Metodología de Encuestas
Volumen 10, 2008, 39-58
ISSN: 1575-7803

ANÁLISIS DE ENCUESTAS CON PREGUNTAS CONDICIONADAS

Amaya Zárraga **Beatriz Goitisolo**
Departamento de Economía Aplicada III,
Universidad del País Vasco-Euskal Herriko Unibertsitatea
Teléfono: 946013740
Fax: 946013754
email: amaya.zarraga@ehu.es

RESUMEN. El análisis de correspondencias múltiples (ACM) es frecuentemente utilizado para el análisis de encuestas. Esta técnica estudia la relación entre las variables cualitativas, recogidas en la encuesta sobre una misma población, cuando éstas han sido codificadas en forma lógica y representadas en lo que se conoce como tabla disyuntiva completa (TDC). Sin embargo, las encuestas pueden contener datos ausentes motivados por ejemplo por la no respuesta y, en casos muy frecuentes, por la existencia de preguntas condicionadas a las que no toda la población tiene que responder. En estos casos, la codificación de los datos en forma lógica se representaría mediante una tabla disyuntiva incompleta (TDI). Tradicionalmente, se crea por cada modalidad ausente una modalidad ficticia para conseguir una TDC y poder ser analizada mediante ACM. Sin embargo, la inclusión de estas modalidades puede alterar los resultados del análisis. En este trabajo se propone el análisis, con las transformaciones adecuadas, de la TDI.

PALABRAS CLAVE. Análisis de Correspondencias, Tablas Disyuntivas Incompletas

ABSTRACT. Multiple Correspondence Analysis (MCA) is frequently used for analysing surveys. This method studies the relationship between several categorical variables defined with respect to a certain population, the data being codified in a complete disjunctive table (CDT). However, the main sources of information include those surveys in which it is usual to find a certain number of absent data and conditioned questions that do not need to be answered by the whole population. In these cases, the data are codified in an incomplete disjunctive table (IDT). The creation for each absent data variable of a non-answer category would allow for a CDT, and, consequently, the analysis by MCA. However, the inclusion of these dummy modalities could alter the results. This paper proposes the appropriate transformations to analyze and IDT.

KEYWORDS. Correspondence Analysis, Incomplete Disjunctive Table

Recibido: 7 de febrero 2008
Revisado: 28 de abril 2008
Aceptado: 15 de mayo 2008

Introducción

El estudio factorial de la información procedente de encuestas es habitualmente llevado a cabo a través del análisis de correspondencias múltiples de tablas disyuntivas completas. En este análisis se impone a todos los individuos la obligación de pertenecer a alguna de las modalidades de cada cuestión. Es decir, se impone que todas las preguntas sean respondidas.

Sin embargo, es muy frecuente que las encuestas contengan preguntas condicionadas. En esta situación un individuo debe responder o no a una pregunta dependiendo de cuál haya sido su respuesta a una cuestión anterior. Por ejemplo, se le pregunta si sabe o no inglés; a continuación, y sólo si ha respondido saber inglés, se le pregunta su nivel de inglés en determinados aspectos. Pueden coexistir, además, varios grupos de preguntas condicionadas (los que saben inglés, francés, sólo los que tienen familiares y se relacionan con ellos contestarán la frecuencia con que lo hacen, etc). En este caso, la no respuesta se agrupa en un determinado número de cuestiones (aquellas cuya respuesta está condicionada por una pregunta anterior), caracteriza a determinados grupos de individuos y da lugar a una tabla disyuntiva incompleta.

En este trabajo se propondrá una metodología para el análisis de este tipo de tablas.

Notación

Se considera la tabla de datos que recoge en forma lógica y disyuntiva las respuestas de un conjunto de individuos a un conjunto de preguntas o cuestiones, poseyendo cada una de ellas un conjunto finito de modalidades de respuesta. Si los individuos responden a todas las preguntas, la tabla así obtenida es la TDC que denotaremos por Z. Se dirá que tal tabla de datos es disyuntiva incompleta, y se denotará por Z^*, cuando los individuos no dan respuesta a una o más de las cuestiones preguntadas.

	$q=1$		…	q	…	$q=Q$
	$j=1$ …	$j=J_1$		j		
1	1	0				
2	0	0				
3	0	1				
⋮						
i				z_{ij}		
⋮						
n						

donde:

Q = {1,…,q,…,Q} es el conjunto de variables a las cuales debe responder el individuo

J_q = {1,…,j,…,J_q} es el conjunto de modalidades de la variable q $\in Q$

J = {1,…,j,…,J} es el conjunto de modalidades de todas las variables

= $\bigcup_{q=1}^{Q} J_q$

I = {1,…,i,…,n} es el conjunto de individuos

z_{ij} = $\begin{cases} 1 & \text{si el individuo } i \in I \text{ responde a la modalidad } j \in J \\ 0 & \text{en otro caso} \end{cases}$

$z_{i.}$ = $\sum_{i \in I} z_{ij}$ es el número de cuestiones a las que responde el individuo $i \in I$

$z_{.j}$ = $\sum_{i \in I} z_{ij}$ es el número de individuos que eligen la modalidad $j \in J$

Se denotará $z_{.j}^{q}$ cuando interese dejar constancia de la variable $q \in Q$ a la que pertenece dicha modalidad

z = $\sum_{i \in I} \sum_{j \in J} z_{ij}$ es el total de la tabla

En las tablas disyuntivas incompletas -al igual que en las completas analizadas mediante el ACM clásico- las variables siguen estando definidas a través de un conjunto de modalidades a las que el individuo debe responder sobre su pertenencia ($z_{ij} = 1$) o no ($z_{ij} = 0$).

Pero, en ocasiones, esas modalidades correspondientes a una misma variable no están definidas en forma completa, es decir, el individuo puede no pertenecer a ninguna de ellas; en otras ocasiones a pesar de estar definidas en forma completa el individuo puede no revelar a que modalidad pertenece; en ambos casos:

$$z_{ij} = 0 \qquad \begin{matrix} i \in I \\ q \in Q \end{matrix} \qquad \forall j \in J_q$$

Por ello será necesario definir también una variable que toma el valor 1 si el individuo $i \in I$ responde a la cuestión $q \in Q$ y 0 en caso contrario:

$$z_{i.}^{q} = \sum_{j \in J_q} z_{ij} \qquad \forall q \in Q \qquad \forall i \in I$$

Y se denotará por z_q el número de individuos que han respondido a la cuestión $q \in Q$:

$$z_q = \sum_{j \in J_q} z_{.j} \qquad \forall q \in Q$$

En resumen, las siguientes relaciones que se dan en las TDC dejan de cumplirse en las TDI:

$$z_{i.}^{q} = 1 \qquad \forall q \in Q \qquad \forall i \in I$$

$$z_q = n \qquad \forall q \in Q$$

$$z_i = Q \qquad \forall i \in I$$

$$z = nQ$$

Las frecuencias relativas conjuntas y marginales se definen como es habitual:

$$f_{ij} = \frac{z_{ij}}{z} \qquad \forall i \in I \qquad \forall j \in J$$

$$f_{i.} = \frac{z_{i.}}{z} = \sum_{j \in J} f_{ij} \qquad \forall i \in I$$

$$f_{.j} = \frac{z_{.j}}{z} = \sum_{i \in I} f_{ij} \qquad \forall j \in J$$

así como los perfiles fila $i, i \in I$:

$$\frac{z_{ij}}{z_{i.}} \qquad \forall j \in J$$

y los perfiles columna $j,\ j \in J$:

$$\frac{z_{ij}}{z_{.j}} \qquad \forall i \in I$$

que forman las nubes $N(I) \in R^J$ y $N(J) \in R^n$ respectivamente.

Problemática que surge al aplicar ACM clásico a las TDI

En las TDI la marginal sobre I no es constante, a diferencia de las TDC en las que esta marginal es constante e igual a $1/n$.

La aplicación de la distancia χ^2 entre dos perfiles fila i e $i' \in I$ sería:

$$d^2(i,i') = \sum_{j\in J} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2 = \sum_{j\in J} \frac{z}{z_{.j}} \left(\frac{z_{ij}}{z_{i.}} - \frac{z_{i'j}}{z_{i'.}} \right)^2$$

Si los individuos i e i' no contestan al mismo número de preguntas, entonces $z_{i.}$ difiere de $z_{i'.}$ y la distancia χ^2 aumenta también con las respuestas comunes. Este es, por tanto, un concepto de distancia no deseable puesto que no reflejaría la similaridad entre individuos -en términos de modalidades comunes elegidas- buscada en un análisis de correspondencias.

La distancia χ^2 entre dos perfiles columna j y $j' \in J$ sería:

$$d^2(j,j') = \sum_{i\in I} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right)^2 = \sum_{i\in I} \frac{z}{z_{i.}} \left(\frac{z_{ij}}{z_{.j}} - \frac{z_{ij'}}{z_{.j'}} \right)^2$$

En esta distancia χ^2 cada individuo tendría una ponderación distinta dependiendo del número de respuestas elegidas. No parece lógico asignar menos importancia a aquellos individuos que responden a la totalidad de las preguntas frente a quienes no lo hacen.

Por tanto, la aplicación directa del análisis de correspondencias clásico no es adecuada al estudio de las tablas disyuntivas incompletas.

Metodología propuesta: análisis factorial con marginal modificada

Una vez descartada la aplicación del análisis clásico se propone un método de análisis de las TDI que minimice la influencia de la no respuesta sobre el análisis. Una solución evidente desde el punto de vista analítico sería eliminar aquellos individuos que no responden a todas las cuestiones, obteniendo de esta forma una TDC. Con esta alternativa perderíamos la información referente a esos individuos en el resto de las cuestiones; información que puede ser importante e implicar a un gran número de individuos en el caso de las preguntas condicionadas.

Una práctica habitual es crear para cada variable con datos ausentes una modalidad de no respuesta, obteniendo de esta forma una TDC a la que se puede aplicar el análisis clásico.

En los casos en los que la no respuesta se debe a la existencia de preguntas condicionadas, la inclusión de una modalidad de no respuesta en cada cuestión no sería adecuada, puesto que se estaría creando una serie de modalidades todas ellas con el mismo perfil e idéntico a una de las modalidades de la pregunta condicionante (no saben inglés) o a una combinación lineal de ellas ("no tienen familiares" y "tienen pero no se relacionan"). Esto podría perturbar los resultados hasta el punto de llegar a crear uno de los primeros ejes del análisis, como así ocurre en la aplicación a la Encuesta de Condiciones de Vida de 1989 de la Comunidad Autónoma de Euskadi presentada en Goitisolo y Zárraga .

En consecuencia la metodología que se propone, basada en el análisis de correspondencias con marginal modificada (Escofier, 1981) consiste en sustituir la marginal real de la tabla ($f_{i.} = z_{i.}/z$ que no es constante) por una marginal constante ($g_{i.} = 1/n$) en todo el análisis.

En las secciones que siguen se tratará con detalle lo adecuado de esta sustitución así como sus consecuencias sobre el análisis.

Nube de individuos: $N(I)$

El punto i se representa en R^J por el perfil $f_{ij}/g_{i.} = n\, z_{ij}/z$. Este perfil es diferente del perfil obtenido en el análisis de correspondencias múltiples clásico (z_{ij}/Q), al ser el efectivo total de la tabla (z) distinto de nQ .

La distancia cuadrática propuesta entre dos individuos i e i' es:

$$d^2(i,i') = \sum_{j\in J} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{g_{i.}} - \frac{f_{i'j}}{g_{i'.}} \right)^2 = \frac{n^2}{z} \sum_{j\in J} \frac{1}{z_{.j}} \left(z_{ij} - z_{i'j} \right)^2$$

Se comprueba que únicamente las respuestas diferentes hacen aumentar la distancia, sin tener en cuenta si ambos individuos responden al mismo número de cuestiones o no.

La ponderación de cada modalidad es, al igual que en correspondencias múltiples con datos completos, el inverso de su efectivo (la distancia aumenta en mayor proporción cuando la modalidad poseída por sólo uno de los individuos es rara).

Considerar la distancia anterior entre los puntos i e i' es equivalente a buscar la distancia euclídea habitual en un espacio dotado de métrica $1/f_{.j}$.

Cada punto i está dotado de un peso $g_i = 1/n$ que, a pesar de no venir representado por la marginal f_i del ACM clásico, coincide con el peso que se asigna a los individuos en ACM habitual. Este peso constante significa que todos los

individuos tienen la misma importancia, independientemente del número de cuestiones que han respondido. Es por tanto, más adecuado que $f_{i.}$.

La nube de individuos queda, por tanto, constituida por los perfiles fila $f_{ij}/g_{i.}$, con pesos $g_{i.}$ y métrica $1/f_{.j}$.

La coordenada j-ésima del centro de gravedad de la nube es:

$$G_I(j) = f_{.j} = \frac{z_{.j}}{z} \qquad \forall j \in J$$

que coincide con la correspondiente al centro de gravedad de la nube de individuos en ACM habitual.

Este será también el origen, O_I, de los ejes de máxima inercia que se han de buscar.

Nube de modalidades: *N*(*J*)

El punto $j \in J$ se representa en R^n por el perfil $f_{ij}/f_{.j} = z_{ij}/z_{.j}$, es decir, el mismo que en el ACM clásico.

La distancia cuadrática propuesta entre dos modalidades j y j' es:

$$d^2(j, j') = \sum_{i \in I} \frac{1}{g_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right)^2 = n \sum_{i \in I} \left(\frac{z_{ij}}{z_{.j}} - \frac{z_{ij'}}{z_{.j'}} \right)^2$$

Esta distancia es semejante a la utilizada cuando la tabla es completa y apropiada para el caso de preguntas condicionadas. Equivale a considerar la distancia euclídea en un espacio dotado de métrica $1/g_{i.}$.

Cada punto j está dotado de un peso $f_{.j} = z_{.j}/z$, coincide con el asignado en ACM de tablas completas y supone que cada modalidad tiene una importancia proporcional a la población que representa. Las modalidades tienen un peso en la construcción de los ejes tanto menor cuanto menor sea su efectivo.

La coordenada i-ésima del centro de gravedad de la nube es:

$$G_J(i) = f_{i.} = \frac{z_{i.}}{z} \qquad \forall i \in I$$

Este baricentro coincide en expresión, pero no en coordenadas, con el obtenido en ACM cuando se posee una TDC, puesto que si existe algún dato ausente, esta marginal no es constante:

$$f_{i.} = \frac{z_{i.}}{z} \neq \frac{1}{n}$$

Cuando la tabla es disyuntiva incompleta, la coordenada i-ésima del centro de gravedad de cada cuestión es:

$$G_q(i) = \frac{z_{i.}^q}{z_q} \qquad \forall i \in I \qquad \forall q \in Q \qquad \text{y}$$

donde $z_{i.}^q$ y z_q son las variables definidas en la sección 2.

Si no todos los individuos contestan a esa pregunta entonces:

$$G_q(i) \neq \frac{1}{n} \qquad \forall i \in I$$

puesto que $z_{i.}^q$ es diferente de 1 para aquellos individuos que no responden a la cuestión y z_q difiere de n. Este centro de gravedad dependerá del número de individuos que hayan respondido a la cuestión y de quiénes sean esos individuos. En consecuencia, no es el mismo para todas las variables, ni coincide con el de la nube de todas las modalidades como ocurría en correspondencias de una TDC. En el caso de las preguntas condicionadas sí será el mismo para todas aquellas variables que dependan de la misma pregunta condicionante (y en las que por tanto, coincidan todos los individuos que han de responderlas).

Si todos los individuos contestan a esa cuestión entonces:

$$G_q(i) = \frac{1}{n} \qquad \forall i \in I$$

puesto que $z_{i.}^q$ es 1 para todos los individuos y z_q es n. Pero este punto sólo coincide con el centro de gravedad de la nube si todos los individuos responden a todas las cuestiones (es decir, cuando se tiene una tabla completa), en caso contrario es un punto que representa la marginal modificada $g_{i.}$.

Como en todo análisis factorial el objetivo consiste en buscar el subespacio de R^n sobre el que se maximice la inercia proyectada de la nube y, evidentemente, interesa que la inercia a proyectar esté basada en las modalidades en las que existe más diferencia entre los individuos. Es decir, si una modalidad ha sido elegida por todos los individuos, no interesa que dicha modalidad incremente la inercia por el hecho de que exista un individuo que no ha respondido a otra cuestión. Eso es lo que ocurre si se toma como origen G_J (de i–ésima coordenada $f_{i.}$), mientras que si se traslada la nube, tomando como origen O_J (cuya coordenada i–ésima es $g_{i.}$), las modalidades elegidas por todos los individuos no contribuirán a aumentar la inercia. Por ello, el análisis de la nube de modalidades se hará tomando como origen O_J, es decir, no tomando como origen el centro de gravedad de todas las variables sino el

de aquéllas que son elegidas por todos los individuos. De esta forma la influencia de la no respuesta es minimizada.

Inercias de ambas nubes

La distancia de los puntos de ambas nubes a su origen correspondiente se recoge en las siguientes expresiones:

$$d^2(i, O_I) = \frac{n^2}{z} \sum_{j \in J} \frac{z_{ij}}{z_{.j}} + 1 - 2\frac{nz_{i.}}{z}$$

$$d^2(j, O_J) = \frac{n}{z_{.j}} - 1$$

A partir de estas distancias se pueden obtener las inercias de los perfiles fila y columna:

$$\text{Inercia de } (i) = g_{i.} d^2(i, O_I) =$$

$$= \frac{n}{z} \sum_{j \in J} \frac{z_{ij}}{z_{.j}} + \frac{1}{n} - 2\frac{z_{i.}}{z}$$

$$\text{Inercia de } (j) = f_{.j} d^2(j, O_J) =$$

$$= \frac{1}{z}\left(n - z_{.j}\right)$$

La inercia de una modalidad es mayor cuanto menor sea el número de individuos que la poseen. Es nula si todos los individuos la eligen y es máxima (n/z) si ningún individuo declara pertenecer a ella. Por tanto, al igual que en correspondencias múltiples sin datos ausentes, las modalidades raras pueden perturbar los resultados.

La inercia de una variable:

$$\text{Inercia de } (q) = \frac{1}{z}\left(nJ_q - z_q\right)$$

aumenta al disminuir el número de individuos que responden (z_q), y al incrementar el número de modalidades en que es clasificada (J_q). No es, por tanto, recomendable que este número de modalidades sea elevado (lo mismo que ocurre en el análisis clásico). En los cuestionarios con preguntas condicionadas habrá que tener especial cuidado con las cuestiones que no han de responder todos los individuos.

La suma para todos los puntos de una nube permite calcular su inercia total:

$$\text{Inercia total} = \frac{nJ}{z} - 1$$

Se puede comprobar que las inercias totales de ambas nubes son iguales y son función del número de respuestas ausentes (a través de z) pero no de su distribución en la tabla.

Obtención de los factores de ambas nubes

En el ACM clásico el análisis se puede hacer bien a través de las nubes centradas o bien a través de las nubes no centradas eliminando el primer eje (asociado a un valor propio igual a 1), que corresponde a la dirección de unión entre el origen del espacio y el baricentro de las nubes.

En el análisis con la marginal modificada esta relación no se mantiene y se deberá considerar siempre la nube trasladada al origen.

Obtener la sucesión de ejes ($u_s, s \in S = \{1, \cdots, s, \cdots, S\},\ S \le J$) que maximizan la inercia proyectada de la nube $N(I)$ equivale a:

$$\text{Maximizar:} \quad u_s^T M X^T P X M u_s$$

$$\text{Sujeto a:} \quad u_s^T M u_s = 1$$

$$u_s^T M u_t = 0 \qquad \forall t < s$$

donde:

- X es una matriz ($n \times J$) de término general:

$$x_{ij} = \frac{f_{ij}}{g_{i.} f_{.j}} - 1 \qquad \forall i \in I \quad j \in J$$

- Al igual que en análisis de correspondencias clásico cada elemento de esta matriz contiene las desviaciones entre la tabla de datos f_{ij} y una tabla de término general que corresponde a la hipótesis de independencia. La diferencia con el análisis clásico radica en que la frecuencia relativa marginal correspondiente a las filas es impuesta en función del número de filas en lugar de obtenida a partir de los datos.

- M es una matriz diagonal correspondiente a la métrica del espacio:

$$m_j = f_{.j} \qquad j \in J$$

- P es la matriz (también diagonal) de pesos:

$$p_i = g_{i.} \qquad i \in I$$

Se puede demostrar (Escofier y Pagès, 1992) que la nube definida por las filas de la matriz X, con la métrica y los pesos considerados, es isomorfa de la definida en la sección 4.1. Ambas nubes mantienen las mismas distancias entre dos puntos cualesquiera.

La resolución de este problema lleva a la diagonalización de la matriz $X^T PXM$ de orden ($J \times J$) cuyo término general es:

$$a_{jj'} = n\sum_{i\in I}\frac{f_{ij}f_{ij'}}{f_{.j}} - f_{.j'} \qquad j, j' \in J \tag{1}$$

Las proyecciones de la nube de individuos sobre los ejes de máxima inercia resultantes son:

$$F_s = XMu_s \qquad s \in S$$

Su i-ésima coordenada adopta la expresión:

$$\begin{aligned} F_s(i) &= \sum_{j\in J}\left(\frac{nf_{ij}}{f_{.j}} - 1\right) f_{.j}u_{sj} = \\ &= n\sum_{j\in J} f_{ij}u_{sj} - \sum_{j\in J} f_{.j}u_{sj} \qquad \begin{matrix} i \in I, \\ s \in S \end{matrix} \end{aligned}$$

Al diagonalizar la matriz $XMX^T P$ cuyo término general es:

$$d_{ii'} = n\sum_{j\in J}\frac{f_{ij}f_{i'j}}{f_{.j}} - f_{i'.} - f_{i.} + \frac{1}{n} \qquad i, i' \in J$$

se obtienen los ejes $v_s, s \in S$ [1] que maximizan la inercia proyectada de la nube $N(J)$ y tras premultiplicar dicha matriz por $X^T P$ las proyecciones $G_s, s \in S$ de dicha nube cuya j-ésima coordenada puede expresarse:

$$\begin{aligned} G_s(j) &= \sum_{i\in I}\left(\frac{nf_{ij}}{f_{.j}} - 1\right)\frac{1}{n}v_{si} = \\ &= \sum_{i\in I}\frac{f_{ij}}{f_{.j}}v_{si} - \frac{1}{n}\sum_{i\in I} v_{si} \qquad \begin{matrix} j \in J, \\ s \in S \end{matrix} \end{aligned}$$

Relaciones entre los factores

Los factores de ambas nubes se relacionan a través de las expresiones:

$$F_s = \frac{1}{\sqrt{\lambda_s}}XMG_s \qquad s \in S \tag{2}$$

[1] Las relaciones de dualidad entre ambos espacios, que se verifican en todo análisis de correspondencias, permiten establecer que los subespacios de ajuste, asociados a valores propios no nulos, son de idéntica dimensión.

$$G_s = \frac{1}{\sqrt{\lambda_s}} X^T P F_s \qquad s \in S \tag{3}$$

En el análisis de correspondencias con marginal modificada, igual que ocurre en el análisis clásico para la cantidad $f_{i.}$, los factores F_s están centrados para la cantidad $g_{i.}$:

$$\sum_{i \in I} g_i F_s(i) = 0 \qquad s \in S$$

Aplicando la fórmula de transición (3):

$$\begin{aligned} G_s(j) &= \frac{1}{\sqrt{\lambda_s}} \sum_{i \in I} \left(\frac{f_{ij}}{f_{.j} g_{i.}} - 1 \right) g_{i.} F_s(i) = \\ &= \frac{1}{\sqrt{\lambda_s}} \sum_{i \in I} \frac{f_{ij}}{f_{.j}} F_s(i) \qquad \begin{matrix} j \in J, \\ s \in S \end{matrix} \end{aligned}$$

que coincide con la relación baricéntrica del análisis de correspondencias.

Sin embargo, a diferencia del análisis de correspondencias clásico los factores G_s no están centrados por la cantidad $f_{.j}$ porque el análisis se hace tomando como origen un punto diferente al centro de gravedad.

$$\sum_{j \in J} f_{i.} G_s(j) = \sum_{i \in I} v_{si} f_{i.} - \sum_{i \in I} g_{i.} v_{si} \qquad s \in S$$

Si la marginal impuesta difiere de la marginal propia de la tabla esta cantidad es distinta de cero.

Por ello los factores F_s no pueden interpretarse como el baricentro de los G_s como en análisis clásico. Según la fórmula de transición (2):

$$\begin{aligned} F_s(i) &= \frac{1}{\sqrt{\lambda_s}} \sum_{j \in J} \left(\frac{f_{ij}}{f_{.j} g_{i.}} - 1 \right) f_{.j} G_s(j) = \\ &= \frac{1}{\sqrt{\lambda_s}} \left\{ \sum_{j \in J} \frac{f_{ij}}{g_{i.}} G_s(j) - \sum_{j \in J} f_{.j} G_s(j) \right\} \qquad \begin{matrix} i \in I, \\ s \in S \end{matrix} \end{aligned}$$

El segundo sumatorio corresponde a la proyección del centro de gravedad de $N(J)$, que al no haber sido tomado como origen de los ejes es diferente de 0.

Benali y Escofier afirman que "Este término, en la práctica es casi nulo, lo que permite interpretar como en correspondencias múltiples clásico la abscisa de un individuo como el baricentro de las modalidades que ha elegido". Hacen referencia a TDI en las que el efectivo de datos ausentes representa una proporción reducida en relación al total (caso de datos ausentes por olvido distribuidos de forma aleatoria a lo largo de la tabla). Sin embargo, puede alterar los resultados y su interpretación

cuando se considera nulo en una tabla de datos en la cual la proporción de no respuesta es grande o corresponde a ciertos grupos de individuos como es el caso de tablas de datos con preguntas condicionadas.

Lo cierto es que este segundo sumatorio es el mismo para todos los individuos, aunque difiere para los distintos ejes, por lo que se podría trasladar los factores $F_s(i)$ de tal forma que en la representación superpuesta de ambas nubes un individuo siga estando representado en el baricentro de las modalidades que posee:

$$F_s^*(i) = \frac{1}{\sqrt{\lambda_s}} n \sum_{j \in J} f_{ij} G_s(j) \qquad \begin{array}{l} i \in I, \\ s \in S \end{array}$$

Número de ejes

En el análisis de correspondencias de una TDC el número de ejes S es igual al número de modalidades activas menos el número de variables, porque todas las modalidades correspondientes a cada una de las variables se encuentran restringidas al mismo hiperplano. En el análisis de correspondencias con marginal modificada de una TDI, las modalidades de una misma cuestión no cumplen ningún tipo de restricción por lo que pueden existir tantos ejes como número de modalidades activas exista en el análisis. Si existen cuestiones con datos completos, sus modalidades mantendrán la misma restricción que en el análisis clásico por lo que la cantidad de ejes disminuirá en ese número de cuestiones completas.

La existencia de preguntas condicionadas en el análisis también reduce la cantidad de ejes, puesto que los individuos que han de responder a una pregunta condicionada vienen determinados por la respuesta a una modalidad (o combinación de ellas) anterior.

Tasas de inercia

En el ACM, bien se realice a través de la TDC o bien a partir de la tabla de Burt, se estudian las relaciones entre cada par de cuestiones y se revela el escaso interés de los valores propios y, en consecuencia, de las tasas de inercia como medida de la información explicada por cada uno de los factores.

Benzécri propone por ello la siguiente corrección de los valores propios:

$$\lambda_s^* = \left(\frac{Q}{Q-1}\right)^2 \left(\lambda_s - \frac{1}{Q}\right)^2 \qquad s \in S^*$$

para aquellos valores λ_s superiores a $1/Q$, siendo λ_s los valores propios resultantes del análisis de la tabla disyuntiva completa. Al estar los valores propios λ_s comprendidos entre 0 y 1, el término $(Q/(Q-1))^2$ permite obtener unos valores propios corregidos comprendidos también entre 0 y 1, que hacen posible su

comparación con otros análisis. Esta modificación de los valores propios lleva a definir las tasas de inercia proyectada:

$$\tau_s^* = \frac{\lambda_s^*}{\sum_s \lambda_s^*} \qquad s \in S^*$$

Esta corrección encuentra su justificación, para el caso de más de dos cuestiones, en la equivalencia entre los análisis de la TDC y de la tabla de Burt. En éste último análisis se introducen en la diagonal principal tablas que cruzan una cuestión consigo misma haciendo incrementar la inercia total y donde el resto de las tablas de contingencia aparecen dos veces. Por esta razón Greenacre (1993, 2006) propone la siguiente corrección de las tasas de inercia:

$$\tau_s^* = \frac{\lambda_s^*}{\frac{Q}{Q-1}\left(\text{Inercia (B)} - \frac{J-Q}{Q^2}\right)} \qquad s \in S^*$$

donde el denominador representa la inercia media de la tabla de Burt eliminando la parte de inercia motivada por los bloques diagonales de la tabla.

Una razón alternativa para calcular las tasas de inercia corregidas se encuentra en el estudio del caso particular donde las Q cuestiones son independientes dos a dos. A pesar del nulo interés del análisis de correspondencias clásico cuando se conoce (en ocasiones únicamente tras los resultados del análisis) la independencia de las cuestiones, su aplicación proporciona una inercia total no nula y $(J-Q)$ factores con valores propios asociados iguales a $1/Q$ (Zárraga, 1989), demostrando que los valores propios del análisis (exista o no independencia) recogen una inercia trivial debida a un efecto de estructura o de construcción de la tabla disyuntiva completa.

Cuando existen datos ausentes, el análisis del caso en el cual las cuestiones son independientes (Zárraga y Goitisolo, 1999), en la forma definida, revela que las tasas de inercia calculadas como los valores propios entre la inercia total, no son una buena medida de la asociación entre las cuestiones recogida por cada eje, por ello se propone, Autor: beatrizen el caso general en que las cuestiones no son independientes, calcular los valores propios y las tasas de inercia de la siguiente forma:

$$\lambda_s^* = \left(\frac{z}{n(Q-1)}\right)^2 \left(\lambda_s - \frac{n}{z}\right)^2 \qquad \forall\, \lambda_s > \frac{n}{z}, \quad s = 1, \cdots, J$$

$$\tau_s^* = \frac{\lambda_s^*}{\sum_s \lambda_s^*} \qquad s = 1, \cdots, J$$

donde λ_s es el valor propio obtenido en el análisis con marginal modificada de la TDI. Estos nuevos valores propios y tasas de inercia coinciden, si la tabla es disyuntiva completa, en la que $z = nQ$, con los propuestos por Benzécri (1979).

Evidentemente también se podría utilizar la corrección propuesta por Greenacre (1993, 2006) sin más que expresar λ_s^* como porcentaje de la inercia media de la pseudo-tabla de Burt (Zárraga y Goitisolo, 2000) una vez eliminada la inercia de los bloques diagonales de la misma.

Ayudas a la interpretación

Como en todo análisis factorial la metodología propuesta permite el cálculo de la contribución de los puntos a la formación de los ejes y de su calidad de representación sobre los mismos, medida a través de las contribuciones absolutas y relativas:

$$\mathrm{CTA}_s(i) = \frac{\frac{1}{n} F_s^2(i)}{\frac{1}{n}\sum_{i\in I} F_s^2(i)} \qquad \mathrm{CTR}_s(i) = \frac{F_s^2(i)}{d^2(i, O_I)}$$

$$\mathrm{CTA}_s(j) = \frac{f_{.j} G_s^2(j)}{\sum_{j\in J} f_{.j} G_s^2(j)} \qquad \mathrm{CTR}_s(j) = \frac{G_s^2(i)}{d^2(j, O_J)}$$

Ejemplo ilustrativo

En esta sección se presenta un breve ejemplo para ilustrar el efecto de analizar encuestas con preguntas condicionadas mediante el análisis de correspondencias clásico de la TDC y con la metodología propuesta en la sección 4. Ambos análisis han sido realizados con Splus.

Los datos analizados provienen de la Encuesta sobre Equipamiento y Uso de Tecnologías de Información y Comunicación en los Hogares (TIC-H), realizada en 2007 por el Instituto Nacional de Estadística (www.ine.es). De esta encuesta se ha elegido para el ejemplo el bloque IV correspondiente a "Uso de ordenador e Internet por los niños (10 a 15 años)" en los hogares encuestados. Por simplicidad se ha retenido la respuesta del informante de la encuesta sobre el primer niño del hogar. Se dispone por tanto de información relativa a 3204 niños. El bloque de uso de ordenador e Internet contiene 19 preguntas (entre paréntesis, la etiqueta identificativa en los planos factoriales) relativas a:

- uso de ordenador en los tres últimos meses (Ordenador).

- finalidad del uso de ordenador: ocio, música, juegos, etc. (OOcio), trabajos escolares (OEscolar) y otros usos (OOtros).
- lugar desde el que se usa el ordenador: su vivienda (OLVivienda), vivienda de familiares o amigos (OLAmigos), centro de estudios (OLEstudios), centros públicos (OLPublico), cibercafés o similares (OLCibercafes).
- uso de Internet en los tres últimos meses (Internet).
- finalidad del uso de Internet: ocio, música, juegos, etc. (IOcio), trabajos escolares (IEscolar) y otros usos (IOtros).
- lugar desde el que ha usado Internet: su vivienda (ILVivienda), vivienda de familiares o amigos (ILAmigos), centro de estudios (ILEstudios), centros públicos (ILPublico), cibercafés o similares (ILCibercafes).
- disponibilidad de teléfono móvil (Movil).

Las preguntas anteriores tienen como modalidades de respuesta sí y no. Entre ellas hay dos que condicionan la respuesta de cuestiones subsiguientes. En concreto, si a la pregunta ¿Ha utilizado el ordenador en los 3 últimos meses? el informante responde no, no debe de responder a las siguientes preguntas sobre el uso del ordenador. De modo similar, si la respuesta del informante a la pregunta ¿Ha utilizado Internet desde cualquier lugar en los últimos 3 meses? es no, no tiene que responder a las restantes cuestiones sobre el uso de Internet.

La codificación de las respuestas a estas 19 preguntas en forma disyuntiva, como se indica en la sección 2, da lugar a una tabla disyuntiva incompleta.

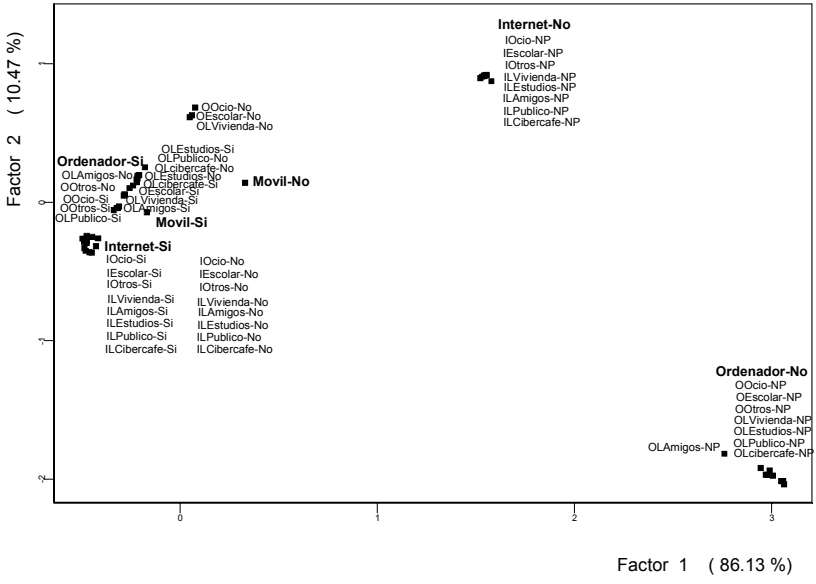
Para analizar esta tabla mediante el análisis de correspondencias múltiples clásico, es preciso crear, para cada una de las cuestiones a las que el informante de la encuesta no debe de responder, una modalidad de "no tiene que responder" o "no procede" (NP). Así se crea y analiza una tabla disyuntiva completa.

En este caso se crean 16 modalidades de este tipo, 8 correspondientes a las preguntas sobre el uso del ordenador, prácticamente todas ellas con el mismo perfil e igual al perfil de la modalidad que condiciona la respuesta – no ha usado el ordenador en los últimos tres meses, (Ordenador-No)-, y 8 correspondientes a las preguntas sobre la conexión a Internet, con idéntico perfil e igual al de la modalidad que condiciona la respuesta –no ha utilizado Internet en los últimos tres meses, (Internet-No)-.

Estas modalidades ficticias crean inercia en la nube de puntos y participan como si se tratara de modalidades activas en la determinación de los ejes factoriales.

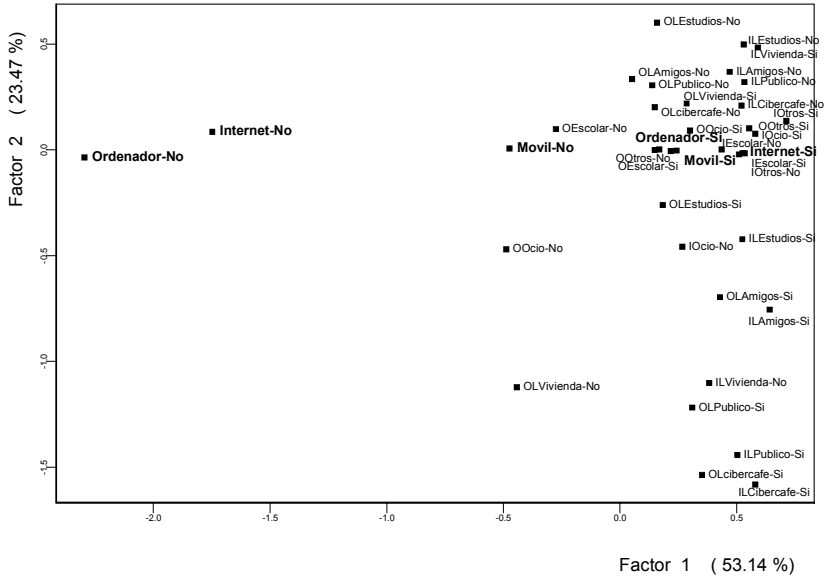
Consecuencia de ello es, como se observa en el primer plano factorial del análisis de la TDC (gráfico 1), que los dos primeros ejes factoriales están determinados por estas modalidades. La inercia proyectada sobre los dos primeros ejes (aplicada la corrección propuesta por Benzécri indicada en la sección 4.5) representa el 86,13% y el 10,47% de la inercia total, respectivamente y las 16 modalidades contribuyen en un 74,37% y 73,18% a la creación de los dos primeros ejes, respectivamente. La contribución de las restantes modalidades es escasa y, en consecuencia, no permiten alcanzar el objetivo del análisis, describir el uso de ordenador e Internet por los niños.

Gráfico 1: Análisis de la TDC



El análisis de la tabla disyuntiva incompleta, con la metodología propuesta, da solución al problema anterior. Los ejes factoriales están creados por las modalidades de respuesta efectiva de los individuos. Así, se observa en el primer plano factorial, gráfico 2, cómo el primer eje, que recoge el 53,14% de la inercia total, representa "el uso de ordenador e Internet" y opone a los individuos que han utilizado estas tecnologías en los últimos tres meses, con diferentes finalidades y en diversos lugares frente a quienes no lo han hecho. El segundo eje factorial, con una tasa de inercia de 23,47%, permite matizar "el lugar de uso del ordenador e Internet". Así se observa que en el primer cuadrante del plano están representados los individuos que usan el ordenador y se conectan a Internet desde la vivienda propia frente a quienes los hacen desde los otros lugares contemplados en la encuesta, que se proyectan en el cuarto cuadrante.

Gráfico 2: Análisis de la TDI



En consecuencia, el análisis factorial de encuestas que incluyen preguntas condicionadas no debe realizarse introduciendo en la tabla de datos modalidades ficticias de relleno que pueden distorsionar los resultados, como se ve en este ejemplo, sino tratando las respuestas dadas por los individuos con una metodología apropiada como es el análisis de la tabla disyuntiva incompleta con la marginal modificada.

Conclusiones

La presencia de preguntas condicionadas en las encuestas conlleva que la tabla en la que se codifican en forma lógica y disyuntiva las respuestas de los individuos sea una tabla incompleta. Esto es debido, evidentemente, a que las preguntas condicionadas por una pregunta anterior no tienen que ser respondidas por todos los individuos. La inclusión de una modalidad ficticia que represente "el no tener que responder" da lugar a una tabla completa pero el análisis puede resultar distorsionado puesto que pueden aparecer relaciones entre variables debidas únicamente al grupo de individuos que no ha tenido que responder. Parece, por tanto, adecuado trabajar con la tabla disyuntiva incompleta. Ahora bien, el análisis de correspondencias de esta tabla revela una deficiencia en el cálculo de la distancia χ^2 entre los individuos puesto que en este caso la distancia aumenta no sólo con las respuestas diferentes de los individuos sino también con aquellas modalidades

elegidas por los dos individuos si el número de preguntas respondidas por ambos no coincide.

La metodología propuesta permite subsanar esta deficiencia ya que la nueva distancia χ^2 únicamente va a aumentar con las respuestas no comunes de los individuos. Además la metodología permite buscar una relación entre las cuestiones al margen de la ausencia de respuesta.

Agradecimientos: *Este trabajo ha sido financiado por el Grupo de la UPV/EHU: IT-321-07*

Referencias

Benali, H. Escofier, B. (1987). Stabilité de l' analyse factorielle des correspondances multiples en cas de données manquantes et de modalités à faibles effectifs. *Revue de Statistique Appliquée* **XXXV**(1), 41-51.

Benzécri, J. (1979). Sur le calcul des taux d' inertie dans l' analyse d' un questionnaire, addendum et erratum à. *Les Cahiers de l' Analyse des Données* **IV**(3), 377-378.

Escofier, B. (1981). Traitement des questionnaires avec non réponse, analyse des correspondances avec marge modifiée et analyse multicanonique avec contrainte. *INRIA* **82**.

Escofier, B. Pagès, J. (1992). *Análisis factoriales simples y múltiples. Objetivos, métodos e interpretación*. Servicio Editorial Universidad del País Vasco.

Goitisolo, B. Zárraga, A. (1998). Application of the incomplete disjunctive tables study to the C.A.V. living conditions survey. In *Analyses Multidimensionnelles des donnèes.* Fernandez-Aguirre, K. and Morineau, A.(Eds). Cisia-Ceresta, pp. 301-313.

Greenacre, M. (1993). *Correspondence Analysis in Practice*. Academic Press.

Greenacre, M. (2006). From simple to multiple correspondence analysis. In *Multiple Correspondence Analysis and Related Methods*. M. Greenacre, J. Blasius (Eds.). Chapman & Hall/CRC, Boca Raton, Fl, pp. 327-350.

Zárraga, A. (1989). *Análisis de correspondencias múltiples por bandas. Aplicación al estudio de una gran encuesta. Tesis Doctoral*, Universidad del País Vasco.

Zárraga, A. Goitisolo, B. (1999). Independencia entre las cuestiones en el análisis factorial de tablas disyuntivas incompletas con preguntas condicionadas. *Qüestiió* **23**(3), 465-488.

Zárraga, A. Goitisolo, B. (2000). Estudio comparativo de análisis alternativos de tablas disyuntivas incompletas. *Documentos de trabajo. Biltoki* (8), 1-30.