The Fifth Annual Conference of the International Association for Mathematical Geology*. Ed. Lippard, S.J. and Næss, A. and Sinding-Larsen, R., Trondheim (Norway), 1, 211–215.

Vives, S. y Villarroya, A. (1996): La combinació de técniques de geometria diferencial amb anàlisi multivariant clàssica: Una aplicació a la caracterització de les comarques catalanes. *Qüestiió*, 20 (3), 449–482.