La práctica del análisis de correspondencias

MICHAEL GREENACRE

Catedrático de Estadística en la Universidad Pompeu Fabra

Separata del capítulo 21

Análisis de correspondencias de subgrupos

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

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

www.fbbva.es

CAPÍTULO 9

Análisis de correspondencias de subgrupos

A menudo conviene analizar sólo una parte de la matriz de datos, dejando fuera algunas filas y/o columnas. Así, por ejemplo, puede que nos convenga analizar de forma separada las columnas en grupos que formemos siguiendo algún criterio sustantivo. O puede ocurrir, por ejemplo, que sea conveniente excluir del análisis categorías con valores perdidos. En estas situaciones podríamos aplicar directamente el AC a la submatriz de interés; pero, al proceder de esta manera, puede ocurrir que los valores marginales de uno o ambos márgenes de la submatriz sean distintos de los de la matriz original y, en consecuencia, cambien los perfiles, las masas y las distancias. Sin embargo, en el *análisis de correspondencias de subgrupos*, el procedimiento de análisis que presentamos en este capítulo para la determinación de masas y distancias χ^2 de cualquier submatriz, utilizamos los valores marginales originales de la matriz completa. Este tipo de análisis tiene muchas ventajas; así, por ejemplo, nos permite descomponer la inercia total de la matriz original en las distintas submatrices, de manera que la información, contenida en la matriz de datos original queda recogida en las submatrices analizadas.

Contenido

Análisis de consonantes y vocales en los datos sobre autores

Los datos sobre autores en lengua inglesa de la tabla de la imagen 10.6 son un buen ejemplo de tabla que podemos dividir de forma natural (el alfabeto inglés está formado por 21 consonantes y 5 vocales). En el capítulo 10, vimos que la inercia total de esta tabla era muy baja, 0,01873 pero, sin embargo existía una clara estructura en las filas (los 12 textos de seis autores). Podría ser interesante ver el resultado del AC de vocales y de consonantes de forma separada. Una posibilidad sería analizar sin más las dos submatrices; la submatriz de 12 × 21 de frecuencias de consonantes y la submatriz de 12 × 5 de frecuencias de vocales. No obstante, proceder de esta manera implicaría recalcular los valores de los perfiles de cada texto en relación con los valores marginales de las nuevas submatrices. Por ejemplo, en el análisis de consonantes, calcularíamos las frecuencias relativas de *b*, *c*, *d*, *f*, ..., etc. en relación con el número total de consonantes del texto, y no en relación con el número total de letras. De esta manera, la masa de cada texto sería proporcional al recuento de consonantes, y no al número total de letras. Los perfiles de las consonantes se mantendrían invariables, sin embargo, las distancias χ^2 entre ellos serían distintas ya que éstas dependen de las masas de las filas que, como hemos comentado, han cambiado.

El análisis de subgrupos mantiene invariables los valores marginales originales

El análisis de correspondencias de subgrupos es una aproximación alternativa, con muchas ventajas. En dicho análisis para el cálculo de masas y de distancias utilizamos los valores marginales de la matriz original. Ello conlleva introducir una modificación muy simple en el algoritmo de cálculo del AC: todo lo que tenemos que hacer es suprimir los cálculos de las sumas marginales «locales» de la submatriz seleccionada, y mantener los valores de los cálculos realizados con la matriz original.

AC de subtablas: análisis de consonantes, biplot estándar

Aplicando el AC de subgrupos a la tabla de frecuencias de consonantes (págs. 315-316), obtenemos el mapa de la imagen 21.1. Aquí en vez de mostrar el mapa simétrico o el asimétrico, mostramos el biplot estándar del AC (cap. 13). Dado que los textos están en coordenadas principales, las distancias entre puntos son aproximadamente distancias χ^2. En el cálculo de las distancias χ^2 sólo tenemos en cuenta las consonantes; dejamos fuera las vocales. Expresamos las consonantes en coordenadas estándares multiplicadas por las raíces cuadradas de las correspondientes frecuencias relativas de las consonantes (es decir, las frecuencias relativas de las 26 letras; recordemos que las sumas marginales son siempre las de la tabla original). En cada eje, las raíces cuadradas de las longitudes de los vectores de consonantes son proporcionales a sus contribuciones al eje. Razón por la cual la letra *y* es tan importante en el segundo eje (más del 50%). Los biplots estándares funcionan igual de bien para tablas con inercias bajas o altas, y en este ejemplo, con una inercia extremadamente baja, es particularmente útil. Comparando este mapa con el mapa asimétrico de la imagen 10.7, vemos que las letras apuntan más o menos hacia las mismas direcciones. También vemos que la configuración de los textos es bastante similar. La inercia total del mapa es de 0,01637, valor exactamente igual a la suma de las inercias de las consonantes del análisis completo. En la página 111

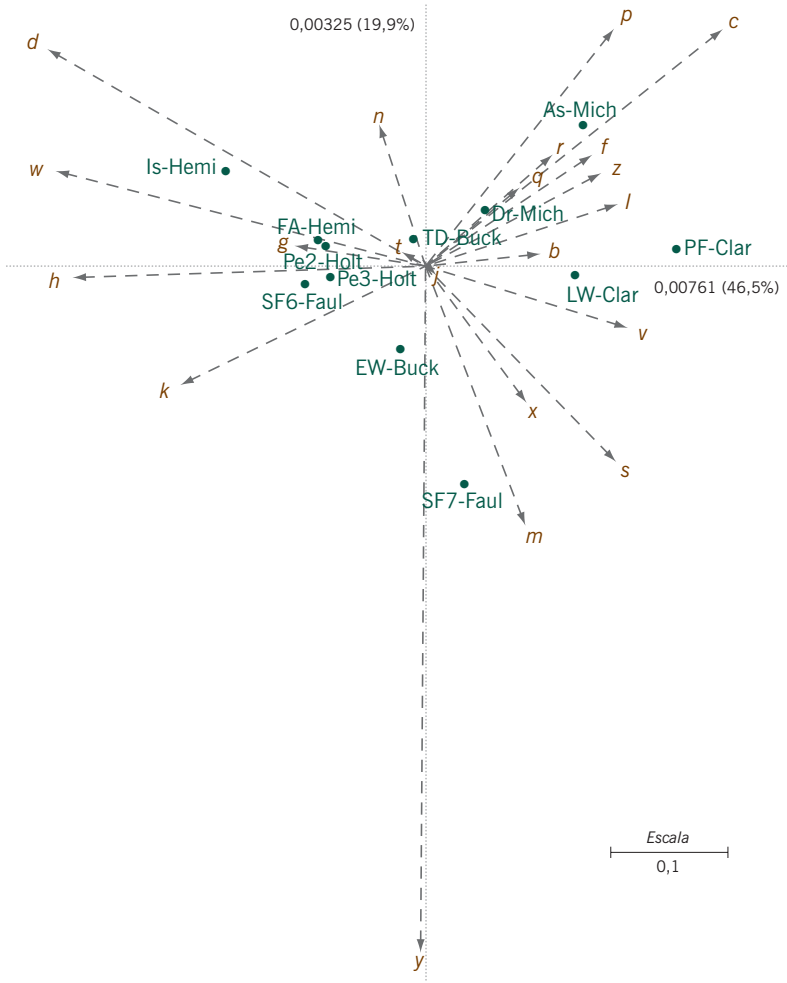


Imagen 21.1:
AC del subgrupo de consonantes del ejemplo de los autores; biplot estándar de filas, es decir, filas (textos) en coordenadas principales y columnas (letras) en coordenadas estándares multiplicadas por la raíces cuadradas de las masas de las columnas

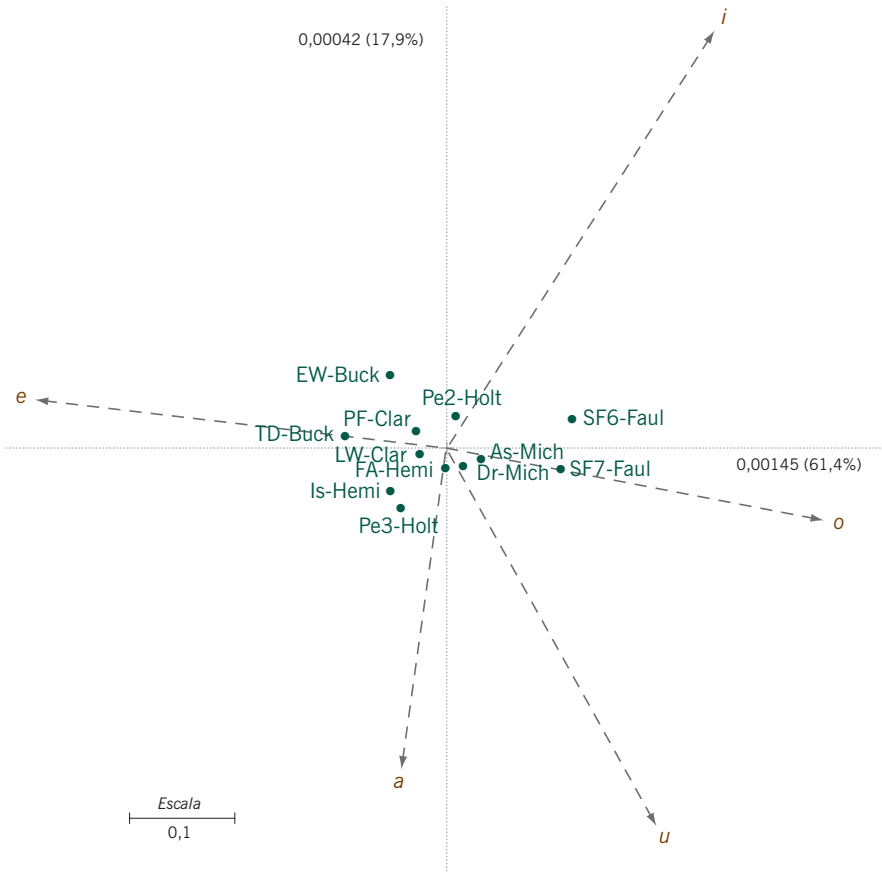
vimos que la inercia total de la tabla completa era de 0,01873; por tanto, la inercia atribuible a las consonantes es del 87,4% (0,01637 con relación a 0,01873) de la inercia total. Así, pues, una vez visto que la mayor parte de la inercia es atribuible a las consonantes, no debe sorprendernos que los mapas del análisis completo de la imagen 10.7 y en el análisis de subgrupos de la imagen 21.1 sean muy similares.

AC de subgrupos: análisis de vocales, biplot estándar

Hemos visto que podemos descomponer la inercia total de la tabla original en inercia de consonantes e inercia de vocales de la siguiente manera:

inercia total = inercia de consonantes + inercia de vocales
0,01873 = 0,01637 + 0,00236
(87,4%) (12,6%)

Imagen 21.2:
Análisis de subgrupos de vocales en el ejemplo sobre los autores; biplot estándar de filas, es decir, filas (textos) en coordenadas principales y columnas (letras) en coordenadas estándares multiplicadas por las raíces cuadradas de las masas de las columnas



A pesar de que las vocales son letras relativamente más frecuentes (las cinco vocales representan el 38,3% del total de letras, mientras que las 21 consonantes representan el 61,7%), la inercia de la submatriz de vocales es mucho menor, sólo el 12,6% de la inercia total original. En el mapa de la imagen 21.2 mostramos, igual que antes para las consonantes, el biplot estándar de las vocales. La menor dispersión de los textos con relación a los vectores de las letras es muy aparente y contrasta con el mapa de la imagen 21.1. Sin embargo, bastantes pares de textos siguen hallándose muy cerca. Podemos observar que las letras *e* y *o* se hallan en posiciones opuestas, a la izquierda y a la derecha, respectivamente. Asimismo se hallan en posiciones opuestas los textos de Buck y los de Faulkner. De los seis autores, los textos de Holt parecen ser los más distintos. En el capítulo 25 veremos las pruebas de permutaciones que nos permitirán contrastar la significación de estos resultados. Anticipándonos un poco, podemos indicar que el emparejado de textos en los mapas es altamente significativo tanto para las consonantes como para las vocales.

ACM de subgrupos

Cuando trabajemos con datos categóricos multivariantes, la idea de dividir la matriz original en submatrices y luego aplicar el ACM es muy útil para investigar si existen estructuras en determinadas submatrices. Así, en datos procedentes de encuestas puede ser interesante, desde un punto de vista sustantivo, centrarnos en un determinado subgrupo de respuestas. Por ejemplo, centrarnos sólo en las categorías de acuerdo, en una escala de acuerdo/desacuerdo con cinco respuestas posibles, o centrarnos únicamente en las respuestas intermedias («ni de acuerdo ni en desacuerdo»), o en respuestas no sustantivas («no sabe», «no contesta», «otras», etc.). O, simplemente, podría ser que quisiéramos excluir las respuestas no sustantivas y concentrarnos sólo en las que sí lo son. En todos estos casos, el análisis de subgrupos nos permitirá ver de forma más clara la relación entre este tipo especial de respuestas y las variables demográficas, lo que posiblemente no ocurriría si analizáramos todas las respuestas conjuntamente. La posibilidad de hacer submatrices nos permite, para diferentes grupos de categorías, dividir la variabilidad de los datos en partes que, luego, podemos visualizar separadamente. La manera de realizar el ACM de subgrupos consiste en llevar a cabo el AC de subgrupos a las partes adecuadas de la matriz binaria o de la matriz de Burt, como veremos a continuación.

Análisis de subgrupos de la matriz binaria

Volvamos a los datos sobre el trabajo de las mujeres que introducimos en el capítulo 17 y analizamos en el capítulo 18 utilizando el ACM. Las cuatro preguntas tienen prevista una categoría, etiquetada en los mapas con el símbolo *?*, para las respuestas del tipo «no sabe» y las respuestas perdidas. Estas categorías tienen un papel muy prominente en el primer eje principal del ACM (imagen 18.2). Vamos realizar un análisis de subgrupos restringido a las respuestas sustantivas de las cuatro variables, etiquetadas como *T* (trabajo a tiempo completo), *t* (trabajo a tiempo parcial) y *C* (permanecer en casa), prescindiendo así de las columnas de la matriz binaria correspondientes a respuestas no sustantivas *?*; en el análisis utilizaremos los valores marginales de filas y de columnas de la matriz binaria original. Dado que en este caso en particular, las sumas de las filas de la matriz binaria son iguales a 4, en el análisis de subgrupos para la ponderación de las filas (encuestados) mantendremos este valor. Los valores de los perfiles seguirán siendo ceros o $\frac{1}{4}$. Es decir, los encuestados con cuatro respuestas sustantivas tendrán cuatro valores $\frac{1}{4}$ en sus perfiles, mientras que los que tengan tres respuestas sustantivas tendrán tres valores $\frac{1}{4}$ y así sucesivamente. Por el contrario, si simplemente prescindiéramos de las columnas no sustantivas y lleváramos a cabo el AC ordinario sobre la matriz binaria, tendríamos valores de $\frac{1}{3}$ para los encuestados con tres respuestas sustantivas, $\frac{1}{2}$ con dos, y 1 con sólo una. Además, sería imposible calcular los perfiles de los casos con cuatro respuestas no sustantivas. Cosa que no ocurre en el AC de subgrupos. Este tipo de casos con sólo ceros se sitúan en el origen en el mapa. La inercia total del AC de subgrupos, con 12 categorías, es de 2,1047. Dado que la inercia total la matriz binaria completa es de 3, vemos que

Imagen 21.3:
Matriz de Burt de las cuatro variables categóricas de la imagen 18.4, arreglada de manera que todas las categorías correspondientes a respuestas no sustantivas (?) se hallan en la últimas filas y columnas. Todas las respuestas sustantivas (T, t y C), 12 × 12, se hallan en la parte superior izquierda, mientras que la esquina inferior izquierda de 4 × 4 contiene la concurrencia de las respuestas no sustantivas («no sabe/valores perdidos»)

1T	1t	1C	2T	2t	2C	3T	3t	3C	4T	4t	4C	1?	2?	3?	4?
2501	0	0	172	1107	1131	355	1710	345	1766	538	40	0	91	91	157
0	476	0	7	129	335	16	261	181	128	293	17	0	5	18	38
0	0	79	1	6	72	1	17	61	14	21	38	0	0	0	6
172	7	1	181	0	0	127	48	4	165	15	0	1	0	2	1
1107	129	6	0	1299	0	219	997	61	972	239	13	57	0	22	75
1131	335	72	0	0	1646	24	989	573	760	616	84	108	0	60	186
355	16	1	127	219	24	379	0	0	360	14	1	7	9	0	4
1710	261	17	48	997	989	0	2084	0	1348	567	23	96	50	0	146
345	181	61	4	61	573	0	0	642	202	286	73	55	4	0	81
1766	128	14	165	972	760	360	1348	202	1959	0	0	51	62	49	0
538	293	21	15	239	616	14	567	286	0	897	0	45	27	30	0
40	17	38	0	13	84	1	23	73	0	0	97	2	0	0	0
0	0	0	1	57	108	7	96	55	51	45	2	362	196	204	264
91	5	0	0	0	0	9	50	4	62	27	0	196	292	229	203
91	18	0	2	22	60	0	0	0	49	30	0	204	229	313	234
157	38	6	1	75	186	4	146	81	0	0	0	264	203	234	465

la inercia se ha descompuesto en 2,1047 (70,2%) para las categorías sustantivas y 0,8953 (29,8%) para las no sustantivas. Las inercias principales y los porcentajes de inercia de las dos primeras dimensiones de la submatriz analizada son de 0,5133 (24,4% del total de 2,1047) y de 0,3652 (17,4%) sobre este mismo total. Por tanto, el porcentaje global de inercia explicada por la solución bidimensional es del 41,8%. Igual que ocurría en el ACM, estos porcentajes son artificialmente bajos. Como vimos en el capítulo 19, y como veremos a continuación, podemos mejorar el mapa implementando un ajuste de los factores de escala de los ejes.

Análisis de subgrupos de la matriz de Burt

Igual que vimos en el ACM, podemos mejorar el mapa del AC de subgrupos llevando a cabo el análisis en la parte apropiada de la matriz de Burt. Para ilustrar este procedimiento, consideremos la matriz de Burt que vimos en la imagen 18.4 del capítulo 18. Podemos reacomodar dicha matriz de manera que todas las categorías de la submatriz de interés se hallen, como mostramos en la imagen 21.3, en la parte superior izquierda de la tabla concatenada. Así, la submatriz de interés de 12 × 12 está formada por cuatro tablas con tres respuestas sustantivas en cada una de ellas. Ahora, las cuatro respuestas no sustantivas se hallan en las últimas cuatro filas y cuatro columnas de la matriz. El AC de subgrupos da una inercia total de 0,6358, y unas inercias principales (y porcentajes) de 0,26354 (41,4%) y de 0,1333 (21,0%) para las dos primeras dimensiones. Igual que ocurría con el ACM,

obtenemos una mejora del mapa con relación al AC de subgrupos llevado a cabo en la matriz binaria. Ahora llegamos a explicar el 62,4% de la inercia, mientras que con el análisis anterior llegábamos a explicar el 41,8% de la inercia. Fijémonos también en que la relación entre el AC de subgrupos en la matriz binaria y en la matriz de Burt es la misma que vimos para el ACM habitual: las inercias principales en el AC de subgrupos en la matriz de Burt son los cuadrados de las de la matriz binaria, así, por ejemplo, $0{,}2635 = 0{,}5133^2$.

Análisis de subgrupos con una solución e inercias ajustadas

El problema de las bajas inercias es el mismo que vimos con el ACM. Efectivamente, en la diagonal de la tabla concatenada de la imagen 21.3, podemos ver matrices diagonales de 3×3 que interfieren en los resultados del AC del subgrupo de interés. Al igual que antes, es posible ajustar el resultado mediante análisis de la regresión, de manera que se ajusten de forma óptima las matrices que se hallan fuera de la diagonal. Esto implica disponer en forma de vector los elementos de las seis tablas situadas fuera de la diagonal, cada una de ellas con nueve elementos, como si fuera un vector de 54 elementos y, así, constituir los elementos de la variable *«y»* de la regresión. Tenemos que expresar estos elementos como en (19.2), es decir, como cocientes de contingencia menos 1. Formaremos las dos variables *«x»* (para la solución bidimensional) multiplicando las correspondientes coordenadas estándares. Hallaremos los valores óptimos de los factores de escala como anteriormente, por mínimos cuadrados ponderados (cap. 19), y así obtenemos un ajuste de $R^2 = 0{,}849$. Desafortunadamente en este caso no parece que exista atajo alguno, como ocurría para el ACM [ecuaciones (19.5) y (19.7)]. A partir de la regresión de mínimos cuadrados ponderada obtenemos los factores de escala 0,3570 y 0,1636, que hemos utilizado para obtener las coordenadas principales y el mapa de la imagen 21.4. Los cuadrados de estos factores de escala son las inercias principales, 0,1275 y 0,0268, que mostramos en los ejes. El porcentaje de inercia explicado por la solución bidimensional ajustada, R^2, es del 84,9% (como vimos anteriormente, no podemos calcular los porcentajes de los ejes individuales ya que la solución no es anidada).

Puntos adicionales en el AC de subgrupos

El procedimiento de representación de puntos adicionales depende de cómo hayamos dividido en grupos las filas o las columnas. Así, por ejemplo, en el caso de los datos de textos de diferentes autores, hemos dividido las letras en dos grupos, y analizamos la submatriz de vocales (imagen 21.2). En este caso, las filas (textos) que no hemos dividido en grupos están centradas. Sin embargo, las columnas, que sí se han dividido, no lo están. Si quisiéramos proyectar la letra *Y* sobre el mapa del AC del grupo de las vocales, para no tener que centrar el perfil de la *Y*, lo que haríamos es utilizar las coordenadas centradas en cero ϕ_{ik} (es decir, los vértices de las filas). Es decir, la media ponderada usual proporciona las coordenadas principales (véase el cap. 12) y la fórmula específica de transición (14.2) aplicable a este caso (para una solución bidimensional) es:

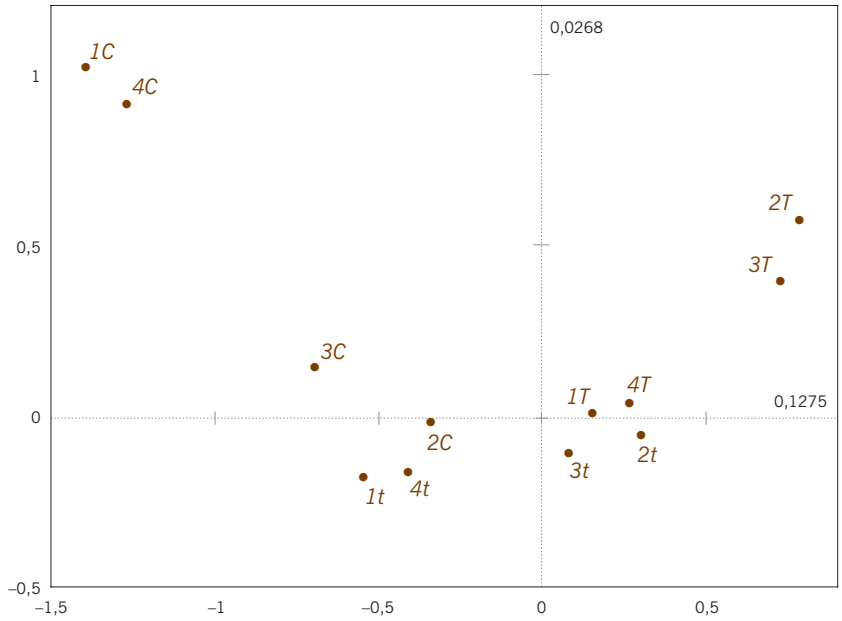


Imagen 21.4:
Mapa del AC de subgrupos de las respuestas categóricas sustantivas (excluidas las respuestas no sustantivas). Hemos ajustado la solución para hallar el mejor ajuste de las tablas de fuera de la diagonal, lo que lleva a una mejora considerable del ajuste total, explicándose el 84,9% de la inercia

$$\sum_i y_i \phi_{ik} \qquad k = 1,2 \tag{21.1}$$

donde y_i es el i-ésimo valor de perfil de Y. Por otra parte, si quisiéramos proyectar un nuevo texto con valores de perfil t_j en el subgrupo de interés (su suma es igual a la proporción de vocales en ese texto, no es 1), tendríamos que centrar los datos con relación a los valores originales c_j del centroide, antes de llevar a cabo el producto escalar con las coordenadas estándares de las columnas γ_{jk}:

$$\sum_j (t_j - c_j)\gamma_{jk} \qquad k = 1,2 \tag{21.2}$$

Fijémonos que para situar un punto adicional en el AC de subgrupos y en el AC habitual, se puede hacer siempre este tipo de centrado. Sin embargo, no es necesario cuando las coordenadas estándares cumplen que $\sum_i r_i \phi_{ik} = 0$ y $\sum_j c_j \gamma_{jk} = 0$, lo que ocurre cuando la suma se lleva a cabo con relación a todos los datos.

Puntos adicionales en el ACM de subgrupos

Los encuestados (filas) de la matriz binaria, así como cualquier agrupación de filas, por ejemplo, según el nivel de educación, género, etc., se pueden representar como puntos adicionales. Igual que en el ACM usual, representamos las categorías de las variables adicionales en los centroides de los puntos de los encuestados que se hallan en estos grupos. En el mapa de la imagen 21.5 mostramos las posiciones de varias categorías demográficas con relación a los mismos ejes principales de la imagen 12.4.

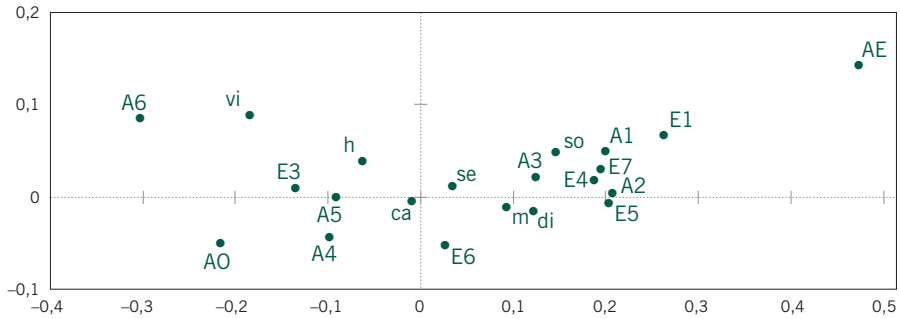


Imagen 21.5:
Posiciones de los puntos adicionales en el mapa de la imagen 21.4

Nota: El significado de las abreviaciones se puede consultar en el capítulo 17, página 161; AO y AE indican Alemania Occidental y del Este, respectivamente

RESUMEN: Análisis de correspondencias de subgrupos

1. La idea en el AC de *subgrupos* es visualizar una submatriz de filas o de columnas (o de ambas) en subespacios del espacio original que se obtiene con todos los datos. En este procedimiento, mantenemos el centroide original en el centro del mapa, también mantenemos las masas originales y los pesos de las distancias χ^2.
2. Dado que en el análisis de subgrupos se mantienen las propiedades del espacio original, podemos descomponer la inercia total original en partes que corresponden a las inercias de las distintas submatrices que componen la matriz original.
3. Podemos implementar fácilmente el AC de subgrupos dejando de calcular los valores marginales de las submatrices, de manera que en todos los cálculos habituales del AC utilizamos los valores marginales originales (masas).
4. Aplicar el ACM a determinadas submatrices de categorías nos proporciona una estrategia de análisis que puede ser muy útil en el análisis de datos procedentes de cuestionarios. Por ejemplo, podemos omitir los valores perdidos. Podemos centrarnos, para todas las respuestas, en un determinado tipo de categorías, y así visualizar las dimensiones de esta submatriz sin interferencias de las otras categorías.
5. Igual que en el ACM habitual, podemos aplicar el ACM de subgrupos a las matrices binaria y de Burt. A continuación podemos redimensionar los resultados para así optimizar el ajuste de las submatrices. Esto permite mejorar mucho los porcentajes de inercia explicados en el mapa.
6. Podemos añadir puntos adicionales en el mapa de una determinada submatriz. En el ACM de subgrupos esta posibilidad nos permite poder relacionar determinadas respuestas con categorías demográficas.