La práctica del análisis de correspondencias

MICHAEL GREENACRE

Catedrático de Estadística en la Universidad Pompeu Fabra

Separata del capítulo 17

Tablas concatenadas

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

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

www.fbbva.es

Capítulo 17

Tablas concatenadas

Habitualmente, la investigación en ciencias sociales, exige trabajar con multitud de variables. Es frecuente que en los cuestionarios de las encuestas encontremos respuestas a muchas preguntas, así como muchas características demográficas que, a menudo queremos relacionar con las opiniones de la gente. El AC tiene como ventaja su capacidad para visualizar simultáneamente muchas variables. Sin embargo, como mostramos en el capítulo anterior, debido al gran número de combinaciones de las categorías, existe un límite en el número de variables que podemos codificar interactivamente. Cuando nos encontramos en esta situación, un procedimiento alternativo consiste en codificar los datos en forma de tablas *concatenadas* o *compuestas*. Ello nos permite interpretar en un mapa conjunto las relaciones entre variables demográficas y variables de opinión. En este capítulo veremos ejemplos sobre esta aproximación, tanto para diversas características demográficas como para varias preguntas.

Contenido

Diversas variables demográficas, una pregunta

Vamos a ampliar el conjunto de datos del capítulo 16 relacionado con la opinión de la gente sobre el trabajo de las mujeres. Aparte de país (24 categorías, véanse las abreviaciones en la imagen 16.4), género (2 categorías, *h* y *m*) y grupo de edad (6 categorías, de A1 a A6), vamos a incluir el estado civil (5 catego-

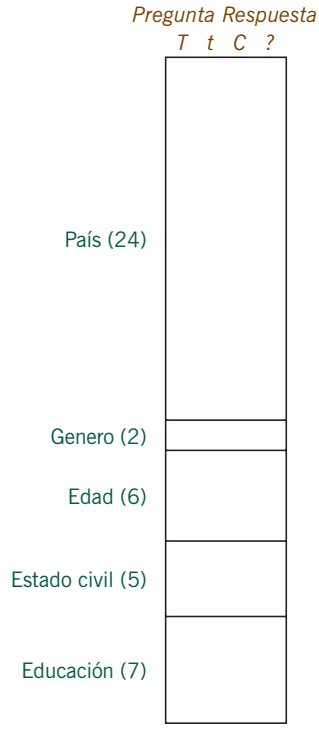


Imagen 17.1: *Tablas concatenadas de contingencia que hemos obtenido cruzando cinco variables demográficas por la respuesta a la pregunta sobre el trabajo de las mujeres (T = tiempo completo, t = tiempo parcial, C = permanecer en casa, ? = no sabe/no contesta)*

rías) y educación (7 categorías), tendremos en total cinco variables demográficas. Las definiciones y las abreviaciones de las dos variables adicionales son las siguientes:

— Estado civil: ca (casado), vi (viudo), di (divorciado), se (separado), so (soltero)

— Educación: E1 (sin educación formal), E2 (educación primaria incompleta), E3 (educación primaria), E4 (educación secundaria incompleta), E5 (educación secundaria), E6 (educación universitaria incompleta), E7 (educación universitaria)

Concatenación como alternativa al codificado interactivo

Codificar interactivamente las cinco variables es completamente imposible. ¡El número de combinaciones sería: 24 × 2 × 6 × 5 × 7 = 10080! Como alternativa, podemos cruzar cada variable demográfica con las respuestas a las preguntas y luego concatenar las tablas de contingencia resultantes, una encima de la otra, como mostramos en la imagen 17.1. La tabla de arriba es la correspondiente a la tabla de la imagen 16.4, con los países como filas. A continuación la tabla con los dos géneros como filas, luego las seis filas de los grupos de edad, y así sucesivamente. Este tipo de codificación no nos permitirá analizar interacciones, la podemos contemplar como una especie de AC medio de las cinco tablas individuales.

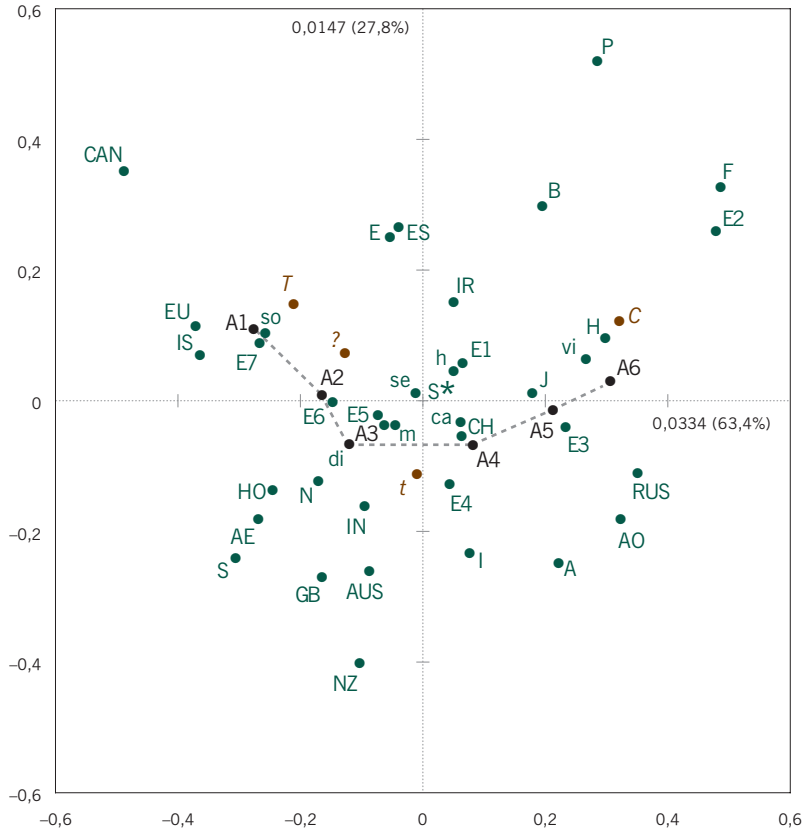


Imagen 17.2:
Mapa simétrico del AC correspondiente a la agrupación de las cinco tablas de contingencia que se muestran de forma esquemática en la imagen 17.1; inercia total = 0,05271, porcentaje de inercia del mapa: 91,2%

AC de tablas concatenadas

Aplicando el AC a la matriz de 44 × 4, correspondiente a las tablas concatenadas, obtenemos el mapa de la imagen 17.2. Las posiciones relativas de las cuatro respuestas categóricas, *T*, *t*, *C* y *?*, aparecen casi como en el mapa de la imagen 16.8. Comparando con los mapas de las imágenes 16.5 y 16.7 vemos que se ha producido una ligera rotación en las posiciones (en el epílogo trataremos sobre rotación). Cada categoría demográfica define un perfil que se sitúa en el mapa con relación a las cuatro respuestas categóricas. Son de especial interés las siguientes características del mapa:

- Podemos unir las categorías de las variables ordinales como, por ejemplo, edad. Vemos, como era de esperar, que edad sigue una trayectoria curva con relación a las respuestas *T*-*t*-*C*, de las más liberales a las más tradicionales.
- Educación sigue un esquema similar pero de derecha a izquierda, excepto para la categoría E1 (sin educación formal) que se halla cerca de la media.
- Las categorías correspondientes al estado civil, muestran a so (solteros) en el lado liberal, y a vi (viudos) en el lado tradicional, probablemente estas variables están correlacionadas con el grupo de edad.

- Los puntos correspondientes a género, h y m se hallan opuestos entre sí con relación a la media. En general, muestran las diferencias entre hombres y mujeres en los distintos países (vimos diferencias específicas en la imagen 16.7).
- De todas las variables demográficas, las diferencias entre países siguen siendo las más importantes.
- Países como España, Eslovenia, Irlanda y Bulgaria, que se hallan dentro del arco, son países polarizados con valores por encima de la media tanto para *T* (trabajo a tiempo completo) como para *C* (permanecer en casa).
- Vemos que el punto correspondiente a no respuesta *?* se halla en el lado liberal del mapa. Es decir, su perfil con relación a las variables demográficas es más similar a *T* que a *C* (en el capítulo 21 veremos que Canadá, tiene un alto porcentaje de no respuestas).

Limitaciones de la interpretación del análisis de tablas concatenadas

Es importante que nos demos cuenta de que el mapa de la imagen 17.2 únicamente muestra asociaciones entre variables demográficas y respuestas a las preguntas. No muestra relaciones entre las variables demográficas entre sí. Con las tablas concatenadas, no obtenemos información sobre la relación entre edad, educación y país. El hecho de que el grupo de edad más joven A1, el nivel de educación más alto E7, y países como Canadá, Estados Unidos e Israel se hallen todos en el lado izquierdo no significa que en estos países predominen jóvenes de nivel de educación más alto. Dado que las variables se relacionan de forma separada con las respuestas, la interpretación correcta es que los porcentajes de respuesta con relación a *T* (trabajo a tiempo completo) del grupo de edad más joven, el nivel de educación más elevado y los mencionados países se hallan por encima de la media. Para confirmar cualquier relación entre las variables demográficas entre sí, tendríamos que cruzarlas y luego analizar la tabla de contingencia resultante.

Descomposición de la inercia en tablas concatenadas

Un resultado muy útil para este capítulo y capítulos siguientes es que cuando cruzamos y agrupamos los mismos individuos, como podemos ver en la imagen 17.1, la inercia total del AC de la tabla concatenada es la media de las inercias de los AC de cada tabla. Lo podemos comprobar calculando las inercias de cada una de las cinco tablas de la imagen 17.1:

Tabla	*Inercia*
País	0,14558
Género	0,00452
Edad	0,04216
Estado civil	0,02675
Educación	0,04221
Media	*0,05224*

La inercia total del análisis de la tabla concatenada es de 0,05271, algo superior al valor medio que aparece en la tabla anterior. Ello es debido a que faltan datos de algunas variables demográficas. Los totales de las tablas varían de 30.471 para educación (a título indicativo señalemos que, en la muestra española, la respuesta a educación se codificaron como «no disponible» para todos los individuos de la muestra) a 33.590 (la muestra completa) para edad y país. El hecho de que los totales de las distintas tablas no sean iguales hace que aumente ligeramente la inercia total de la tabla compuesta, ya que existen pequeñas diferencias en los totales de las columnas de las distintas tablas. Para que la descomposición anterior sea exacta, el total de todas las tablas —y, en consecuencia, también los valores marginales— tienen que ser iguales. La tabla de inercias anterior también muestra que la mayor inercia se produce entre países. Por tanto, la relación entre las respuestas a las preguntas y los países debe dominar los resultados.

Concatenación de tablas por filas y por columnas

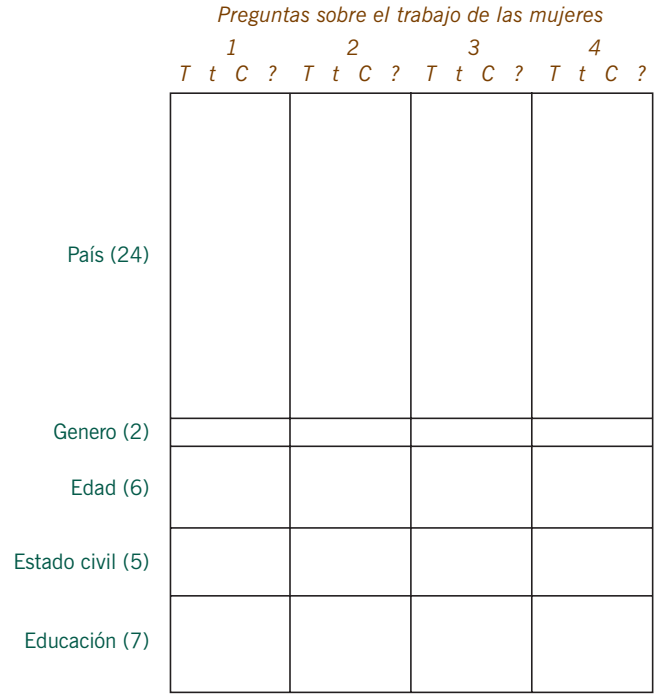
La concatenación de tablas se puede ampliar introduciendo nuevas preguntas que hayamos cruzado con los datos demográficos. En la encuesta ISSP, de la que hemos obtenido estos datos, había cuatro preguntas relacionadas con la opinión de los encuestados sobre el trabajo de las mujeres. Cada una de ellas con las mismas cuatro respuestas: trabajo a tiempo completo, trabajo a tiempo parcial, permanecer en casa y una categoría adicional que incluye todos los tipos de «no respuesta». En concreto, las preguntas hacían referencia a los siguientes supuestos sobre la situación de las mujeres: (1) casadas con hijos, (2) con un hijo en edad preescolar en casa, (3) con un hijo en edad escolar en casa (la pregunta que hemos estado analizando hasta ahora) y (4) con todos los hijos viviendo fuera de casa. Podemos cruzar cada una de las cinco variables demográficas con cada una de estas cuatro preguntas, con lo que obtenemos 20 tablas de contingencia, que asimismo podemos concatenar por filas y por columnas, como mostramos de forma esquemática en la imagen 17.3.

AC de tablas concatenadas por filas y por columnas

Aplicando el AC a las 20 tablas agrupadas en cinco filas y cuatro columnas obtenemos el mapa de la imagen 17.4. Hemos representado las categorías por puntos. Son de especial interés las siguientes características del mapa:

- Las 16 columnas forman una estructura en arco incluso más clara que la que vimos anteriormente, que se extiende desde *2T* y *3T*, arriba a la izquierda, hacia *3t*, *4t* y *2C*, en la parte baja, y luego sube hacia *4C* y *1C*, arriba a la derecha. Este es el resultado típico del AC cuando existe lo que los ecólogos llaman un *gradiente* en los datos. Aquí el gradiente se produce en la dispersión de opiniones, de liberales a tradicionales. A lo largo de este gradiente, podemos ordenar, de forma aproximada, las respuestas categóricas, de la siguiente manera (por el momento omitimos las no respuestas de la discusión *(?)*):

Imagen 17.3:
Tablas concatenadas de contingencia que se han obtenido cruzando cinco variables demográficas con las respuestas a las preguntas sobre el trabajo de las mujeres (T = tiempo completo, t = tiempo parcial, C = permanecer en casa, ? = no sabe/no contesta)



2T 3T 2t y 4T 1T 3t 4t 2C 1t 3C 4C 1C

lo que muestra cómo las columnas se alinean de las extremadamente liberales a la izquierda (las mujeres pueden trabajar a tiempo completo incluso con hijos en casa) a las extremadamente tradicionales a la derecha (las mujeres deben permanecer en casa incluso si los hijos no viven en ella).

- La mayor parte de las filas se sitúa a lo largo de esta curva. Sin embargo existe una sustancial dispersión en la segunda dimensión, que opone a países con una opinión polarizada en sus respuestas (parte superior del mapa, concretamente España) con países con la mayor parte de respuestas en las categorías intermedias del gradiente (parte inferior del mapa, como por ejemplo Austria y Alemania Occidental).
- Los cuatro puntos correspondientes a las no respuestas se hallan juntos formando un pequeño grupo, justo a la izquierda de la media —de hecho, estos puntos quedan mejor representados en la tercera dimensión de este análisis—. Es decir, hay que imaginarlos como si salieran verticalmente de la hoja, la tercera dimensión ordena los grupos demográficos con relación a los porcentajes de no respuesta en las cuatro preguntas.

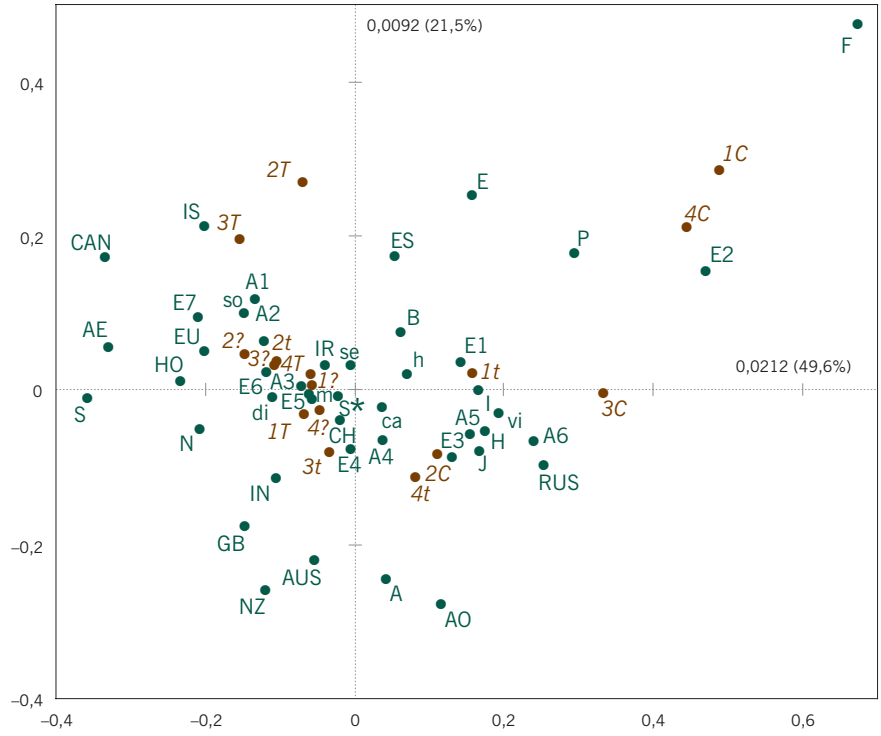


Imagen 17.4:
Mapa simétrico del AC correspondiente a la concatenación de las 20 tablas de contingencia que mostramos de forma esquemática en la imagen 17.3; inercia total = 0,0427; porcentaje de inercia del mapa = 71,1%

Descomposición de la inercia de la tabla concatenada

Aquí también podemos aplicar los resultados que vimos anteriormente sobre la descomposición de la inercia. En primer lugar, si todas las tablas de la tabla concatenada tuvieran exactamente el mismo número de individuos, la inercia de la tabla concatenada sería, exactamente, la media de las inercias de las distintas tablas de contingencia que la forman. Veámoslo de manera más formal. Supongamos que $\mathbf{N}_{qs}$, $q = 1,...,Q$, $s = 1,...,S$ son tablas de contingencia que cruzan, dos a dos, Q variables categóricas, con S variables categóricas, todas ellas con el mismo número n de individuos (en nuestro ejemplo, $Q = 5$ y $S = 4$). Sea $\mathbf{N}$ la tabla concatenada formada uniendo vertical y horizontalmente las Q × S tablas. Entonces:

$$\text{inercia}\,(\mathbf{N}) = \frac{1}{QS}\sum_{q=1}^{Q}\sum_{s=1}^{S}\text{inercia}\,(\mathbf{N}_{qs}) \qquad (17.1)$$

Este resultado es aproximado si hay pérdida de datos en alguna de las tablas de contingencia. En el ejemplo que nos ocupa hemos añadido los valores perdidos de las cuatro preguntas sobre el trabajo de las mujeres a las categorías *?* respectivas, que incluyen respuestas como, por ejemplo, «no sabe». Existen muy pocos valores perdidos correspondientes a variables demográficas de los diferentes países. La falta de algunos valores hará que el resultado de (17.1) no se cumpla exacta-

mente. En efecto, la inercia de la tabla concatenada **N** (lado derecho de (17.1)) aumentará en una pequeña cantidad ε, debido a diferencias en las frecuencias marginales, de manera que (17.1) se convierte en:

$$\text{inercia}(\mathbf{N}) = \frac{1}{QS}\sum_{q=1}^{Q}\sum_{s=1}^{S}\text{inercia}(\mathbf{N}_{qs}) + \varepsilon \qquad (17.2)$$

En la tabla de la imagen 17.5 tenemos los valores de las inercias de todas las tablas de contingencia, así como las medias de las filas, las columnas y la media total. Como era de esperar, la inercia total de la tabla concatenada es ligeramente superior (0,04273) a la media de las inercias de las tablas que la componen (0,04239), la diferencia es del 0,8%. En la tabla de la imagen 17.6 mostramos las inercias de la tabla de la imagen 17.5 expresadas en tantos por mil con relación al valor de 0,04273 × 20 (el valor a la izquierda de (17.2), multiplicado por $QS = 20$) para hacer más cómodo el análisis, igual que hicimos cuando interpretamos las contribuciones a la inercia en el capítulo 11. Estos resultados muestran que en promedio los países explican el 65,6% de la inercia del análisis de la tabla concatenada, seguida por educación (13,3%) y edad (11,6%). La pregunta 3 es la que, en general, presenta mayores inercias (30,6% de la inercia total) —es decir, para esta pregunta existen más diferencias entre los grupos demográficos—, mientras que, para la pregunta 4, las inercias son generalmente menores (21,6% de la inercia total). El total de esta tabla, 992, indica, que la contribución de ε a la inercia es del 0,8% [ecuación (17.2)]. Podemos explicar este valor por las pequeñas diferencias existentes entre las sumas marginales de las 20 tablas de contingencia, diferencias ocasionadas por la pérdida de datos.

Los mapas sólo muestran las asociaciones entre las respuestas y los grupos demográficos, no entre las respuestas entre sí

Una vez más hacemos hincapié en los límites de la interpretación de mapas como el de la imagen 17.4. Con relación a las cuatro preguntas, debemos recordar que no estamos analizando las asociaciones *entre* estas preguntas, sino que lo que analizamos son las asociaciones entre estas variables y las variables demográficas. El análisis de las asociaciones entre variables respuesta es el tema del próximo capítulo, que trata sobre análisis de correspondencias múltiples.

Imagen 17.5:
Inercias de las 20 tablas de contingencia que integran la tabla concatenada de la imagen 17.4; se dan los valores de las medias de las filas y de las columnas, así como el de la media global

VARIABLE	*Pregunta 1*	*Pregunta 2*	*Pregunta 3*	*Pregunta 4*	*Media*
País	0,15268	0,12834	0,14558	0,13410	*0,14018*
Género	0,00821	0,00336	0,00452	0,00484	*0,00523*
Edad	0,01033	0,03359	0,04216	0,01266	*0,02469*
Estado civil	0,00529	0,01341	0,02675	0,00869	*0,01354*
Educación	0,02306	0,02380	0,04221	0,02430	*0,02834*
Media	*0,03991*	*0,04050*	*0,05224*	*0,03692*	*0,04239*

Variable	*Pregunta 1*	*Pregunta 2*	*Pregunta 3*	*Pregunta 4*	*Total*
País	179	150	170	157	*656*
Género	10	4	5	6	*24*
Edad	12	39	49	15	*116*
Estado civil	6	16	31	10	*63*
Educación	27	28	49	28	*133*
Total	*234*	*237*	*306*	*216*	*992*

Imagen 17.6:
Contribuciones a la inercia, expresadas en tantos por mil, de las 20 tablas de contingencia que integran la tabla concatenada; el 0,8% restante es la inercia adicional debido a la diferencia en los valores marginales de las columnas ocasionadas por los valores perdidos

RESUMEN:
Tablas concatenadas

1. Una posibillidad para analizar la relación entre variables demográficas y las respuestas a varias preguntas es agrupar todas las tablas que cruzan los dos conjuntos de variables, y analizar la *tabla concatenada* resultante mediante AC.
2. Debemos interpretar el mapa del AC de la tabla compuesta teniendo presente que la información que analizamos deriva de un conjunto de relaciones entre las respuestas y cada una de las variables demográficas. En el mapa no existe información específica sobre las relaciones existentes entre las respuestas entre sí o entre las variables demográficas entre sí.
3. Debemos contemplar el mapa del AC de una tabla concatenada como un mapa promedio resultante del AC de cada una de las tablas individuales.
4. Cuando los valores marginales de filas y columnas de todas las tablas que integran la tabla concatenada son idénticos, la inercia total de la tabla concatenada es igual a la media de las inercias de cada una de las tablas (ello se cumple cuando todas las tablas tienen el mismo número de individuos). Sin embargo, cuando en algunas tablas faltan individuos, el resultado anterior es sólo aproximado, ya que la inercia total de la tabla concatenada será ligeramente mayor que la media de las inercias de las tablas que la integran.