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

ANÁLISIS FACTORIAL MÚLTIPLE PARA TABLAS DE CONTINGENCIA: ESTUDIO DE LA MORTALIDAD EN LAS COMUNIDADES AUTÓNOMAS DE ESPAÑA

Mónica Bécue-Bertaut[1], Jérôme Pagès[2], René Álvarez Laverde[1], María Luisa Hernández Maldonado[1]

[1]Departamento de Matemáticas e Investigación Operativa
Universidad Politécnica de Cataluña, Barcelona. España.
E-mail: monica.becue@upc.es; hrene@eio.upc.es; m.luisa.hernandez@upc.es
[2]ENSA/INFSA, Rennes, Francia
E-mail: jerome.pages@agrorennes.educagri.fr

RESUMEN

A partir del fichero de defunciones en España, se construyen las tablas de contingencia cruzando comunidad de residencia y causa de mortalidad por sexo y grandes grupos de edad (tabla múltiple de contingencia). El objetivo es caracterizar las diferentes comunidades en función de la información global y de cada una de las subtablas. La metodología de análisis utilizada es el AFM para tablas de contingencia (AFMTC), extensión del análisis de correspondencias a tablas de contingencia múltiple. Ofrece una síntesis de la tabla múltiple (Descripción global) y una comparación sistemática de las subtablas (descripción parcial) mediante diversos índices.

Palabras clave: Tabla de Contingencia Múltiple, Análisis de Correspondencias, Análisis Factorial Múltiple, Análisis Factorial Múltiple para Tablas de Contingencia, Análisis Intra-tablas.

1. Introducción

El análisis factorial múltiple para tablas de contingencia (Bécue & Pagès, 1999, 2000) facilita la comparación de varias tablas de contingencia con una sola dimensión en común.
El objetivo de este trabajo consiste en mostrar las propiedades de este método mediante su aplicación al estudio la variabilidad de las causas de mortalidad (prematura y no prematura, diferenciando por sexo) entre las regiones autónomas, en referencia la variabilidad de la pirámide de edad.
La diversidad de las regiones de España (industrialización, nivel económico, condiciones climáticas, costumbres dietéticas, etc.) entraña una desigualdad frente a la mortalidad (ver, por ejemplo, Benach *et al*. 2001a, 2001b) que interesa conocer tanto para establecer la planificación de las compañías de seguro como para mejorar la política sanitaria.
El objetivo consiste en contestar a preguntas como:

- ¿existen asociaciones entre regiones y causas, que no provengan solamente de las diferencias entre pirámides de edad?

- ¿existe una relación entre la estructura regional de la mortalidad prematura y la estructura regional de la mortalidad no prematura?

No se aborda la problemática de la sobremortalidad sino la distribución de las causas entre las regiones de manera a establecer, para cada región, las causas predominantes y, para cada causa, las regiones que presentan una alta proporción de dicha causa entre en su mortalidad.

2. Datos

Se utilizan aquí los datos de mortalidad del año 1995 compilados por el Instituto Nacional de Estadística correspondientes a los adultos de 35 años o más. La mortalidad infantil, muy particular, y la de los jóvenes adultos, causada principalmente por los accidentes y el SIDA, merecen estudios propios. No se tienen en cuenta las regiones de Ceuta y Melilla.

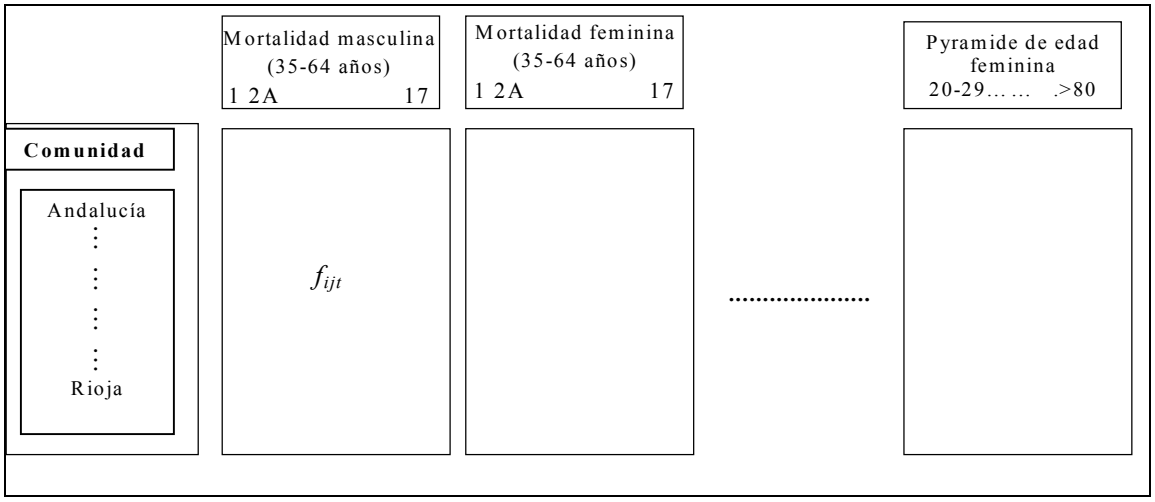

Figura 1. Tabla de contingencia múltiple

17 regiones-filas; 6 grupos de columnas, mortalidad masculina entre 35 y 64; mortalidad femenina entre 35 y 64 años; mortalidad masculina con 65 años y más, mortalidad femenina con 65 años y más; pirámides de edad masculina y femenina.

Datos de mortalidad: Las defunciones se reparten por causa, sexo y región de residencia. Las causas de mortalidad están clasificadas según el ICD (*International Classification of Diseases*) y reagrupadas a un nivel más o menos detallado de esta clasificación, según la importancia numérica de las causas (Tabla 1).
No se tienen en cuenta las causas 14 (*anomalías congénitas*) y 15 (*algunas condiciones originadas en el período perinatal)* casi exclusivas de los niños ni la causa 11 (*complicaciones del embarazo, del recién nacido y del posparto*) muy poco frecuente (menos de 3 casos por región en 1995).
A partir de los datos de mortalidad, se construyen 4 tablas de contingencia, dos para cada sexo (de 35 a 64 años, más de 64 años), con las regiones en fila y las causas de mortalidad en columna.

Pirámides de edad: la distribución de los habitantes por edad (7 categorías, de diez en diez años, desde 20 hasta 79 años, y 80 o más), sexo y región. Los efectivos son las proyecciones realizadas por el INE a partir del censo de 1991.
Los datos de edad llevan a construir 2 tablas de contingencia, una para los hombres, otra para las mujeres, con las regiones en fila y los intervalos de edad en columna.
Información suplementaria: se dispone también de indicadores económicos y sociales sobre las regiones.

Nº	Descripción	Código ICD
1	Infecciones y enfermedades parasitarias	(001-139)
2	Neoplasmas	(140-239)
	Cáncer de estomago	151
	Cáncer de colon	153
	Cáncer de pulmón	161
	Cáncer de pecho	174
	Cáncer de próstata	185
	Otros cánceres	
3	Enfermedades endocrinas, nutricionales, metabólicas e inmunológicas	(240-279)
	Diabetes	250
	SIDA	279.5
	Otras enfermedades endocrinas e inmunológicas	
4	Enfermedades de la sangre y órganos de formación de la sangre	(280-289)
5	Desordenes Mentales	(290-319)
	Demencia	290
	Otros desordenes mentales	
6	Enfermedades del sistema nervioso y de los órganos de los sentidos	(320-389)
	Mal de Alzeimer	331
	Otras enfermedades del sistema nervioso	
7	Enfermedades del sistema circulatorio	(390-459)
	Enfermedad isquémica del corazón	410-414
	Otras enfermedades del corazón	
	Enfermedades cerebro vasculares	430-438
	Arterosclerosis	440
	Otras enfermedades del sistema circulatorio	
8	Enfermedades del Sistema Respiratorio	(460-519)
	Infecciones respiratorias agudas, neumonía e influenza	460
	Enfermedad pulmonar obstructiva crónica	490-496
	Otras enfermedades del sistema respiratorio	
9	Enfermedades del sistema Digestivo	(520-579)
	Cirrosis	571
	Otras enfermedades del sistema digestivo	
10	Enfermedades del sistema genitourinario	(580-629)
11	Complicaciones del embarazo, del recién nacido y del posparto	(630-676).
12	Enfermedades de la piel y del tejido subcutáneo	(680-709).
13	Enfermedades del sistema músculo esquelético y del tejido conectivo	(710-739)
14	Anomalías congénitas	(740-759)
15	Ciertas condiciones originadas en el periodo perinatal	(760-779)
16	Síntomas y signos mal definios.	(780-799)
17	Heridas y envenenamientos	
	Lesiones por accidentes de trafico	E810
	Suicidio	E950
	Otras heridas	

Tabla 1: Causas de mortalidad

3. Análisis de la tabla múltiple

3.1 Análisis de correspondencias separados de las seis tablas

El análisis de correspondencias (AC) (Benzécri, 1973; Escofier & Pagès 1988-1998, Lebart *et al*., 1984, 1988), es un instrumento privilegiado para la descripción de las tablas de contingencia. Se puede analizar cada una de las 6 tablas por separado y comparar las estructuras inducidas sobre las regiones por cada grupo de columnas, pero la comparación es un trabajo arduo y la síntesis de los resultados es compleja.

3.2 Análisis de correspondencias de la tabla yuxtapuesta

Se puede también realizar el AC de la tabla múltiple. Este análisis permitiría contestar a preguntas del tipo:

- ¿una misma causa pero prematura y no prematura se asocia o no a las mismas regiones?
- ¿qué regiones tienen un perfil de mortalidad similar, tanto para la mortalidad prematura como para la no prematura?
- ¿qué regiones tienen una mortalidad similar y una pirámide de edad similar?

Pero se debe notar que la estructura inducida sobre las regiones depende:

- de las diferencias entre perfiles de los márgenes-fila (suma de los valores de la fila) de las diferentes tablas;
- de la importancia relativa de las tablas en el análisis, medida a través las contribuciones de las columnas.

Además, esta metodología no proporciona información relativa a dos puntos importantes:

- la comparación de las diferentes estructuras sobre las regiones inducidas por cada tabla, es decir, ¿qué regiones son a la vez similares desde el punto de vista de un grupo de mortalidad o pirámide de edad y alejadas por otro grupo de mortalidad o pirámide de edad?.
- La definición de una estructure sobre los grupos de columna: ¿qué grupos inducen una estructura parecida? ¿Cuáles inducen una estructura diferente?

Para tratar estas preguntas, la metodología de referencia es el AC separada de cada tabla pero, como se comentó, la comparación de los resultados puede ser muy laboriosa y a veces inextricable.

Además, el AC de la tabla yuxtapuesta no proporciona información relativa a dos puntos importantes:

- La comparación de las estructuras sobre las regiones inducidas por cada tabla, es decir, qué regiones son a la vez próximas desde el punto de vista de un grupo de mortalidad o pirámide de edad y alejadas por otro grupo de mortalidad o pirámide de edad.
- La definición de una estructure sobre los grupos de columna: ¿qué grupos inducen una estructura parecida? ¿Cuáles inducen una estructura diferente?

Para contestar a estas preguntas, proponemos utilizar el AFMTC, análisis factorial múltiple para tablas de contingencia, propuesta por Bécue y Pagès (1999, 2000), como una extensión del AFM (Escofier y Pagès, 1998; 1998). Se presentan en la sección 3.3 los principios básicos del AFMTC.

3.3 AFMTC de la tabla yuxtapuesta

Notación. f_{ijt} : frecuencia relativa asociada a la fila i y columna j de la tabla t; un índice sustituido por un punto indica la suma sobre este índice.

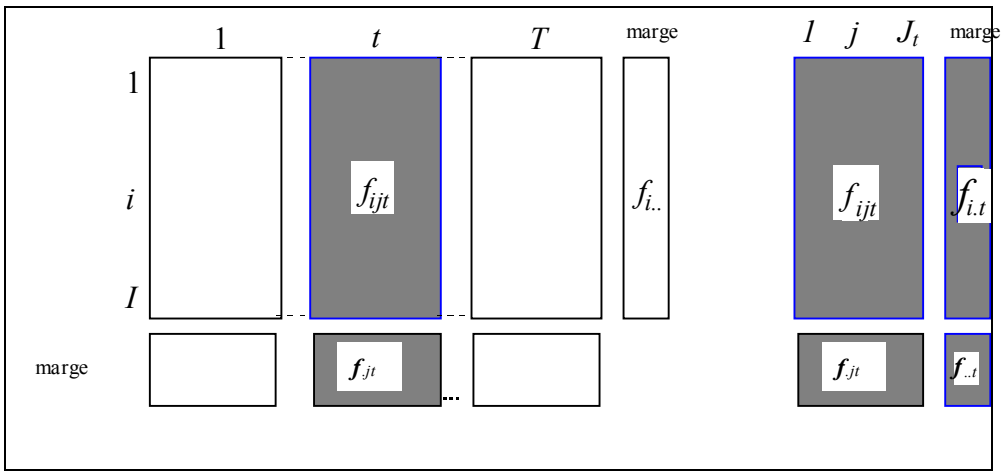

Figure 2. Tabla de contingencia múltiple y márgenes: notación

Análisis separados: se realiza primero un AC de cada tabla t pero imponiendo los márgenes-fila $\{f_{i..},\ i=1,\ldots,I\}$ y los márgenes-columnas, $\{f_{.jt},\ j=1,\ldots J_t\}$ (de hecho, son análisis seudo-separados; la diferencia entre los AC separados y los AC seudo-separados, es decir, con los márgenes impuestos indicados, difieren poco si los márgenes-fila internos $f_{i.t}$ difieren poco).

El AC con márgenes impuestos de la tabla t es equivalente al ACP de la tabla de término general:

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

dando a la fila i, el peso $f_{i..}$ y a la columna j,t, el peso $f_{.jt}$. Las filas tiene así un mismo peso en todos los análisis, igual al peso medio calculado sobre el conjunto de las tablas.

Análisis global de la tabla múltiple El AFMTC es equivalente a realizar un ACP no reducido de la tabla yuxtapuesta, de término general dado por (1), dando a la fila i el peso $f_{i..}$ y a la columna (j,t) el peso $f_{.jt}/\lambda_1{}^t$. Esta ponderación permite equilibrar la importancia las tablas en el análisis global.

Este análisis ofrece resultados:

- similares a los del AC aplicado a las tablas yuxtapuestas (principalmente, una representación global de las filas-regiones y de las columnas-causas y intervalos de edad)

- específicos de las tablas múltiples, principalmente, la representación superpuesta de las estructuras de las regiones inducidas 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.

4. Resultados

4.1 Análisis separados

Los primeros valores propios de los análisis separados valen: 0.0391 (mortalidad en hombres de 35 a 64 años), 0.0059 (mortalidad en mujeres de 35 a 64 años), 0.0714 (mortalidad en hombres con 65 años y más), 0.0086 (mortalidad en mujeres con 65 años y más), 0.007 (pirámide de edad masculina) y 0.0007 (pirámide de edad femenina). La gran diferencia entre los valores propios de los análisis separados justifica la necesidad de equilibrar la influencia de los grupos.

Las estructuras sobre las regiones inducidas por las dos pirámides de edad, por separado, son bidimensionales, con un primer eje muy dominante. Las estructuras sobre las regiones inducidas por las cuatro tablas de mortalidad, por separado, son más complejas (4 ó 5 direcciones de dispersión según las tablas).

4.2 Análisis global

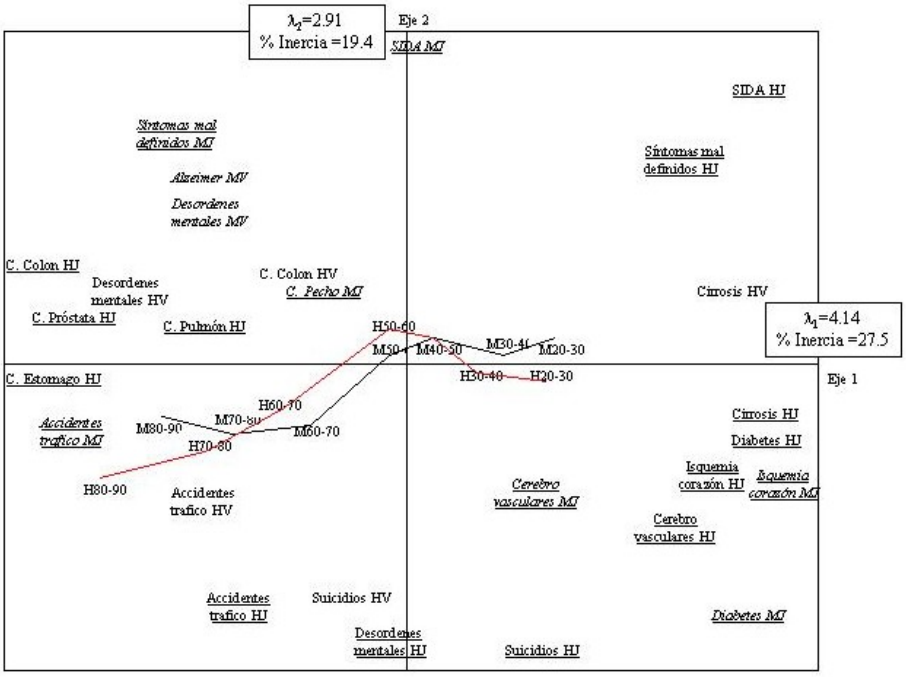

Figura 3. Primer plano factorial: trayectoria de las pirámides de edad y de algunas de las causas de mortalidad

El primer valor propio del análisis global vale 4.14 (27.5% de la inercia total). La secuencia de los valores propios sugiere la presencia de cinco ejes interpretables. Estos cinco ejes acumulan 72.7% de la inercia.
Los seis grupos contribuyen la construcción del primer eje con un 15.2%, 19%, 13%, 13.4%, 19.5% y 20% de la inercia, respectivamente. Esta dirección de dispersión es común a los seis grupos. La contribución al segundo eje de los grupos de mortalidad es, globalmente, igual a 81.3%. Esta dirección de dispersión es común a los seis grupos, pero es importante solamente en los grupos de mortalidad.

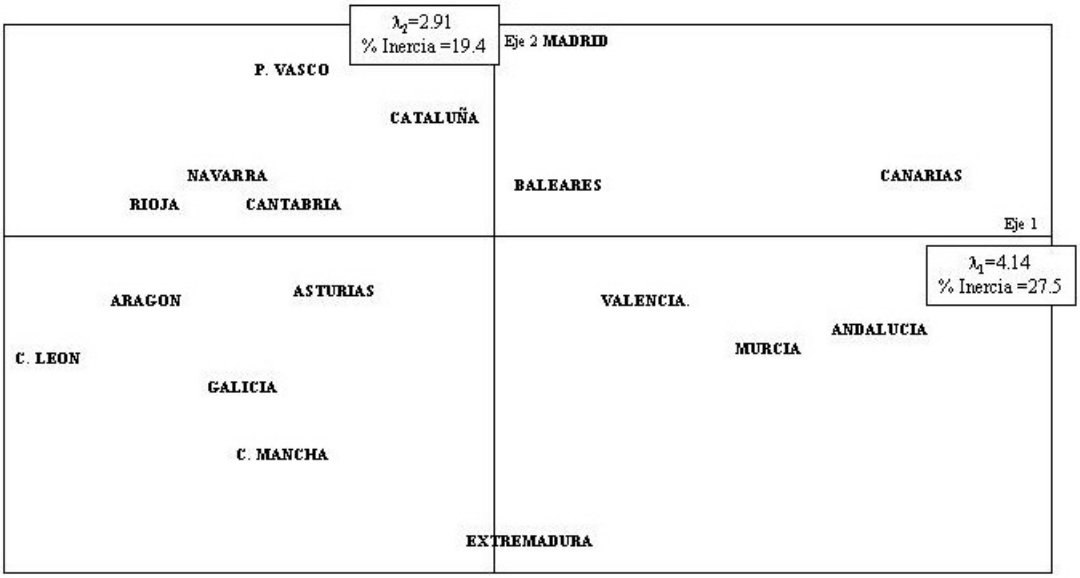

Figura 4. Primer plano factorial: representación de las comunidades autónomas

Las representaciones de las filas y de las columnas sobre los planos factoriales se pueden superponer y deben interpretarse simultáneamente.

Primer eje

El primer eje ordena los intervalos de edad en su orden natural. Opone regiones en las cuales existe un exceso de jóvenes adultos ("regiones jóvenes": Islas Canarias, Andalucía y Murcia) con regiones que tienen más personas mayores que la media ("regiones mayores": Asturias, Castilla y León).
Dicho eje opone también las causas de mortalidad frecuentes en las "regiones jóvenes" (como mortalidad prematura por enfermedad isquemia del corazón tanto en hombre como en mujeres, mortalidad prematura por SIDA en hombres, mortalidad prematura por diabetes en mujeres, infecciones en mujeres, cáncer de colón en hombres mayores) con las causas de mortalidad frecuentes en las "regiones mayores" (mortalidad prematura en hombres por cáncer de colón, cáncer de estomago y otro tipos de cáncer, mortalidad prematura por accidentes de tráfico en mujeres, mortalidad por síntomas mal definidos tanto en hombres como en mujeres mayores y demencia en mujeres mayores).
El primer eje tiene una correlación positiva con los índices de desempleo (0.69), hacinamiento (0.57) y grado de analfabetismo (0.70) (Figura 5).

Segundo eje

El segundo eje opone el las comunidades con un exceso de adultos de edad intermedia (40-59 años), principalmente, Madrid, País Vasco y Cataluña, con las comunidades con un déficit de personas en estas edades y, para muchas de ellas, un exceso de personas mayores, principalmente Extremadura, Galicia y Castilla la Mancha.

Esta oposición viene acompañada de un incremento de mortalidad por SIDA (hombres y mujeres jóvenes), lesiones y envenenamientos (mujeres jóvenes y mayores), desordenes mentales (mujeres mayores), síntomas mal definidos (hombres y mujeres) en el primer grupo de regiones. En el segundo grupo de regiones, se nota un incremento de las siguientes causas de mortalidad: enfermedad obstructiva pulmonar crónica (mujeres y hombres jóvenes), suicidios (hombres jóvenes y mayores, mujeres jóvenes), enfermedades cerebro-vasculares (mujeres jóvenes y mayores) e infecciones (mujeres jóvenes).

El segundo eje tiene una correlación muy fuerte con el PIB (0.94), fuerte con el porcentaje de diplomados superiores sobre la población de egresados del sistema escolar en los últimos 10 años (0.81) y menos fuerte con el analfabetismo (-0.56).

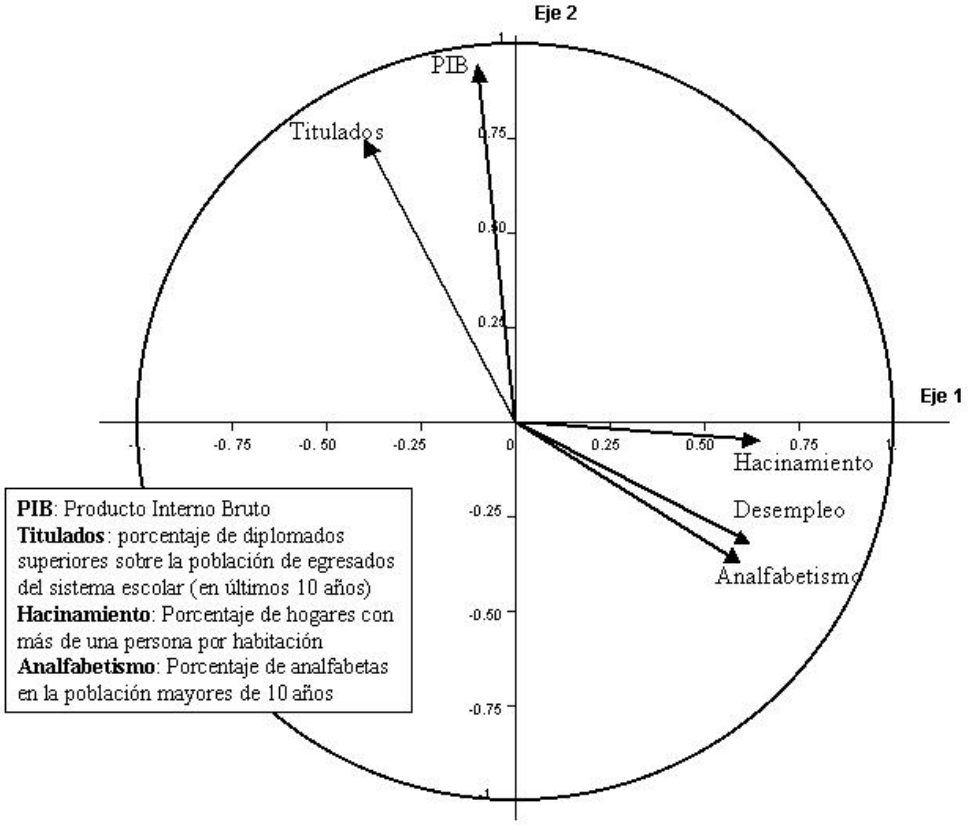

Figura 5. Proyección de las variable ilustrativas sobre el primer plano factorial

Ejes 3, 4 y 5

Los siguientes ejes ponen en evidencia rasgos específicos de algunas regiones, como, por ejemplo, la alta incidencia de las enfermedades del sistema respiratorio en los hombres y las mujeres mayores en Galicia, y la alta incidencia de estas mismas enfermedades pero sólo para los hombres mayores en Asturias y Aragón.

4.3 Superposición de las estructuras parciales

Para comparar las estructuras sobre las regiones inducidas por los 6 grupos de columnas, el AFMTC ofrece una superposición de las filas-regiones, descritas par el conjunto de los grupos (representación global) y por cada una de los grupos (representación parcial).
La figura 5 ofrece un extracto de la representación superpuesta en la cual se puede notar, por ejemplo, que:

- Madrid y Andalucía tienen una estructura de edad similar (los puntos parciales Mad5, Mad6, And5 y And6, que representan, respectivamente Madrid y Andalucía según las pirámides de edad de los hombres (grupo 5) y de las mujeres (grupo 6) están muy próximos) pero la estructura de la mortalidad es muy diferente (los puntos parciales que representan la mortalidad femenina prematura en Madrid y Andalucía, Mad3 y And3 tienen posiciones muy alejadas).
- La mortalidad masculina prematura en el País Vasco, representada por el punto parcial Pva1, difiere mucho de lo que se podría esperar por su pirámide de edad masculina, dado que Pva5 ocupa una posición no muy alejada de Mad5.
- En Extremadura, la mortalidad de las mujeres mayores (Ext4) y en menor medida la de los hombres mayores (Ext2) obedece a tendencias claramente distintas a las seguidas por la mortalidad prematura en esta misma región (Ext 3 y 4).

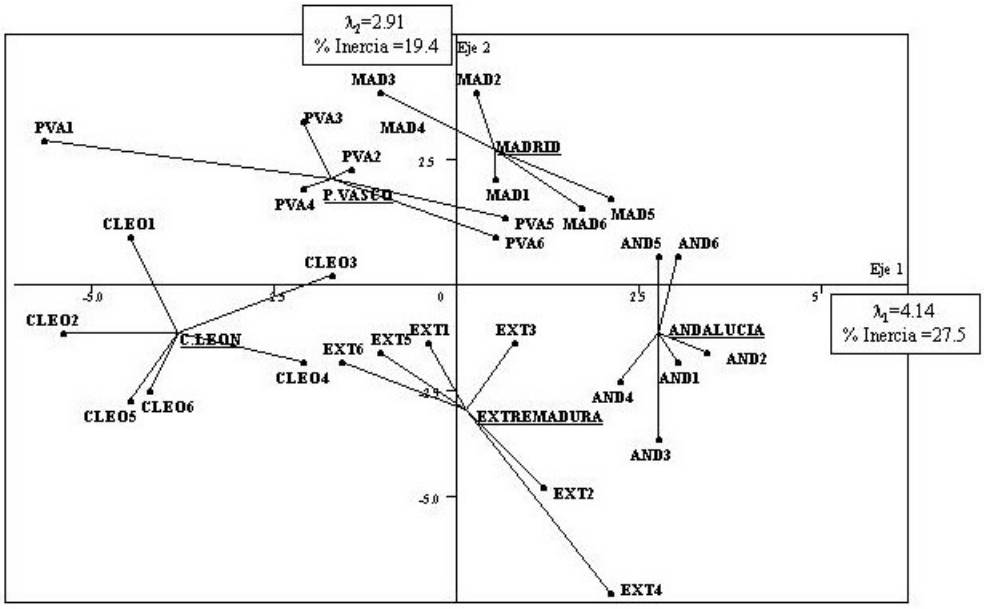

Figura 5. Extracto de la superposición de las representaciones global y parciales

5. Conclusiones

El AFMTC ofrece una herramienta descriptiva flexible para la comparación de tablas de contingencia con una sola dimensión en común. Numerosas extensiones son posibles.

Se pueden citar aquí dos que abren el camino a aplicaciones particularmente interesantes:

- La posibilidad de introducir tablas de variables de distinto tipo (continuas, cualitativas y de tipo frecuencia) como activas en un mismo análisis. Así, por ejemplo, se podrían involucrar características, continuas o cualitativas, de las regiones (indicadores económicos, medidas del sistema sanitario, etc.), para estudiar la mortalidad en referencia a dichas características en lugar de hacerlo en referencia a las pirámides de edad de las regiones.
- Obtener, como en el AFM, un gráfico en el cual cada grupo venga representado por un único punto. Este gráfico permite una comparación sintética de los grupos y una evaluación rápida de su mutua semejanza, lo que es particularmente útil cuando el número de grupos es elevado. Permitiría, por ejemplo, extender el estudio presentado aquí y analizar no sólo un año sino también la evolución de la mortalidad a lo largo de varios años.

El conjunto de estas posibilidades constituye una metodología completa para el análisis exploratorio de un conjunto de individuos o grupos de individuos descritos por datos de diferente tipo.

6. Software

AFMTC ha sido programado por el segundo autor en lenguaje Fortran (dentro del sistema ADDAD) pero se puede obtener como un programa DOS independiente, pidiéndolo a los autores. Se puede también utilizar el sistema SPAD-V (DECISIA, París), a condición de preparar los datos y los pesos de forma adecuada.

Referencias

Bécue, M., Pagés, J. (1999). Intra-Set Multiple Factor Analysis. Application to textual data. In: *Proc. of the 9th International Symposium on Applied Stochastic Models and Data Analysis,* J. Jansen et al. (Eds.). Universidade de Lisboa. Lisboa, pp. 51.60.

Bécue, M., Pagès, J. (2000). Analyse factorielle multiple intra-tableaux. Application à l'analyse simultanée de plusieurs questions ouvertes. In: *JADT2000, 5ème Journées Internationales d'Analyse statistique de Données Textuelles,* Rajman M. et Chappelier J.C. (Eds.). EPFL, Lausanne, pp. 425-432.

Benach, J., Yasui, Y., Borrell, C., Rosa, E., Pasarín, M.I., Benach, N. Español, E., Martínez J.M., Daponte, A., (2001a). Atlas of mortality in small areas in Spain (1987-1995). Universitat Pompeu Fabra. Barcelona.

Benach, J., Yasui, Y., Borrell, C., Sáez M., Pasarin, M.I., (2001b). Material deprivation and leading causes of death by gender: evidence from a nationwide small area study. *J. Epidemiol. Community Health*. 55, 239-245.

Benzécri, J.P., (1973). Analyse des Données, vol. 1: Analyse des correspondances. Dunod. Paris.

Escofier, B., Pagès, J., (1988-1998). Analyses factorielles simples et multiples; objectifs, méthodes et interprétation. Dunod. Paris.

Escofier, B., Pagès, J. (1994). Multiple factor analysis: afmult package. Comput. Statist. Data Anal. 18, 121-140.
Lebart, L., Morineau, A., Warwick, K.M. (1984). Multivariate descriptive statistical analysis. Wiley, New York.
Lebart, L., Morineau, A., Piron, M. (1998). Statistique exploratoire multidimensionnelle. Dunod. París.