# La práctica del análisis de correspondencias

**MICHAEL GREENACRE**

Catedrático de Estadística en la Universidad Pompeu Fabra

Separata del capítulo 23

## Recodificación de datos

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

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



www.fbbva.es

CAPÍTULO 23

# Recodificación de datos

Hasta ahora, en los 22 capítulos anteriores hemos trabajado con tablas de frecuencias; con tablas simples (capítulos 1 a 16 y 22) o con tablas compuestas (capítulos 17 a 21). En este capítulo vamos a tratar con datos de naturaleza distinta. Veremos cómo recodificarlos, o transformarlos, para que podamos visualizarlos utilizando AC. Este procedimiento fue muy bien desarrollado por Benzécri que, antes de visualizar los datos utilizando AC, utilizaba distintos procedimientos para transformarlos. Los datos que veremos en este capítulo derivan de escalas de grados, de preferencias, de comparaciones por pares o de escalas continuas. En todos estos casos deberemos recordar el paradigma fundamental del AC: el AC analiza recuentos. Por tanto, si somos capaces de transformar los datos a algún tipo de recuento, entonces es probable que sea apropiado aplicar AC. Para comprobar que la transformación es apropiada, tendremos que confirmar que tienen sentido los conceptos de perfil, de masa y de distancia $\chi^2$.

## Contenido



### Escalas de grados

En el capítulo 20 vimos una escala de grados típica, una escala de cinco puntos de acuerdo/desacuerdo que utilizamos en el ejemplo sobre ciencia y medio ambiente:

| □ | □ | □ | □ | □ |
|---|---|---|---|---|
| *Muy de acuerdo* | *Bastante de acuerdo* | *Ni de acuerdo ni en desacuerdo* | *Algo en desacuerdo* | *Muy en desacuerdo* |

Analizamos estos datos como si se tratara de variables categóricas nominales. Con este fin creamos una variable binaria para cada categoría y para recodificar los datos. No aplicamos el AC a los datos expresados en la escala original de 1 a 5. En la escala original, el concepto de perfil no tendría sentido ya que el perfil de la respuesta [1 1 1 1] —muy de acuerdo con las cuatro afirmaciones— y el de la respuesta [5 5 5 5] —muy en desacuerdo con las cuatro afirmaciones— serían iguales. Otro tipo de escalas de grados que encontramos a menudo en ciencias sociales y en investigación de mercados son:

— Una escala de nueve puntos (añadimos una categoría extra entre los puntos de la escala de cinco puntos):

| □ | □ | □ | □ | □ | □ | □ | □ | □ |
|---|---|---|---|---|---|---|---|---|
| *Muy de acuerdo* | | *Bastante de acuerdo* | | *Ni de acuerdo ni en desacuerdo* | | *Algo en desacuerdo* | | *Muy en desacuerdo* |

— Una escala de importancia de cuatro puntos:

| □ | □ | □ | □ |
|---|---|---|---|
| *Nada importante* | *Bastante importante* | *Muy importante* | *Extremadamente importante* |

— Una escala semántica diferencial de siete puntos en una encuesta sobre la satisfacción de los clientes:

| *Servicio antipático* | □ | □ | □ | □ | □ | □ | □ | *Servicio simpático* |
|---|---|---|---|---|---|---|---|---|

— Una escala de grados continua (por ejemplo, una escala de 0 a 10):

*Muy insatisfecho* 0 ——————————————— 10 *Muy satisfecho*

En este último ejemplo el encuestado puede escoger cualquier valor entre 0 y 10, incluso, si quiere, con decimales. Sin embargo, seguimos considerando los datos como procedentes de una escala de grados. En consecuencia, llevaremos a cabo la recodificación de forma similar a la de los ejemplos anteriores. Fijémonos, sin embargo, en que cuando el número de puntos de las escalas es grande es poco manejable utilizar variables binarias para codificar el ACM.

Doblado de la escala de grados

El *doblado* es el procedimiento de recodificación que utilizamos habitualmente en AC para datos procedentes de escalas de grados. Lo que hacemos es redefinir las escalas de grados como un par de escalas complementarias; el polo «positivo» o «elevado», y el polo «negativo» o «bajo». Antes de llevar a cabo el doblado es recomendable que el extremo inferior de la escala de grados sea igual a cero. Así, por ejemplo,

| *Pregunta* | | | | *Pregunta A* | | *Pregunta B* | | *Pregunta C* | | *Pregunta D* | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *A* | *B* | *C* | *D* | *A–* | *A+* | *B–* | *B+* | *C–* | *C+* | *D–* | *D+* |
| 2 | 3 | 4 | 3 | 1 | 3 | 2 | 2 | 3 | 1 | 2 | 2 |
| 3 | 4 | 2 | 3 | 2 | 2 | 3 | 1 | 1 | 3 | 2 | 2 |
| 2 | 3 | 2 | 4 | 1 | 3 | 2 | 2 | 1 | 3 | 3 | 1 |
| 2 | 2 | 2 | 2 | 1 | 3 | 1 | 3 | 1 | 3 | 1 | 3 |
| 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| ⋮ | ⋮ | ⋮ | ⋮ | | ⋮ | | ⋮ | | ⋮ | | ⋮ |
| … y así para las 871 filas | | | | | | | | | | | |

**Imagen 23.1:** *Datos originales correspondientes a las variables sobre la ciencia y el medio ambiente, y codificado doble, para los cinco primeros encuestados de $N = 871$ (muestra de Alemania Occidental)*

si trabajamos con las escalas de 1 a 5 y de 1 a 7, restaríamos 1 a los valores de las escalas para obtener las escalas de 0 a 4 y de 0 a 6, respectivamente. Estas nuevas escalas definen directamente el polo positivo. Damos por supuesto que los valores altos corresponden a valores sustantivamente altos de la escala (por ejemplo, mucha satisfacción, mucha importancia, muy de acuerdo). En las escalas de acuerdo/en desacuerdo que vimos anteriormente, los valores altos corresponden a un gran desacuerdo. Por tanto, en este caso, para evitar confusiones antes de proseguir el análisis tendríamos que haber invertido la escala. Obtenemos el polo negativo restando a $M$ —el mayor valor del polo positivo (4, 6 u 8 para las escalas de grados, y 10 para la escala 0-10 que vimos anteriormente)— los valores de la escala del polo positivo. En la tabla de la imagen 23.1 hemos ilustrado este procedimiento para los datos sobre ciencia y medio ambiente que vimos en el capítulo 20. En la tabla de la imagen 23.1 mostramos las primeras 10 filas de datos y sus homólogos doblados. Por ejemplo, el primer valor del encuestado 1 es un 2, restando 1 obtenemos el valor 1, su valor doblado es 3. Por tanto, para la pregunta 1, aparecen los valores 1 y 3 en la primera columna de doblado que hemos etiquetado como *A*– y *A*+ para indicar que los valores calculados cuantifican el desacuerdo y el acuerdo, respectivamente, con relación a la primera pregunta. De manera similar, el valor original 3 de la segunda pregunta se convierte en un 2 y en un valor doblado de 2, es decir, valores iguales para los polos de desacuerdo y de acuerdo, *B*– y *B*+, y así sucesivamente.

Paradigma de recuento

Podemos considerar los valores doblados como si fueran recuentos. Efectivamente, los valores doblados 1 y 3 indican el número de puntos de la escala que quedan por debajo y por encima, respectivamente, del valor observado 1 (en la escala que empieza por cero). En la escala original, la respuesta 2 («bastante de acuerdo» tiene, en la escala, un punto por debajo [1] y tres por encima [3, 4 y 5]). De la misma forma, la respuesta 3, «ni de acuerdo ni en desacuerdo», se halla en el centro de la escala, ya que tiene dos puntos por encima y dos por debajo. Es decir, la tabla de datos doblados sustituye los datos originales midiendo la asociación entre cada encuestado y los polos de acuerdo y desacuerdo de la escala de grados. Es necesario medir esta asociación con ambos polos: si utilizáramos los valores de un

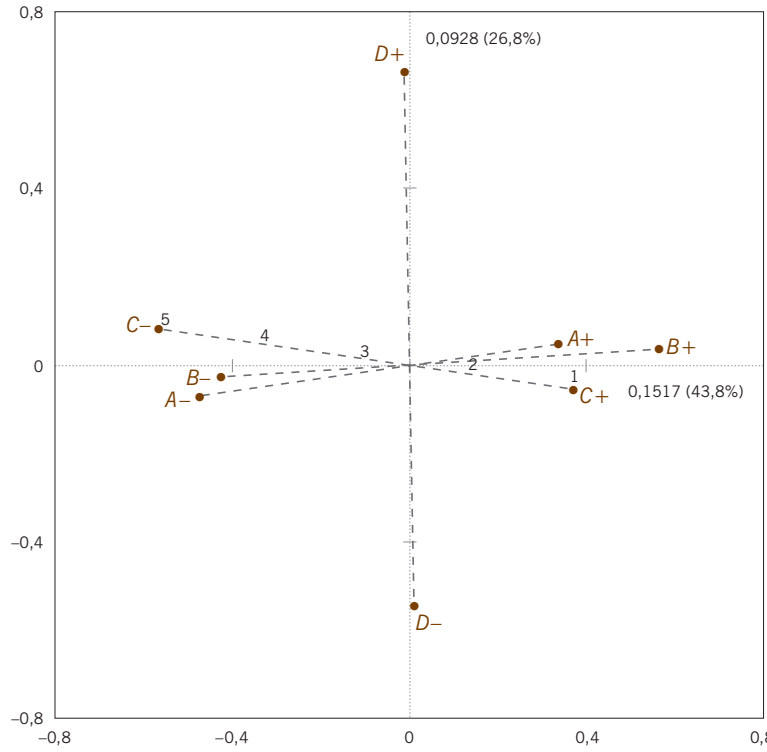



**Imagen 23.2:** *AC correspondiente al codificado doble de los datos sobre* ciencia *y* medio ambiente, *que muestra sólo los valores derivados del codificado doble. El porcentaje de inercia explicada es del 70,6%. En cada uno de los ejes, podemos imaginar la escala de rangos a intervalos iguales, conectando los polos (es decir, la escala de 1 a 5, de la pregunta C). La media de cada pregunta se halla exactamente en el origen*

solo polo, los perfiles del AC no tendrían sentido ya que, por ejemplo, estar muy de acuerdo o muy en desacuerdo con todas las preguntas tendrían los mismos perfiles y, por tanto, la misma posición en el mapa.

Mapa del AC con la escala de grados doble

Apliquemos el AC a la tabla de 8 columnas y 871 filas situada a la derecha de la imagen 23.1. Todas las filas dan la misma suma (16 en este ejemplo). Por tanto, las masas de los encuestados (filas) son iguales y, efectivamente, no tiene porqué haber diferencias en los pesos de los encuestados. Los cuatro pares de columnas dan la misma suma, por tanto, hay cuatro restricciones lineales en las columnas y no solamente una como en el AC habitual. En consecuencia, la dimensionalidad de la matriz de datos es 8 – 4 = 4. La inercia total y su descomposición en los cuatro ejes principales es la siguiente:

$$0{,}3462 = 0{,}1517\ (43{,}8\%) + 0{,}0928\ (26{,}8\%) + 0{,}0529\ (15{,}3\%) + 0{,}0488\ (14{,}1\%)$$

En el mapa de la imagen 23.2 hemos representado las columnas en coordenadas principales. Para cada pregunta tenemos dos puntos. Como muestran las líneas de trazo discontinuo, los polos positivos se hallan, con relación al origen, opuestos a sus homólogos negativos. Vemos claramente que la pregunta *D* se halla fuera de las alineaciones que muestran las otras tres preguntas. Lo mismo que vimos

en el capítulo 20. Quizás habríamos esperado que *D*– estuviera a la derecha y que *D*+ estuviera a la izquierda; en cualquier caso estas variables forman casi un ángulo recto con las restantes.

Los ejes de la escala de grados tienen el origen en la media

Los cuatro «ejes» de las escalas de grados pasan por el origen del mapa. Para recuperar la escala original, podemos subdividir las líneas discontinuas entre polos en cuatro intervalos iguales, y etiquetar los cinco puntos resultantes, como mostramos para la pregunta *C* utilizando las etiquetas de la escala original de 1 a 5 (1 corresponde a «muy de acuerdo»). Para todas las preguntas se cumple que el origen del mapa corresponde a la media de los valores de las respuestas en sus correspondientes escalas. Así, podemos ver en el mapa que las medias de las respuestas a las preguntas *A* y *C* se hallan más en el lado de acuerdo (+) de la escala de grados (para la pregunta *C* la media es 2,58), mientras que las medias de las preguntas *B* y *D* se hallan ligeramente en el lado de desacuerdo. Otra forma de verlo sería imaginar que los pesos de los puntos extremos de cada eje de escala de grados es proporcional a la media de los valores de su polo (de esta manera *C*+ se halla más próximo al origen que *C*–, porque es «más pesado»).

Correlaciones aproximadas por los cosenos de los ángulos

Los cosenos de los ángulos formados por los cuatro ejes del mapa de la imagen 23.2 son, aproximadamente, las correlaciones entre variables. Por tanto, las variables *A*, *B* y *C* estarán correlacionadas positivamente ente sí, mientras que no lo están con *D*. Los coeficientes de correlación de las cuatro variables son los siguientes:

| *Preguntas* | *A* | *B* | *C* | *D* |
|---|---|---|---|---|
| *A* | 1 | 0,378 | 0,357 | 0,036 |
| *B* | 0,378 | 1 | 0,436 | 0,016 |
| *C* | 0,357 | 0,436 | 1 | –0,062 |
| *D* | 0,036 | 0,016 | –0,062 | 1 |

Ello concuerda con nuestra deducción visual. Debido a que el mapa sólo explica el 70% de la inercia de los datos, no hemos recuperado de forma exacta los valores de las correlaciones. Por ejemplo, *B* y *C* deberían formar un ángulo más pequeño que *A* y *B*. Lo veríamos de forma más precisa en una representación tridimensional de los ejes de las escalas de grados.

Posiciones de filas y puntos adicionales

Igual que en el AC habitual, cada encuestado tiene un perfil y, por tanto, una posición en el mapa. No obstante, como ocurre en el ACM con datos procedentes de encuestas de grandes muestras, para nosotros tiene más interés representar como puntos adicionales grupos de individuos que las posiciones de cada individuo. Para ilustrarlo consideremos, para los datos anteriores, los datos clasificados en hombres y mujeres de seis grupos de edad, es decir 12 grupos. Podríamos calcular los valores medios de los mencionados grupos en la escala de grados. Lue-

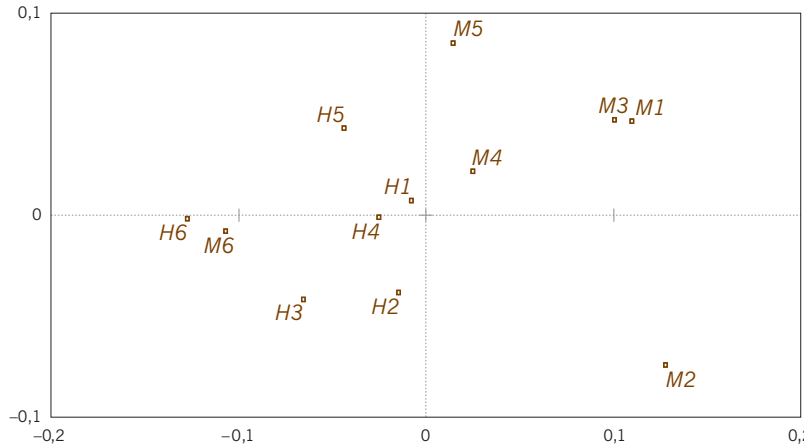



**Imagen 23.3:** *Puntos adicionales correspondientes a hombres y mujeres de cinco grupos de edad. Todas las mujeres se hallan en el lado derecho (acuerdo), mientras que los hombres —con excepción del grupo de mayor edad— se hallan en el lado de desacuerdo*

go los añadimos en el mapa como filas adicionales (dobladas). En el mapa de la imagen 23.3 mostramos sus posiciones. Vemos que todos los grupos de mujeres se sitúan a la derecha del mapa. Es decir, en el lado de acuerdo de las preguntas *A*, *B* y *C*. A excepción del grupo de hombres de mayor edad, los grupos de hombres se hallan en el lado de desacuerdo de estas preguntas, es decir, son menos críticos con el papel de la ciencia en el medio ambiente.

Datos de preferencias

Podríamos considerar los datos de preferencias como un caso especial de datos en escalas de grados y así visualizarlos en AC como estos últimos. En investigación de mercados es habitual pedir a los encuestados que ordenen productos de más a menos según su preferencia, o que ordenen atributos de los productos de más a menos importantes. Así, supongamos, por ejemplo, que tenemos seis productos, de *A* a *F*, y que un determinado individuo los ordena de la siguiente manera:

más preferidos : $B > E > A > C > F > D$ : menos preferidos

De acuerdo con esta ordenación, a cada uno de los seis productos le corresponde los siguientes rangos:

| *A* | *B* | *C* | *D* | *E* | *F* |
|---|---|---|---|---|---|
| 3 | 1 | 4 | 6 | 2 | 5 |

Podemos considerar estos rangos como derivados de una escala de grados de seis puntos. La diferencia con la escala de rangos habitual es que los encuestados se han visto forzados a utilizar una sola vez cada uno de los valores de la escala. Sin embargo, los podemos doblar de la manera habitual. En las etiquetas asignadas a las columnas dobles, + indica mucha preferencia, y – muy poca preferencia:

| *A*– | *A*+ | *B*– | *B*+ | *C*– | *C*+ | *D*– | *D*+ | *E*– | *E*+ | *F*– | *F*+ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 3 | 0 | 5 | 3 | 2 | 5 | 0 | 1 | 4 | 4 | 1 |

**Imagen 23.4:** *Indicadores económicos de la Unión Europea y sus rangos del menor a mayor*

| | *Datos originales* | | | | | *Rangos de los datos* | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Países | *TD* | *PIB* | *CI* | *CCI* | *CLUR* | *TD* | *PIB* | *CI* | *CCI* | *CLUR* |
| Bélgica | 8,8 | 102 | 104,9 | 3,3 | 89,7 | 7 | 7 | 7 | 7,5 | 5,5 |
| Dinamarca | 7,6 | 134,4 | 117,1 | 1 | 92,4 | 5 | 12 | 11 | 1 | 8 |
| Alemania | 5,4 | 128,1 | 126 | 3 | 90 | 3 | 11 | 12 | 6 | 7 |
| Grecia | 8,5 | 37,7 | 40,5 | 2 | 105,6 | 6 | 2 | 2 | 2 | 12 |
| España | 16,5 | 67,1 | 68,7 | 4 | 86,2 | 12 | 4 | 4 | 11 | 3 |
| Francia | 9,1 | 112,4 | 110,1 | 2,8 | 89,7 | 8 | 9 | 9 | 4,5 | 5,5 |
| Irlanda | 16,2 | 64 | 60,1 | 4,5 | 81,9 | 11 | 3 | 3 | 12 | 2 |
| Italia | 10,6 | 105,8 | 106 | 3,8 | 97,4 | 10 | 8 | 8 | 10 | 10 |
| Luxemburgo | 1,7 | 119,5 | 110,7 | 2,8 | 95,9 | 1 | 10 | 10 | 4,5 | 9 |
| Países Bajos | 9,6 | 99,6 | 96,7 | 3,3 | 86,6 | 9 | 6 | 5 | 7,5 | 4 |
| Portugal | 5,2 | 32,6 | 34,8 | 3,5 | 78,3 | 2 | 1 | 1 | 9 | 1 |
| Reino Unido | 6,5 | 95,3 | 99,7 | 2,1 | 98,9 | 4 | 5 | 6 | 3 | 11 |

TD: tasa de desempleo (%); PIB: producto interior bruto per cápita (índice); CI: consumo individual (índice); CCI: cambio en el consumo individual (%); CLUR: costes laborales unitarios reales

De todas formas, es habitual que los encuestados puedan ordenar sólo los objetos más preferidos (por ejemplo, los tres más preferidos). En tal caso, asignamos el mismo rango a todos los objetos no seleccionados (la media de los rangos de las posiciones no seleccionadas). Por ejemplo, si ordenáramos sólo los tres mejores de seis productos, asignaríamos a los tres productos omitidos el rango 5, la media de 4, 5 y 6.

Comparaciones por pares

Las *comparaciones por pares* son un tipo de ordenación más libre que la ordenación por preferencias. Por ejemplo, supongamos que presentamos a los encuestados los 15 pares posibles de seis productos, de *A* a *F*, y le pedimos que elija los pares preferidos. Haríamos el doblado de las respuestas de los encuestados de la siguiente manera:

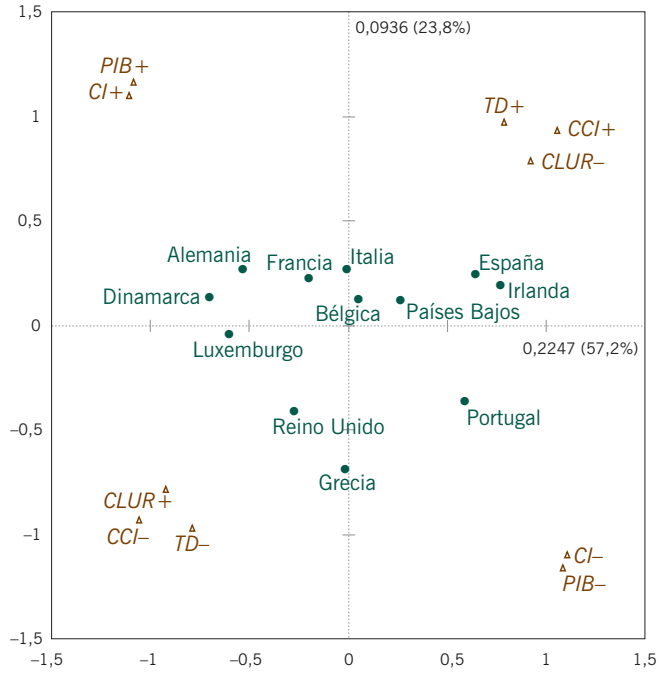
*A*+: número de veces que se ha preferido A a los restantes productos
*A*–: número de veces que se han preferido los restantes productos y no *A* ( = 5 – *A*+),

y así sucesivamente con los demás productos. Luego, igual que antes, aplicamos el AC a los datos doblados.

Conjunto de datos 13: indicadores de la Unión Europea

Mediante AC, también podemos visualizar datos procedentes de escalas continuas, si previamente los recodificamos convenientemente. Con este fin, existen distintas posibilidades. Como ejemplo, consideremos los datos situados a la izquierda de la tabla de la imagen 23.4. Se trata de cinco indicadores económicos de 12 países de la Unión Europea a principios de los años noventa, que se han expresado en distintas escalas. Así, por ejemplo, la tasa de desempleo y el cambio en el consumo individual se expresan como porcentajes.

**Imagen 23.5:**
*Mapa asimétrico del AC correspondiente a la recodificación, mediante rangos, de los indicadores de la Unión Europea. La inercia explicada es del 81,0%*




Recodificación de datos continuos, ordenación y doblado

Para recodificar estos datos, primero expresaremos los valores de cada variable como rangos tal como se muestra en la tabla de la imagen 23.4. Así, por ejemplo, vemos que Luxemburgo tiene el desempleo más bajo y que, por tanto, le corresponde el rango 1, luego Portugal con el rango 2, etc. A los valores iguales les asignaremos también rangos iguales, o sea, la media de sus rangos. Por ejemplo, Francia y Luxemburgo tienen el mismo valor para el *CCI*, están empatados ocupando las posiciones 4 y 5. Les asignaremos como rango la media de 4 y 5, es decir 4,5. Una vez expresados los valores de cada variable como rangos, podemos llevar a cabo el doblado tal como hicimos anteriormente. Es decir, en primer lugar, para obtener el polo positivo restamos 1 a los valores de los rangos. Luego obtenemos el polo negativo como: 11 menos el valor del polo positivo. En la imagen 23.5, mostramos el mapa del AC de la matriz doblada. De nuevo, hemos unido los polos opuestos de cada variable. En este caso observamos que las distancias de los polos al origen son iguales. Ello se debe a que las medias de sus rangos son idénticas (por tanto, bastaría con dibujar el polo positivo). En el mapa podemos ver dos grupos de variables muy poco correlacionadas, pero muy correlacionadas dentro de cada grupo. Así, por ejemplo, fijémonos en que *CLUR* (costes laborales unitarios reales) están muy negativamente correlacionadas con *TD* (tasa de desempleo) y con *CCI* (cambio consumo individual); como estamos utilizando rangos, cuando hablamos de correlaciones, nos referimos a la *correlación por rangos de Spearman*. La posición de cada país depende de los valores de los rangos de cada variable y no de su valor concreto. Por tanto, dado que analizamos

rangos, y no valores originales, el análisis será robusto con relación a las observaciones atípicas. Así, pues, estamos llevando a cabo un AC *no paramétrico*.

Otras posibilidades de recodificación para datos continuos

La expresión de variables continuas como rangos conlleva la pérdida de algo de información. Sin embargo, nuestra experiencia nos muestra que esta pérdida es mínima por lo que respecta a la visualización de los datos. Por el contrario, en muchas situaciones, la robustez de los rangos es una ventaja. De todas formas, si necesitamos toda la información contenida en los datos, existen otras posibilidades. Por ejemplo, una transformación adecuada consiste en estandarizar todas las variables (puntuaciones $z$). Es decir, restamos a cada variable su media y esta diferencia la dividimos por su desviación típica. A continuación, a partir de $z$ creamos dos versiones de cada variable utilizando la remodificación siguiente:

$$\text{valor positivo} = \frac{1+z}{2} \qquad \text{valor negativo} = \frac{1-z}{2} \tag{23.1}$$

A pesar de que obtenemos algunos valores negativos, los valores marginales de filas y columnas se mantienen positivos, e iguales para todas las filas y para todos los pares de columnas dobladas. Por tanto, la ponderación es igual para todos los casos y todas las variables. El AC de esta matriz doblada proporciona un mapa casi idéntico al mapa de la imagen 23.5. Como curiosidad, señalar que hasta donde llega nuestro conocimiento, se trata del único caso de matriz de datos con algunos valores negativos que podemos analizar de forma válida utilizando el AC.

RESUMEN: Recodificación de datos

1. Podemos recodificar datos procedentes de diferentes escalas de medida, de manera que sean adecuados para el AC.
2. Siempre que la matriz recodificada posea perfiles y sumas marginales que tengan sentido en el contexto de la aplicación, el AC proporcionará una visualización correcta de los datos.
3. Uno de los principales procedimientos de recodificación consiste en *doblar* las variables. Es decir, convertir cada variable en un par de variables para que la suma de los pares de variables sea constante.
4. Podemos llevar a cabo el doblado en el caso de escalas de grados, de preferencias y de comparaciones por pares de objetos. Obtenemos mapas en los que cada variable queda representada por dos puntos opuestos con relación al origen. En el caso particular de las variables expresadas en escalas de grados, el origen del mapa indica el valor medio de la variable que estemos considerando en relación con la escala delimitada por los dos polos extremos.
5. Podemos recodificar datos continuos como rangos dobles, lo que nos conduce a una forma no paramétrica del AC. Otra posibilidad es transformarlos en una par de variables continuas, a partir de sus valores estandarizados.