# La práctica del análisis de correspondencias

**MICHAEL GREENACRE**

Catedrático de Estadística en la Universidad Pompeu Fabra

Separata del capítulo 10

## Tres ejemplos más

Primera edición: julio 2008
ISBN: 978-84-96515-71-0

Traducción: Jordi Comas Angelet
Revisión: Carles M. Cuadras Avellana



www.fbbva.es

CAPÍTULO 10

# Tres ejemplos más

Para terminar los primeros 10 capítulos de introducción al AC aplicado a tablas de contingencia de dos entradas, vamos a ver tres ejemplos más: a) una tabla que resume la clasificación de científicos de diez disciplinas de investigación en distintas categorías de financiación; b) una tabla de recuentos de 92 especies marinas en diferentes puntos de muestreo en el fondo marino; y c) un ejemplo lingüístico, en el que se ha llevado a cabo un recuento de las letras del alfabeto en muestras de textos en inglés de seis autores. Con el análisis de estos ejemplos, avanzaremos en la discusión sobre temas relacionados con las representaciones bidimensionales, como la interpretación de las dimensiones, la diferencia entre los mapas asimétricos y los mapas simétricos, y la importancia de la razón de escalas del mapa.

## Contenido



### Conjunto de datos 5: evaluación de investigadores científicos

Los datos proceden de una organización de investigación y desarrollo que clasificó a 796 investigadores científicos en cinco categorías de acuerdo con los recursos financieros de que disponían para su investigación (imagen 10.1). Hemos clasificado los investigadores según su disciplina científica (las 10 filas de la tabla) y según el tipo de financiación (las cinco columnas de la tabla). Asimismo, hemos

**Imagen 10.1:** *Frecuencias de las categorías de financiación de 796 investigadores que solicitaron fondos para la investigación: la categoría A corresponde a los que recibieron más recursos, la D a los que recibieron menos y la E a los que no recibieron*

| | Categoría de financiación | | | | | |
|---|---|---|---|---|---|---|
| Disciplina científica | *A* | *B* | *C* | *D* | *E* | *Suma* |
| Geología | 3 | 19 | 39 | 14 | 10 | *85* |
| Bioquímica | 1 | 2 | 13 | 1 | 12 | *29* |
| Química | 6 | 25 | 49 | 21 | 29 | *130* |
| Zoología | 3 | 15 | 41 | 35 | 26 | *120* |
| Física | 10 | 22 | 47 | 9 | 26 | *114* |
| Ingeniería | 3 | 11 | 25 | 15 | 34 | *88* |
| Microbiología | 1 | 6 | 14 | 5 | 11 | *37* |
| Botánica | 0 | 12 | 34 | 17 | 23 | *86* |
| Estadística | 2 | 5 | 11 | 4 | 7 | *29* |
| Matemáticas | 2 | 11 | 37 | 8 | 20 | *78* |
| *Suma* | *31* | *128* | *310* | *129* | *198* | *796* |
| *Perfil fila medio* | *3,9%* | *16,1%* | *38,9%* | *16,2%* | *24,9%* | |

etiquetado las categorías de financiación como *A*, *B*, *C*, *D* y *E*, de más a menos recursos financieros. En realidad, las categorías de la *A* a la *D* corresponden a investigadores que han disfrutado de recursos de investigación, de la *A* (los que recibieron más) hasta la *D* (los que recibieron menos), mientras que categoría *E* corresponde a los científicos que no consiguieron financiación (es decir, sus proyectos de investigación fueron rechazados).

Descomposición de la inercia

Esta tabla de 10 × 5 se halla exactamente en un espacio tetradimensional. La descomposición de la inercia en los dos primeros ejes principales es la siguiente:

| *Dimensión* | *Inercia principal* | *Porcentaje de inercia* |
|---|---|---|
| 1 | 0,03912 | 47,2% |
| 2 | 0,03038 | 36,7% |

Expresamos la inercia explicada por cada eje como porcentaje. Así, las dos primeras dimensiones explican casi el 84% de la inercia. La suma de las inercias principales es de 0,082879; por tanto, el estadístico $\chi^2$ es igual a 0,082879 × 796 = 65,97. Si hiciéramos una prueba estadística, utilizando la distribución $\chi^2$ con 9 × 4 = 36 grados de libertad, veríamos que se trata de un valor altamente significativo ($p$ = 0,002).

Mapa asimétrico de perfiles fila

En la imagen 10.2 hemos representado el mapa asimétrico de los perfiles fila y los vértices columna. En esta representación gráfica vemos que el grado de la asociación entre las disciplinas científicas y las categorías de financiación es bastante baja; es decir, los perfiles no se alejan demasiado de la media (compárese con las figuras de la imagen 4.2). Esta situación es bastante típica de los datos derivados

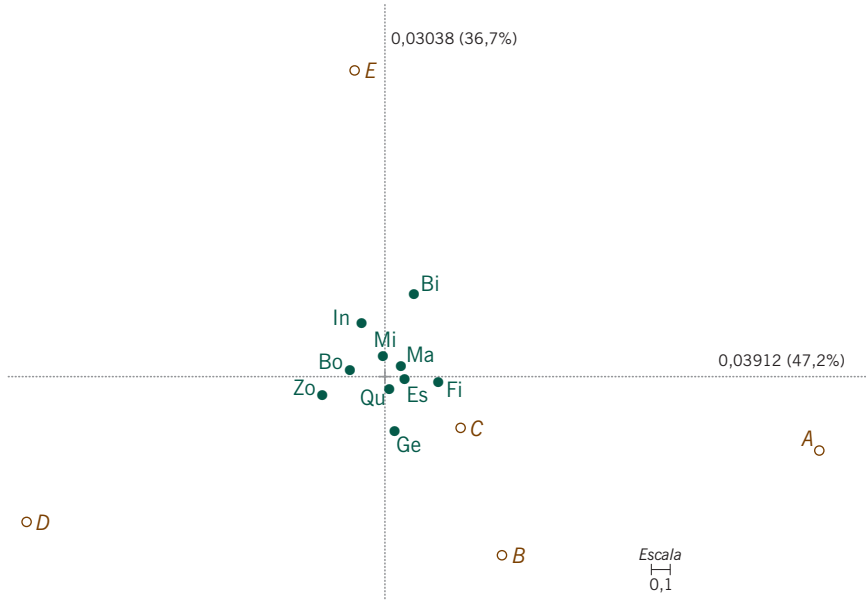
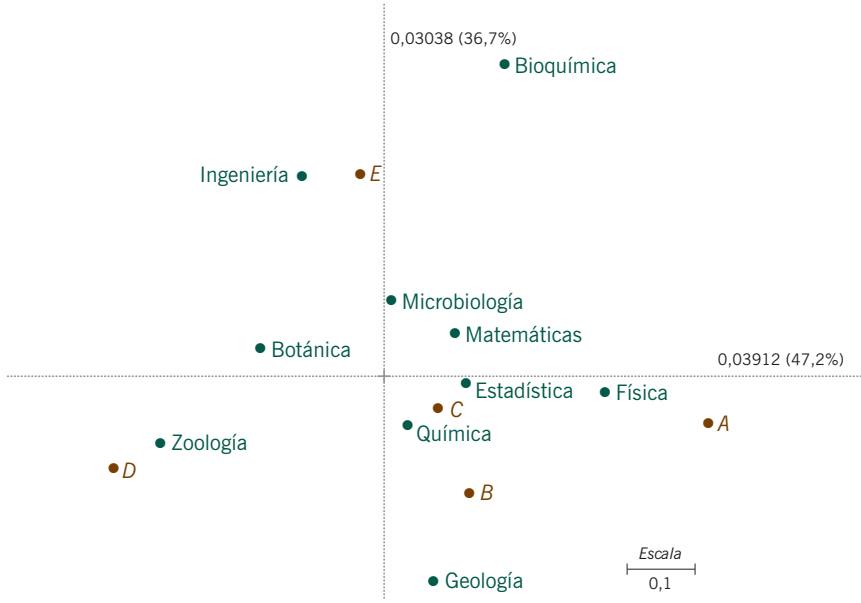



**Imagen 10.2:**
*Mapa asimétrico de los perfiles fila de la tabla 10.1 (datos sobre la financiación de la investigación científica)*

de las ciencias sociales. Por tanto, para esta tabla, el mapa asimétrico no es muy útil ya que todos los perfiles se hallan apelotonados en el centro de la figura —están tan cerca unos de otros que no podemos escribir las etiquetas completas, nos tenemos que conformar con las dos primeras letras de cada disciplina—. Sin embargo, si nos fijamos en las posiciones de los vértices, podemos interpretar fácilmente el espacio. La dimensión horizontal alinea en su orden natural las cuatro categorías de financiación, de la *D* (menor financiación) a la *A* (mayor financiación), las categorías *B* y *C* se hallan juntas, en medio. La dimensión vertical opone la categoría *E* (sin financiación) a las otras categorías, de manera que la interpretación es bastante directa. Cuanto más arriba se halle una disciplina, menos proyectos de los investigadores habrán conseguido financiación. Cuanto más a la derecha se halle una disciplina, más financiación habrán recibido los proyectos de los investigadores. Utilizando la terminología de la investigación de mercados, podríamos decir que «el punto ideal» se hallaría abajo a la derecha: más proyectos de investigación aceptados (abajo) y proyectos con buena financiación (a la derecha). Por tanto, si hiciéramos un estudio de tendencias en función del tiempo, las disciplinas tendrían que desplazarse hacia abajo a la derecha para mostrar una mejora en su estatus de financiación. De momento no existen disciplinas en esta zona, aunque la Física es la que se halla más a la derecha (mayor porcentaje [10 de 114, el 8%] de investigadores de la categoría *A*), pero verticalmente se halla en medio, ya que el porcentaje de investigadores que no han conseguido recursos se halla cerca de la media (26 de 114 no han conseguido financiación, el 22,8%, en comparación con la media global que es de 198 de 796, el 24,9%).

**Imagen 10.3:**
*Mapa simétrico de la tabla de la imagen 10.1 (datos sobre la financiación de la investigación científica)*




Mapa simétrico

En la imagen 10.3 mostramos el mapa simétrico de los mismos datos. La única diferencia entre este último mapa y el mapa de la imagen 10.2 es que, en vez de representar las columnas como vértices, las hemos representado como perfiles. De manera que se produce un cambio de escala que amplía la representación de los perfiles fila. Esta ampliación en la configuración de las disciplinas nos facilita la interpretación de sus posiciones relativas y nos deja más espacio para escribir las etiquetas completas. Ahora podemos ver más fácilmente las posiciones relativas de las disciplinas. Por ejemplo, Geología, Estadística, Matemáticas y Bioquímica se hallan todas en la misma posición, con relación al primer eje, lo que no ocurre con relación al segundo eje. Esto significa que los investigadores de estas disciplinas, cuyos proyectos de investigación han sido financiados, tienen posiciones similares con relación a las categorías de financiación de la *A* a la *D*. Sin embargo, Geología tiene muchos menos proyectos «no aceptados» (el 11,8% en la categoría *E*) que Bioquímica (41,4%). En esta representación simétrica no podemos valorar gráficamente el grado de asociación total (inercia) entre filas y columnas. Solamente podemos valorar la asociación a partir del valor de las inercias principales de cada eje, o a partir de sus raíces cuadradas, es decir, de las correlaciones canónicas en cada eje, concretamente $\sqrt{0{,}039117} = 0{,}198$ y $\sqrt{0{,}030381} = 0{,}174$, respectivamente. Sólo podemos evaluar gráficamente el grado de asociación entre filas y columnas en mapas asimétricos como el mapa de la imagen 10.2 (comparemos de nuevo los distintos grados de asociación que vimos en las figuras de la imagen 4.2).

| | Estación (muestra) | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Especies | E4 | E8 | E9 | E12 | E13 | E14 | E15 | E18 | E19 | E23 | E24 | R40 | R42 |
| *Myri.ocul.* | 193 | 79 | 150 | 72 | 141 | 302 | 114 | 136 | 267 | 271 | 992 | 5 | 12 |
| *Chae.seto.* | 34 | 4 | 247 | 19 | 52 | 250 | 331 | 12 | 125 | 37 | 12 | 8 | 3 |
| *Amph.falc.* | 49 | 58 | 66 | 47 | 78 | 92 | 113 | 38 | 96 | 76 | 37 | 0 | 5 |
| *Myse.bide.* | 30 | 11 | 36 | 65 | 35 | 37 | 21 | 3 | 20 | 156 | 12 | 58 | 43 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| *Eucl.sp.* | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| *Scal.infl.* | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| *Eumi.ocke.* | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| *Modi.modi.* | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |

**Imagen 10.4:** *Frecuencias de 92 especies marinas en 13 muestras (las dos últimas son muestras de referencia); hemos ordenado las especies (filas) en orden descendiente de abundancia total; mostramos las cuatro especies más abundantes y las cuatro menos abundantes*

Interpretación de las dimensiones de los mapas

Tanto si recurrimos a mapas simétricos como asimétricos, la interpretación *dimensional* de los mapas siempre es la misma. Es decir, tenemos que interpretar los ejes uno a uno, como hicimos anteriormente. Así es, por ejemplo, lo habitual en el análisis factorial, cuando utilizamos las posiciones relativas de los puntos —«las variables» de la tabla— para asignar nombres descriptivos a los ejes. Es lo que hicimos anteriormente al describir en primer lugar los ejes a partir de las posiciones relativas de las categorías de financiación para, a continuación, interpretar las posiciones relativas de las disciplinas con relación a los ejes. Sin embargo, en este tipo de interpretación, todas las afirmaciones son relativas. Es decir, no podemos evaluar de forma absoluta las diferencias entre los perfiles de financiación de las distintas disciplinas a no ser que nos refiramos a los datos originales. Dicho de otra manera, con otros datos, aunque obtuviéramos mapas simétricos similares a los de la imagen 10.3, el grado de asociación entre los perfiles de financiación y las disciplinas podría ser mucho mayor (o menor).

Conjunto de datos 6: abundancia de especies en muestras del fondo marino

El AC se utiliza ampliamente para analizar datos en ecología. El segundo ejemplo que presentamos hace referencia a un conjunto de datos usual en biología marina. Los datos, que mostramos parcialmente en la tabla de la imagen 10.4, corresponden a los recuentos de 92 especies marinas identificadas en 13 muestras del fondo marino del mar del Norte. La mayor parte de las muestras se obtuvieron cerca de una plataforma petrolífera que producía una cierta contaminación del fondo marino. Existen dos muestras, utilizadas como referencia, supuestamente no contaminadas, que se obtuvieron lejos de la zona de influencia de la plataforma petrolífera. Estos datos, y en general este tipo de datos biológicos, se caracterizan por presentar una gran variabilidad, que podemos percibir fácilmente inspeccionando la pequeña parte de datos que aquí proporcionamos. La inercia total de esta tabla es de 0,7826, mucho mayor que la de los ejemplos anteriores. En consecuencia, cabe esperar que los perfiles mostrarán mucha más dispersión con relación a los vértices. Fijémonos que, en este ejemplo, no pode-

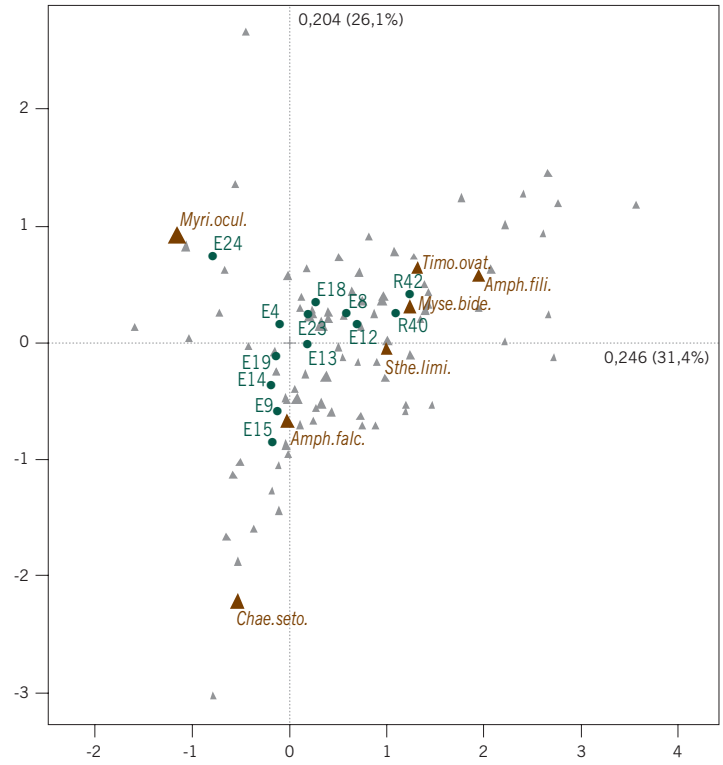



**Imagen 10.5:**
*Mapa asimétrico del AC, con las estaciones de muestreo en coordenadas principales y las especies en coordenadas estándares. Los símbolos de las especies son proporcionales a la abundancia de las especies (masa); hemos etiquetado con las primeras letras de su nombre científico a algunas especies importantes para el análisis, situando la etiqueta al lado de su símbolo triangular. La inercia explicada por el mapa es del 57,5%*

mos utilizar la prueba $\chi^2$, ya que los datos no constituyen una verdadera tabla de contingencia (los recuentos no son independientes, ya que los organismos marinos a menudo se presentan agrupados en los puntos de muestreo).

Mapa asimétrico del AC de los datos sobre abundancia de especies

En el mapa de la imagen 10.5 podemos ver el mapa asimétrico de los perfiles de las muestras (columnas) y de los vértices de las especies (filas). Dado que tenemos 92 especies, no hemos podido etiquetar todos los puntos. Sólo hemos etiquetado las especies que contribuyen de forma más notable a la configuración del mapa, en general son las especies más abundantes. En el capítulo 11 veremos cómo podemos medir la contribución de cada punto; por el momento digamos simplemente que 10 de las 92 especies contribuyen en la construcción de este mapa en más de un 85% (podríamos eliminar las restantes 82 especies sin que el mapa cambiara demasiado). Podemos observar que las estaciones de muestreo describen una curva desde la parte baja a la izquierda (las estaciones de muestreo más contaminadas) hasta la parte superior derecha (las menos contaminadas). Las estaciones de referencia quedan arriba a la derecha. Una excepción es la estación de muestreo 24, que claramente se separa de las restantes, principalmente debido a su gran abundancia en *Myri.ocul.* (*Myriochele oculata*), lo podemos comprobar en la primera fila de la imagen 10.4.

| Libro | Letras | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | *a* | *b* | *c* | *d* | *e* | ⋯ | *w* | *x* | *y* | *z* | *Suma* |
| TD-Buck | 550 | 116 | 147 | 374 | 1.015 | ⋯ | 155 | 5 | 150 | 3 | *7144* |
| EW-Buck | 557 | 129 | 128 | 343 | 996 | ⋯ | 187 | 10 | 184 | 4 | *7479* |
| Dr-Mich | 515 | 109 | 172 | 311 | 827 | ⋯ | 156 | 14 | 137 | 5 | *6669* |
| As-Mich | 554 | 108 | 206 | 243 | 797 | ⋯ | 149 | 2 | 80 | 6 | *6510* |
| LW-Clar | 590 | 112 | 181 | 265 | 940 | ⋯ | 146 | 13 | 162 | 10 | *7100* |
| PF-Clar | 592 | 151 | 251 | 238 | 985 | ⋯ | 106 | 15 | 142 | 20 | *7505* |
| FA-Hemi | 589 | 72 | 129 | 339 | 866 | ⋯ | 225 | 1 | 155 | 2 | *6877* |
| Is-Hemi | 576 | 120 | 136 | 404 | 873 | ⋯ | 250 | 3 | 104 | 5 | *6924* |
| SF7-Faul | 541 | 109 | 136 | 228 | 763 | ⋯ | 160 | 11 | 280 | 1 | *6885* |
| SF6-Faul | 517 | 96 | 127 | 356 | 771 | ⋯ | 216 | 12 | 171 | 5 | *6971* |
| Pe3-Holt | 557 | 97 | 145 | 354 | 909 | ⋯ | 194 | 9 | 140 | 4 | *6650* |
| Pe2-Holt | 541 | 93 | 149 | 390 | 887 | ⋯ | 218 | 2 | 127 | 2 | *6933* |

Abreviaciones:
TD (Three Daughters), EW (East Wind) -Buck (Pearl S. Buck)
Dr (Drifters), As (Asia) -Mich (James Michener)
LW (Lost World), PF (Profiles of Future) -Clar (Arthur C. Clarke)
FA (Farewell to Arms), Is (Islands) -Hemi (Ernest Hemingway)
SF7 y SF6 (Sound and Fury, capítulos 7 y 6) -Faul (William Faulkner)
Pe3 y Pe2 (Bride of Pendorric, capítulos 3 y 2) -Holt (Victoria Holt)

**Imagen 10.6:**
*Recuento de las letras en 12 muestras de textos de libros de seis autores distintos, que muestran datos de 9 de las 26 letras del alfabeto inglés*

Como comentábamos anteriormente, hemos etiquetado las especies más abundantes, que son precisamente las que más determinan el mapa. Fijémonos en que los datos de este ejemplo se adaptan bien a un mapa asimétrico, sin duda debido a que existe una gran variabilidad entre las muestras, típico de los datos en ecología y, por tanto, la inercia es muy grande. ¡En el próximo ejemplo ocurre todo lo contrario!

Conjunto de datos 7: frecuencia de las letras en libros de seis autores

Hemos incluido este sorprendente ejemplo en el paquete **ca** del programa R (véase el apéndice de cálculo, B). Los datos forman una matriz de 12 × 26, las filas representan 12 textos que configuran seis pares, cada par contiene textos de un mismo autor (en la la imagen 10.6 mostramos parte de esta matriz). Las columnas son las 26 letras del alfabeto inglés, de la *a* a la *z*. Los datos son recuentos de letras en muestras constituidas por un texto de cada uno de los libros. Tenemos aproximadamente 6500-7500 recuentos de letras de cada libro o capítulo.

Una de las inercias más bajas, pero con una estructura significativa

Este conjunto de datos tiene una de las menores inercias totales que nunca he visto en mi experiencia con el AC. La inercia total es de 0,01873, lo que significa que los datos se hallan muy cerca de los valores esperados calculados a partir de las frecuencias marginales. Es decir, los perfiles son prácticamente idénticos. En la imagen 10.7, mostramos el mapa asimétrico de estos datos, con las letras en los vértices y los 12 textos formando una especie de pequeña mancha alrededor del origen, lo que indica

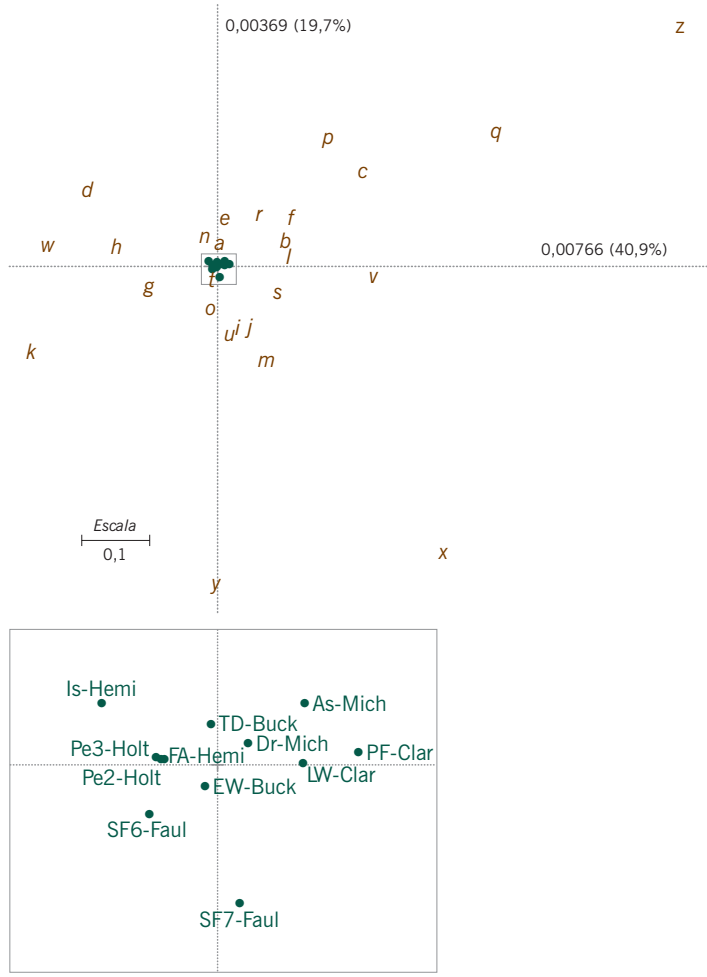



**Imagen 10.7:** *Mapa asimétrico de los datos sobre autores de la imagen 10.6, con las filas (textos) en coordenadas principales. La muy baja inercia de la tabla queda patente por la proximidad de los perfiles fila al centroide. Una «ampliación» del rectángulo situado en el centro del mapa muestra las posiciones relativas de los perfiles fila*

la poca variación que existe entre los textos en términos de distribución de las letras, por otra parte un resultado esperable. Es sorprendente ver cómo al ampliar esta pequeña mancha de puntos, de una pequeña variación, surge una estructura muy marcada. Efectivamente, los pares de textos del mismo autor aparecen juntos, resultado muy significativo desde un punto de vista estadístico (en el capítulo 25 veremos la prueba de permutación que nos permitirá contrastar esta afirmación).

La necesidad de mantener una razón de escalas de los mapas igual a 1

Vamos a realizar un importante comentario final sobre las representaciones gráficas de los mapas bidimensionales en el análisis de correspondencias. Dado que, en los mapas, las distancias son especialmente importantes, es evidente que una unidad de longitud en el eje horizontal debe ser igual a una unidad en el eje vertical. A pesar de que este requisito es obvio, en muchos programas informáticos y

hojas de cálculo se pasa por alto esta consideración y se dibujan diagramas de dispersión con los ejes en escalas distintas. Sabemos que, en general, los puntos presentan poca variación en el segundo eje (vertical); sin embargo, se suelen representar los mapas en rectángulos predefinidos que exageran este segundo eje. La *razón de escalas* del mapa, es decir el cociente entre una unidad de longitud horizontal y una unidad en el eje vertical debería ser igual a 1. Al final del apéndice de cálculo, B, presentamos algunas opciones para generar mapas de buena calidad.

RESUMEN:
Tres ejemplos más

1. Cuando sea posible, es útil contrastar —utilizando la prueba $\chi^2$— si la asociación entre las filas y las columnas de una tabla de contingencia es significativa. Sin embargo, la significación estadística no es un requerimiento crucial para el análisis de mapas. Podemos ver el AC como una manera de expresar datos en forma gráfica para facilitar su interpretación; así tiene sentido representar cualquier tabla.
2. La interpretación *dimensional* de los mapas es siempre igual, tanto si recurrimos a mapas simétricos como a mapas asimétricos. Es decir, tenemos que interpretar los ejes uno a uno. Basamos la interpretación en asignar nombres descriptivos a los ejes principales a partir de las posiciones relativas de los puntos de uno de los dos conjuntos de coordenadas. A continuación, interpretamos las posiciones relativas del otro conjunto de coordenadas con relación a las dimensiones que previamente hemos asignado nombres descriptivos.
3. Los mapas asimétricos van bien cuando la inercia es alta, pero resultan problemáticos cuando la inercia total es pequeña. Ello es debido a que las coordenadas principales se hallan demasiado cerca del origen, lo que complica el etiquetado.
4. Es importante que las utilidades de representación gráfica mantengan la *razón de escalas* de los mapas. Una unidad en el eje horizontal debe aproximarse tanto como sea posible a una unidad en el eje vertical. Cuando las escalas son distintas, las distancias se distorsionan.