Infancia y Aprendizaje, 1987, 39-40, 171-186

Análisis de correspondencias y técnicas de clasificación: Su interés para la investigación en las ciencias sociales y del comportamiento

*Joan Manuel Batista**
Universidad de Barcelona
Joan Sureda
E.S.A.D.E., Barcelona

1. INTRODUCCION

La estadística clásica se ha caracterizado por desarrollar nociones descriptivas e inferenciales de estimación o contrastes de hipótesis en un marco fundamentalmente uni-bidimensional. Sin embargo, los fenómenos psicológicos y sociales, biológicos, físicos o económicos, objeto de estudio de la estadística actual, son complejos: consideran gran número de aspectos, medidos en diferentes tipos de escalas y con diversas finalidades. El objetivo del estudio puede centrarse en los individuos —sus diferencias o similitudes—, o/y en las variables, su interrelación o explicación de una(s) en función de las restantes.

Por tanto, la estadística es hoy multivariante (1), considera múltiples medidas, continuas o no, sobre un conjunto de individuos que pueden provenir de una o más poblaciones. En esencia, todos los métodos pretenden simplificar la complejidad del estudio con una pérdida mínima de información, lo que se logra examinando la dependencia o interdependencia entre las variables implicadas con la ayuda de adecuadas representaciones gráficas.

En este trabajo se introducen las técnicas multivariantes de análisis de correspondencias y de análisis de agrupaciones (Cluster Analysis), también utilizadas en el segundo artículo de J. Palacios publicado en este mismo número. Un sencillo ejemplo de datos ocupacionales en las distintas comunidades autónomas permitirá ilustrar los conceptos introducidos.

2. CONCEPTOS BASICOS

El estudio de las relaciones entre variables cualitativas —medidas en escalas nominales— se ha llevado tradicionalmente a cabo partiendo de una tabla rectangular de números positivos, cuyas filas y columnas representan las distintas

* *Dirección del primer autor*: Departamento de Metodología de las Ciencias del Comportamiento. Facultad de Psicología, C/ Adolf Florensa, s.n. 08028 BARCELONA.

categorías de cada una de las variables relacionadas. A este tipo de matriz se la conoce como tabla de contingencia o tabla cruzada. A continuación, el análisis y verificación de la relación de dependencia se efectúa basándose casi exclusivamente en la prueba X^2 (Test de Chi-Cuadrado), o en otros índices derivados de la misma.

Aunque los trabajos de Hirsfeld (1935), R. Fisher (1940) y L. Guttman (1941) sobre tablas de contingencia pueden considerarse como precursores del método, el Análisis Factorial de Correspondencias o simplemente Análisis de Correspondencias (AC) es una técnica relativamente reciente. El término de AC fue propuesto por primera vez por J. P. Benzecri en 1962, y el desarrollo de esta técnica, poniendo el énfasis en los aspectos algebraicos y geométricos más que en la aplicación inferencial estadística planteada por R. Fisher, se debe a la escuela de estadística francesa.

Inicialmente propuesto como método inductivo para análisis lingüísticos (Benzecri, 1977), cuyos datos eran recogidos en una matriz que disponía en lógica «correspondencia» consonantes en las filas y vocales en las columnas, para poner de manifiesto el sistema de asociaciones entre los elementos de ambos conjuntos siguiendo la filosofía estructuralista, rápidamente se extenderá como técnica estadística descriptiva de análisis multivariable a numerosos campos siempre que la información recogida sobre las variables sea de naturaleza nominal o recodificada como tal (2).

El AC se configura en la década de los 60, en la que las aplicaciones lingüísticas eran escasas y por contra las de la Psicometría americana eran múltiples, algunas con objetivos similares a los del AC. Por todo ello, recibe fuertes influencias: 1.º) del Análisis de Proximidades de Shepard; éste, basándose en una matriz cuyo formato era el propio del AC, ponía en correspondencia Estímulos y Respuestas para representar en un espacio de menor dimensión las relaciones de proximidad entre los elementos de ambos conjuntos; 2°) de los algoritmos de análisis de las tablas de doble entrada de D. Carroll y J. J. Chang; 3°) de la implantación de las doctrinas de Análisis Factorial, cuyos problemas matemáticos de ajuste se habían resuelto en 1936 por Eckart y Young; y 4°) del escalonamiento multidimensional que básicamente analizaba los resultados de pruebas psicotécnicas con el fin de representarlos geométricamente en un espacio de dimensión reducida (Benzecri, 1977). Como consecuencia, muchos de los algoritmos de cálculo y de las estructuras matemáticas del AC son identificables a las utilizadas en Análisis de Componentes Principales (ACP), Escalonamiento Multidimensional (MDS) y Análisis Discriminante o Análisis Canónico.

En la actualidad, del mismo modo que el AC permite analizar la interdependencia entre variables categóricas jugando el mismo papel que el Análisis en Componentes Principales desempeña con variables de tipo continuo, los modelos log-lineales permiten construir modelos explicativos de variables de tipo categórico (Upton, 1978), de forma análoga en que el modelo lineal —Regresión o Análisis de la Varianza— se utiliza para variables continuas.

3. OBJETIVO Y DOMINIO DE APLICACION

Si bien el método puede considerarse bajo la óptica de la reducción de datos (Sánchez Carrión, 1985) y por ello próximo al Análisis en Componentes Principales, la características principal del AC consiste en que obtiene la mejor representación simultánea de los dos conjuntos de datos definidos respectivamente por las filas y las columnas de la matriz de contingencia dada.

Concretamente, el AC consiste en un conjunto de técnicas que en base a una matriz de contingencia X (I, J), cuyo elemento X_{ij} indica el número de veces

que se ha presentado asociada la categoría i-ésima de una variable con la j-ésima de la otra, trata de obtener una representación gráfica de filas y columnas de la matriz de contingencia, de forma que se pongan de manifiesto las relaciones existentes dentro de cada conjunto (filas y columnas) y entre ambos.

En la matriz de partida X(I, J) en cada casilla (i, j) se encuentra un número no negativo que representa la frecuencia absoluta con la que se han presentado asociadas ambas modalidaddes de las dos variables consideradas. Sin embargo, debido fundamentalmente a la distancia utilizada en esta técnica (véase apart. 5.2) puede asimismo aplicarse a tablas de descripcion lógica sobre la presencia o ausencia de un determinado carácter en un individuo, o generalizarse a tablas en las que se consideran en correspondencia más de dos variables nominales (Análisis de Correspondencias Múltiples; ver el ya citado segundo artículo de J. Palacios), y en general a cualquier medida no negativa homogénea y exhaustiva.

Esta técnica resulta especialmente útil para el análisis de matrices de datos de grandes dimensiones, cuya magnitud impide interpretar las relaciones existentes y resulta por ello apropiada para obtener descripciones sintéticas. Esta filofía exploratoria de los datos no debe entenderse como de mera descripción ya que, por ende, estas técnicas posibilitan el análisis, la comprobación y verificación de hipótesis previas, aunque no puedan considerarse técnicas de análisis estadístico confirmatorio en el sentido de Jöreskog (1969).

Los rasgos estructurales que revele el análisis pueden reflejar no tanto los fenómenos que pretendemos estudiar, cuanto el método adoptado en la recogida de datos. La validez de los resultados será tanto mayor cuanto más homogéneos y exhaustivos sean los datos recogidos del fenómeno a un nivel dado.

4. UN EJEMPLO

Considérese la distribución de la ocupación en los sectores Agrícola (Agr), Industrial (Ind), Construcción (Con) y Servicios (Ser) junto con el desempleo (Par) observado en el conjunto de comunidades autónomas del Estado Español (tabla I) (3).

TABLA I

Distribución ocupacional por sectores. (en miles de personas)

COMUNIDAD	Agric.	Industr.	Constr.	Servicio	Paro	TOTAL
ANDALUCIA	261.4	211.4	108.8	732.5	597.5	1911.6
ARAGON	66.2	89.0	23.0	165.8	70.7	414.7
BALEARES	20.9	43.2	26.6	137.1	39.2	267.0
CANARIAS	66.8	43.9	38.5	271.3	142.7	563.2
CANTABRIA	36.4	42.2	10.6	66.3	30.5	186.0
CASTILLA-MANCHA	115.0	84.0	45.1	164.2	76.3	484.6
CASTILLA-LEON	178.7	127.8	52.4	288.7	150.2	797.8
CATALUÑA	114.8	674.0	120.7	895.6	519.3	2324.4
VALENCIA	159.4	312.2	66.4	561.1	281.9	1381.0
EXTREMADURA	69.1	23.1	20.2	97.3	80.9	290.6
GALICIA	455.3	154.4	65.0	322.7	148.7	1146.1
MADRID	23.2	306.7	93.0	932.4	414.3	1769.6
MURCIA	50.8	58.5	19.6	126.5	60.3	315.7
NAVARRA	21.8	47.7	10.3	66.9	35.7	182.4
PAIS VASCO	35.9	232.8	40.1	316.2	192.7	817.7
ASTURIAS	78.6	94.1	24.0	142.8	78.6	418.1
RIOJA	12.0	23.0	5.0	30.2	14.3	84.5
Total España	1766.3	2568.0	769.3	5317.6	2933.8	13355.0

Esta tabla presenta las cifras absolutas de personas que en cada comunidad se adscriben a los sectores ocupacionales reseñados; así por ejemplo, vemos que en Baleares 43.200 personas se dedican a actividades del sector industrial, mientras que esta cifra es de 42.200 en Cantabria. La última columna, bajo el epígrafe TOTAL, presenta la suma de la población activa de cada comunidad; de la misma manera, la última fila (Total España) incorpora el total de personas ocupadas en cada sector.

Debe observarse que el análisis superficial de estos datos puede llevarnos a conclusiones precipitadas y erróneas. Por ejemplo, la comparación de las 43.200 personas del sector industrial en Baleares y las 42.000 en Cantabria no debe llevarnos a inducir que el sector industrial balear es más importante que el cántabro; en general, se debe matizar esta afirmación en función del tamaño de la comunidad. Así, vemos que Cantabria tiene una población activa sensiblemente menor que Baleares (186.000 frente a 267.000). Obviamente todos los sectores baleares tenderán a ocupar más personas que los cántabros; ello no implica el que un determinado sector sea más importante en Baleares que en Cantabria. En efecto, y siguiendo con el mismo ejemplo, nos encontramos que el sector industrial en Cantabria ocupa el 22.69 % de su población (42.2 sobre 186.0) mientras que este mismo sector ocupa sólo el 16.18 % (43.2 sobre 267.0) en Baleares.

A partir de este tipo de datos y considerando el tipo de problemas que puede plantear su análisis, el objetivo del AC consiste en determinar las relaciones de dependencia que existen entre estas dos variables nominales: Tipo de ocupación y Comunidad Autónoma. La dependencia entre variables nominales se plantea siempre en los siguientes términos: 1º) ¿Qué modalidades o grupo de modalidades (4) de una variable (Comunidad Autónoma) se comportan de forma similar respecto de todas las modalidades (Tipo de ocupación) de la otra?, y viceversa, ¿qué comunidades pueden considerarse similares con respecto al tipo de ocupación y qué sectores serían similares respecto a su distribución geográfica?; 2º) ¿cuáles son las modalidades o valores de una variable que explicarían las diferencias observadas en la otra?; por ejemplo, ¿qué sectores explican las diferencias entre dos comunidades?; y 3º) ¿cómo podemos representar gráficamente las conclusiones obtenidas, de forma que las interrelaciones se distorsionen lo menos posible? Esta última pregunta es fundamental, pues ya con una tabla como la del ejemplo, de dimensiones relativamente reducidas ($p = 5$, $q=17$), no es posible descubrir fácilmente la respuesta a las dos primeras preguntas.

5. PERSPECTIVA GEOMETRICA

5.1. Tabla de perfiles

Una tabla de contingencia como la Tabla I, siempre puede leerse de dos maneras: considerando los I puntos-fila (comunidades autónomas) representados en el espacio R^J de los sectores ocupacionales, o alternativamente considerar los J puntos-columna (sectores ocupacionales) en el espacio R^I de las autonomías. Así, por ejemplo, el punto representativo de la comunidad de Aragón tiene por coordenadas el número de empleados en cada sector y el paro observado.

Para evitar que las modalidades de cada variable con marginales más elevados (5) tengan un peso excesivo en el análisis, se consideran como punto de partida las frecuencias relativas en lugar de las absolutas, es decir, la distribución condicionada de la ocupación por sectores para una comunidad dada. Para ello se construyen, a partir de la original X(I, J), las matrices de perfiles de las

autonomías (tabla IIa) dividiendo cada fila por los marginales respectivos. De esta manera, como hemos comentado más arriba, obtenemos una tabla que pone de manifiesto la estructura de la población activa en cada una de las comunidades. Por ejemplo, vemos que Rioja ocupa el 14,2 % de su población activa en la agricultura, el 27,22 % en la industria, el 5.92 % en construcción, etc. Mediante esta transformación se elimina, como se pretendía, el efecto del tamaño de la comunidad para poder representar adecuadamente la similitud o disimilitud estructural entre las distintas comunidades. Es especialmente relevante en el análisis la última fila de esta tabla, que responde a la idea del perfil medio, o en nuestro caso, comunidad autónoma media: distribución en el conjunto del Estado Español de la población activa por sectores (el 13.23 % de la población activa española se ocupa en agricultura, el 19.23 % en industria, etc.).

TABLA IIa

Tabla de perfiles de Comunidades Autónomas

	AGR	IND	CON	SER	PAR
Andalucía (AND)	13.67	11.06	5.69	38.32	31.26
Aragón (ARA)	15.96	21.46	5.55	39.98	17.05
Baleares (BAL)	7.83	16.18	9.96	51.35	14.68
Canarias (CAN)	11.86	7.79	6.84	48.17	25.34
Cantabria (SAN)	19.57	22.69	5.70	35.65	16.40
Castilla-Mancha (C-M)	23.73	17.33	9.31	33.88	15.74
Castilla-León (C-L)	22.40	16.02	6.57	36.19	18.83
Cataluña (CAT)	4.94	29.00	5.19	38.53	22.34
Valenciana (VAL)	11.54	22.61	4.81	40.63	20.41
Extremadura (EXT)	23.78	7.95	6.95	33.48	27.84
Galicia (GAL)	39.73	13.47	5.67	28.16	12.97
Madrid (MAD)	1.31	17.33	5.26	52.69	23.41
Murcia (MUR)	16.09	18.53	6.21	40.07	19.10
Navarra (NAV)	11.95	26.15	5.65	36.68	19.57
País Vasco (PVA)	4.39	28.47	4.90	38.67	23.57
Asturias (AST)	18.80	22.51	5.74	34.15	18.80
Rioja (RIO)	14.20	27.22	5.92	35.74	16.92
TOTAL ESPAÑA	13.23	19.23	5.76	39.82	21.97

Análogamente, se obtendrían los perfiles de los sectores (tabla IIb).

TABLA IIb

Tabla de perfiles de tipos de ocupación

	AGR	IND	CON	SER	PAR	MAR-GINAL
Andalucía (AND)	14.80	8.23	14.14	13.78	20.37	14.31
Aragón (ARA)	3.75	3.47	2.99	3.12	2.41	3.11
Baleares (BAL)	1.18	1.68	3.46	2.58	1.34	2.00
Canarias (CAN)	3.78	1.71	5.00	5.10	4.86	4.22
Cantabria (SAN)	2.06	1.64	1.38	1.25	1.04	1.39
Castilla-Mancha (C-M)	6.51	3.27	5.86	3.09	2.60	3.63
Castilla-León (C-L)	10.12	4.98	6.81	5.43	5.12	5.97
Cataluña (CAT)	6.50	26.25	15.69	16.84	17.70	17.40
Valenciana (VAL)	9.02	12.16	8.63	10.55	9.61	10.34
Extremadura (EXT)	3.91	0.90	2.63	1.83	2.76	2.18
Galicia (GAL)	25.78	6.01	8.45	6.07	5.07	8.58
Madrid (MAD)	1.31	11.94	12.09	17.53	14.12	13.25
Murcia (Mur)	2.88	2.28	2.55	2.38	2.06	2.36
Navarra (NAV)	1.23	1.86	1.34	1.26	1.22	1.37
País Vasco (PVA)	2.03	9.07	5.21	5.95	6.57	6.12
Asturias (AST)	4.45	3.66	3.12	2.69	2.68	3.13
Rioja (RIO)	0.68	0.90	0.65	0.57	0.49	.63

Estas tablas de perfiles traducen las diferencias de comportamiento de las categorías de una variable (autonomías o sectores) en cada categoría (sector o autonomía) de la otra, otorgando el mismo peso a cada modalidad de los dos conjuntos. Por ello, la proximidad de dos puntos-perfiles tanto en R^I (espacio definido por las comunidades) como en R^J (espacio definido por los sectores), puede interpretarse como similitud entre las modalidades que representan. Así, por ejemplo, CAT y PVA en la tabla IIa estarían representados en R^5 (espacio de los sectores) por puntos relativamente muy próximos dada la semejanza con que distribuyen su población activa entre los distintos sectores.

5.2. Distancia X^2 de Benzecri (6)

Si a partir de la tabla IIa se quisiera representar las distancias entre autonomías utilizando la tradicional distancia euclídea, nos encontraríamos con que aquellas modalidades con marginales elevados, como sería el caso del paro, colaborarían de forma decisiva en el cálculo de las distancias. Para evitar este posible efecto distorsionante, se corrigen las proyecciones sobre cada eje =1,...,5, por el factor $1/f_{.j}$, siendo los $f_{.j}$ las frecuencias relativas marginales o perfil del individuo medio, en este caso, la comunidad autónoma media, de la tabla IIa. De esta manera, las distancias en R^5 se define como:

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

Siendo:

- $d(i, i')$ = Distancia entre las comunidades i e i'.
- $f_{.j}$ = Frecuencia relativa marginal del sector j.
- f_{ij} = Frecuencia relativa de ocupados en el sector j en la comunidad autónoma i.
- $f_{i.}$ = Frecuencia relativa marginal de la comunidad i.
- $f_{ij}/f_{i.}$ = Frecuencias condicionadas del sector j en la comunidad i. (Datos de la tabla IIa).

Con esta corrección, de hecho, lo que se consigue es magnificar las diferencias relativamente grandes cuando $f_{.j}$ sea reducido. Así, por ejemplo, las diferencias entre C-M y C-L del orden de un 2.8 % en «CON» y un 2.3 % en «SER» no serán equivalentes, pues estarían ponderadas por los respectivos factores de 1/0.0576 y 1/0.3982. Idéntico razonamiento llevaría a determinar la expresión dual de (1) como la de la métrica adecuada en R^{17} entre sectores ocupacionales.

La distancia de X^2 verifica la propiedad de equivalencia distribucional que la hace especialmente adecuada para tratar con variables nominales. Esta básicamente supone que al agregar dos perfiles idénticos, las distancias entre dos puntos-modalidades de la otra variable permanecen inalterables y sólo varían ligeramente cuando estos perfiles son parecidos, en lugar de idénticos. Así pues, agregar o subdividir (recodificar) categorías homogéneas de una variable no representa modificar los resultados.

En el anexo 1 se resumen los pasos fundamentales de la técnica estadística subyacente en este método multivariante.

6. INTERPRETACION DE RESULTADOS

Como en el ACP, los resultados de un AC se presentan frecuentemente en forma de gráficos, pero a diferencia de los primeros, éstos reflejan las configu-

raciones de ambos conjuntos (filas o comunidades y columnas o sectores) simultáneamente (7), en los planos que forman los pares de ejes principales.

En primer lugar, como en el ACP, un índice global de la calidad de representación se encontrará en el porcentaje de la varianza total (traza de la matriz S en (3) del anexo 1) explicada por los ejes retenidos. En el ejemplo, tabla III, puede verse que sobre el espacio definido por los dos primeros ejes se recoge el 93 % (74.16 % y 18.88 %, respectivamente) de la variabilidad o inercia total. La magnitud de este valor garantiza que casi no se pierde información al considerar sólo dos coordenadas (sobre los ejes principales) en lugar de las cinco originales (sobre los sectores ocupacionales); por ello la representación gráfica (Figura 1) en dos dimensiones está apenas distorsionada, y reflejará claramente las proximidades entre perfiles, por otro lado, difícilmente observables en las tablas de perfiles originales.

TABLA III

Varianza o inercia total = 0.1328

FACTOR	VALOR PROPIO (8)	% VARIANZA	% VARIANZA ACUM.
1	0.0985	74.16	74.16
2	0.0251	18.88	93.04
3	0.0077	5.80	98.84
4	0.0015	1.16	100.00

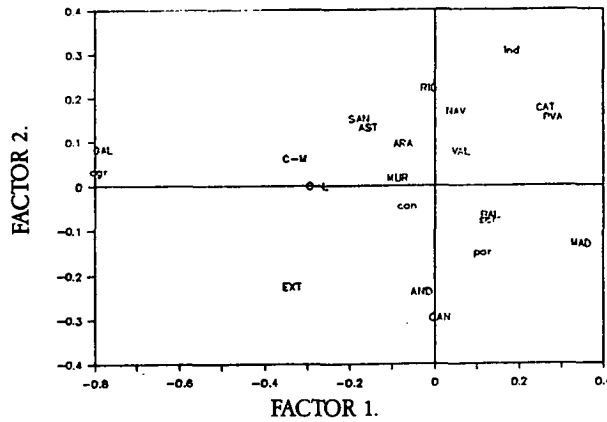
Como en ACP, es legítimo interpretar las distancias entre elementos de un mismo conjunto, filas o columnas. Por ejemplo, la posición alejada de GAL respecto del resto o las proximidades de SAN y AST, CAT y PVA o AND y CAN, indican respectivamente una distribución de la ocupación sectorial muy distinta del resto y perfiles similares dos a dos. No obstante, la interpretación de la distancia es tanto más factible cuanto más «periférico» sea el individuo, puesto que la distancia al origen se interpretará como desviación con respecto al individuo medio (M) por lo que el punto estará mejor caracterizado. Por el contrario, en la Fig. 1, el perfil de Murcia (MUR) se encuentra próximo al centro porque sin duda su distribución ocupacional por sectores es poco diferenciada; por el contrario, GAL está muy polarizada en un sector y por ello se aparta de la comunidad promedio. Análogo razonamiento conduce a considerar el sector Construcción (CON) como homogéneamente repartido en las distintas comunidades; su perfil es pues parecido al perfil del sector medio (distribución población activa por comunidades); por el contrario, es la Agricultura el sector con quizás mayor heterogeneidad en su comportamiento (concentración en determinadas comunidades y muy poco peso específico en otras).

El primer factor viene caracterizado fundamentalmente por el sector agricultura. Este es el primer gran hecho diferenciador entre comunidades. En la figura 1 puede observarse que las proyecciones de las distintas comunidades de izquierda a derecha reproducen la importancia de este sector en cada una de ellas. Los elementos extremos de esta polarización vienen representados por Galicia y Madrid.

El segundo gran elemento diferenciador o factor dos, aunque básicamente definido por el comportamiento del sector industria, viene a representar una cierta oposición entre éste y el paro. En este sentido, las comunidades que mejor reflejan este enfrentamiento son: Cataluña, por un lado (alto porcentaje del sector industrial y relativo bajo nivel de paro), frente a Andalucía y Canarias, con relativa baja importancia del sector industrial y alto nivel de paro.

FIGURA 1

Plano factorial 1-2

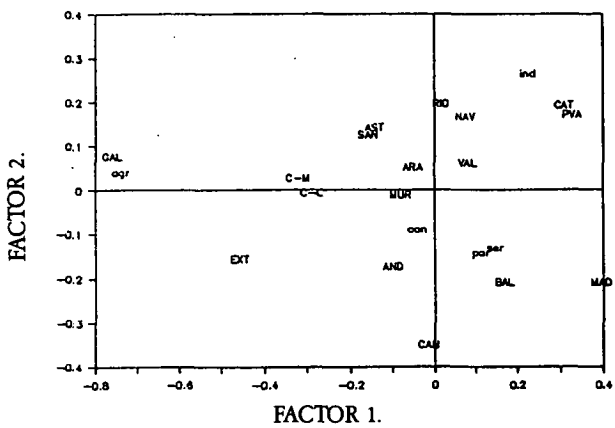


Así, puede decirse que la proximidad entre dos puntos-comunidad revela una distribución sectorial similar de la ocupación, mientras que dos sectores cercanos describen sectores de ocupación cuyo comportamiento es parecido en el conjunto de las autonomías. Por último, y aunque las nubes de puntos filas y columnas no se encuentran inicialmente en el mismo espacio, las relaciones baricéntricas mencionadas permiten interpretar la posición de cada autonomía con respecto al conjunto de sectores y, viceversa, cada sector respecto todas las comunidades. Aparecen en esta línea de interpretación comunidades autónomas industrializadas como RIO, NAV, CAT y PVA, de servicios como MAD y BAL, aquellas que sobre todo se caracterizan por ser poco industrializadas y con un alto índice de desempleo respecto a la media, como es el caso de AND, EXT y CAN.

Al interpretar la posición de un elemento de un conjunto en función de su posición relativa respecto a los elementos del otro conjunto, debe tenerse muy presente que, salvo distorsiones inevitables del análisis, cada punto se sitúa como baricentro de todos los elementos del otro conjunto. Es decir, Madrid y Baleares, por ejemplo, se situarán como centros de gravedad de todos los sectores en función de sus propios perfiles. Por ello, aunque lo que más caracteriza a ambas comunidades es el fuerte peso del sector servicios, la posición más alejada de Madrid debe entenderse como un alejamiento del sector agricultura, ya que éste apenas tiene incidencia en esta autonomía.

FIGURA 2

Plano factorial 1-2



Una vez obtenido el espacio que los ejes retenidos determinan, la interpretación puede enriquecerse posicionando en este mismo gráfico(s) o espacio «los individuos o variables suplementarios», es decir aquellos que no se han utilizado en el cálculo —variables no homogéneas con el resto o individuos pertenecientes a un grupo de control—. La figura 2 muestra los resultados del AC en el que el sector Paro se ha eliminado del análisis y se ha considerado como suplementario.

Las tablas Va y Vb proporcionan las coordenadas de las comunidades autónomas y de los sectores en los dos ejes o factores considerados, es decir, recogen la misma información no visual que suministra la figura 1. Sin embargo, la interpretación correcta de estos gráficos no sólo requiere tener en cuenta la proximidad de los puntos que se proyectan al plano definido por los ejes principales, sino también el papel que desempeña cada punto en la determinación de los ejes. Las precauciones en las interpretaciones anteriores deben extremarse si el porcentaje de varianza explicado por los dos ejes considerados en el plano es relativamente bajo. En general, se utilizan para este cometido dos series de índices: 1) *Contribuciones absolutas,* que miden el nivel de participación de cada elemento fila o columna en la construcción o definición del eje, lógicamente la CA_k (i) = 1; y 2) *Contribuciones relativas*, que determinan la proximidad del punto al eje mediante el cuadrado del coseno del ángulo que forma el punto respecto al eje; esta medida puede ser interpretada geométricamente como el cuadrado del coeficiente de correlación del elemento y el eje que se considere, es decir, la influencia del eje en la explicación de la distancia del punto al origen. En ésta se cumple que CR_K (i) = 1.

TABLA Va

Coordenadas comunidades en el espacio factorial

	FACTOR 1	FACTOR 2
Andalucía	—.0305	—.2373
Aragón	—.0729	.0950
Baleares	.1300	—.0647
Canarias	.0122	—.2954
Cantabria	—.1769	.1514
Castilla-Mancha	—.3293	.0609
Castilla-León	—.2774	.0010
Cataluña	.2625	.1748
Valencia	.0614	.0753
Extremadura	—.3375	—.2256
Galicia	—.7806	.0818
Madrid	.3474	—.1314
Murcia	—.0866	.0172
Navarra	.0509	.1677
País Vasco	.2789	.1521
Asturias	—.1559	.1302
Rioja	—.0136	.2215

TABLA Vb

Coordenadas factoriales sectores ocupacionales

	FACTOR 1	FACTOR 2
Agricultura	—.7930	.0337
Industria	.1790	.3057
Construcción	—.0647	—.0435
Servicios	.1244	—.0704
Paro	.1123	—.1489

A título de ejemplo, en la tabla VI incluimos los valores más relevantes de las contribuciones absolutas y relativas de los sectores y comunidades. Como se observa, pueden darse situaciones como la de C-L, comunidad que queda casi totalmente explicada por el factor 1 cuando prácticamente no ha contribuido a su formación. Este tipo de situaciones debe explicarse por la poca masa de elemento en cuestión (en nuestro caso, el 5.97 % de la población activa), que hace que el factor «agricultura» explique muy bien su distancia al origen, aunque, esta comunidad, debido a su poca masa, no haya colaborado apenas en la formación del eje.

TABLA VI

Contribuciones absolutas y relativas

	FACTOR 1		FACTOR 2	
	CA	CR	CA	CR
Andalucía			.3213	.7534
Canarias			.1468	.9276
Cantabria		.5673		
Castilla-Mancha		.8117		
Castilla-León		.9874		
Cataluña	.1218	.6718	.2120	
Valencia				.5188
Extremadura		.6379		
Galicia	.5309	.9841		
Madrid	.1624	.7931		
Murcia		.7090		
Navarra				.9001
País Vasco		.7324		
Asturias		.5826		
Rioja				.9864
AGRICULTURA	.8444	.9978		
INDUSTRIA		.2531	.7168	.7380
SERVICIOS		.5521		
PARO		.2320	.1941	.4079

7. TECNICAS DE CLASIFICACION

Debido al desarrollo de los computadores y a la relevancia científica de la clasificación en cualquier disciplina, en las últimas décadas se han desarrollado diversos algoritmos encaminados a determinar agrupaciones naturales de los individuos sujetos al análisis. Intentos de recoger los métodos desarrollados hasta el momento pueden encontrarse en Aldernderfer y Basfiell (1984), Benzecri (1973), Everitt (1980) y Gordon (1981).

El término anglosajón «Cluster Analysis» se utiliza para describir un conjunto de algoritmos, en general no estadísticos, que tienen por finalidad explorar una matriz de datos X (N × p) que incluya información con respecto a p características medidas sobre N individuos. Todos estos métodos tienen como objetivo común clasificar o diseccionar un conjunto de N individuos en subgrupos que difieren entre sí y que están formados, cada uno de ellos, por individuos homogéneos. Son sus medidas en las p características las que determinan las diferencias entre individuos. A diferencia de otras técnicas de clasificación, como el análisis discriminante, que asigna individuos a grupos previamente establecidos, aquí se trata de identificar estos grupos en primer lugar.

Las técnicas de análisis multivariante mencionadas, cuyo objetivo es reducir la dimensión del fenómeno analizado, pueden constituir un primer paso del análisis de los datos. En efecto, las nuevas variables obtenidas en este proceso de reducción, ejes, componentes, factores o dimensiones, por su mayor relevancia conceptual, permiten en posteriores aplicaciones sustituir a las primitivas variables originales sin apenas pérdida de información.

Esta idea, sugerida hace ya algunas décadas (Moser y Wolf, 1961), se está convirtiendo en práctica habitual del investigador que se enfrenta a grandes conjuntos de datos. Ejemplos de esta aplicación secuencial de técnicas multivariantes a diversas disciplinas pueden encontrarse en Batista y Estivill (1983), Daling y Tamura (1970), y Everitt, Gourlag y Kendall (1971).

En el ejemplo de la ocupación sectorial por comunidades, propuesto en los puntos anteriores, se han logrado determinar e interpretar dos factores que permiten caracterizar a las comunidades mediante dos únicas coordenadas en lugar de las cinco originales. Este análisis de correspondencias previo, no sólo ha clarificado la estructura de la matriz de datos facilitando la interpretación de las interrelaciones entre ocupaciones y comunidades, sino que además proporciona un marco adecuado para clasificar las comunidades exclusivamente en base a la información «esencial» que estos ejes suministran.

8. ETAPAS BASICAS DE UN ALGORITMO DE CLASIFICACION

A partir de la matriz base del análisis, sean variables originales o factores derivados, puede medirse la similitud entre cualquier par de individuos en función de los valores que aquéllos tomen sobre éstos. Para evaluar la similaridad o distancia entre individuos se dispone de una gran cantidad de índices que resumen en un solo valor la contribución de todas las variables; uno de los más utilizados es la distancia euclídea. La elección del índice dependerá de la naturaleza de los datos. La obtención de una matriz de similaridades o distancias supone el primer paso de todo proceso de clasificación.

La tabla VII corresponde a la matriz de distancias euclídeas entre comunidades obtenidas a partir de las puntuaciones factoriales.

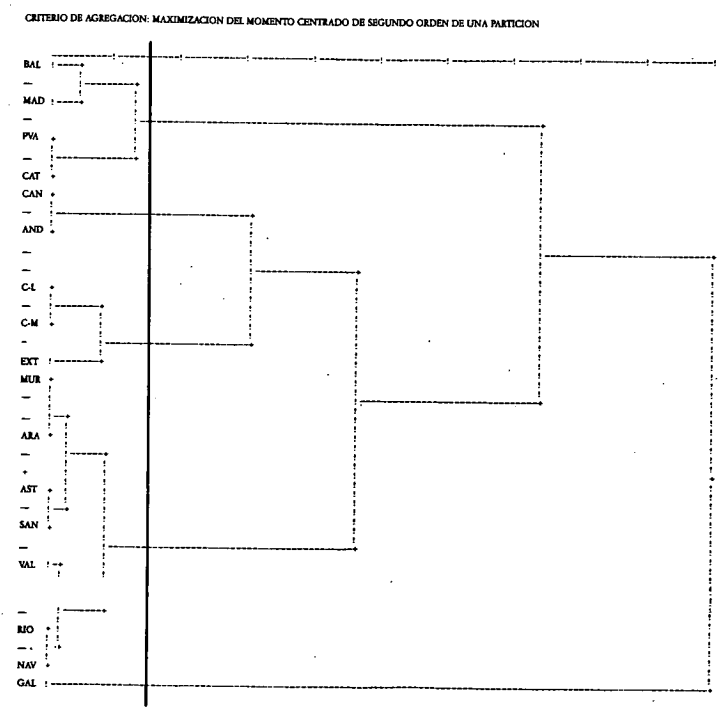
TABLA VII

Matriz de distancias

	AND	ARA	BAL	CAN	SAN	C-M	C-L	CAT	VAL	EXT	GAL	MAD	MUR	NAV	PVA	AST	RIO
AND																	
ARA	.33																
BAL	.24	.26															
CAN	.07	.40	.26														
SAN	.42	.12	.38	.49													
C-M	.42	.26	.48	.49	.18												
C-L	.34	.23	.41	.41	.18	.08											
CAT	.51	.34	.27	.53	.44	.60	.57										
VAL	.33	.14	.16	.37	.25	.39	.35	.22									
EXT	.31	.42	.49	.36	.41	.29	.23	.72	.50								
GAL	.82	.71	.92	.88	.61	.45	.51	1.05	.84	.54							
MAD	.39	.48	.23	.37	.60	.70	.64	.32	.35	.69	1.15						
MUR	.26	.08	.23	.33	.16	.25	.19	.38	.16	.35	.70	.46					
NAV	.41	.14	.25	.46	.23	.39	.37	.21	.09	.55	.84	.42	.20				
PVA	.50	.36	.26	.52	.46	.61	.58	.03	.23	.72	1.06	.29	.39	.23			
AST	.39	.09	.35	.46	.03	.19	.18	.42	.22	.40	.63	.57	.13	.21	.44		
RIO	.46	.14	.32	.52	.18	.35	.34	.28	.16	.55	.78	.50	.22	.08	.30	.17	
	AND	ARA	BAL	CAN	SAN	C-M	C-L	CAT	VAL	EXT	GAL	MAD	MUR	NAV	PVA	AST	RIO

A partir de la matriz de similaridades, un algoritmo de agregación seleccionado de entre las diversas familias (9) existentes proporcionará, salvo algunos algoritmos excepcionales, una estructura arborescente denominada dendrograma que representa gráficamente el proceso de agrupación. La figura 3 muestra el dendrograma correspondiente a nuestro ejemplo, en el cual puede observarse en el eje horizontal las distintas comunidades (individuos), y en el vertical los niveles de similaridad en los cuales se poducen las agregaciones; a mayor ordenada mayor distancia o menor similaridad entre los elementos que se agrupan de acuerdo al índice seleccionado. Esto permite determinar distintas agrupaciones en clases según el nivel de similaridad —ordenada del dendrograma— que se considere. Por ejemplo, analizando nuestro árbol de la figura 3, puede distinguirse en el nivel más bajo cómo los pares de comunidades más similares se agrupan (PVA-CAT, CAN-AND, CL-CM, MUR-ARA, AST-SAN y RIO-NAV). A este primer nivel de agregación se pasa de tener 17 individuos (comunidades) a disponer de esos 6 primeros grupos y 5 comunidades aisladas que configurarían una primera clasificación. Avanzando en este proceso de agregación, que supone por tanto admitir mayor heterogeneidad dentro de los grupos a cambio de reducir su número, se llegaría a una única clase que incluiría todas las comunidades. Obviamente, debe llegarse a un compromiso entre el número de clases a retener y la heterogeneidad dentro de las mismas. En este caso, se ha optado por detener este proceso de agregación al nivel señalado por la línea continua horizontal, que determina cinco clases. La composición de las mismas se ha superpuesto en el gráfico que los factores obtenidos en el AC determinan (ver figura 4).

FIGURA 3

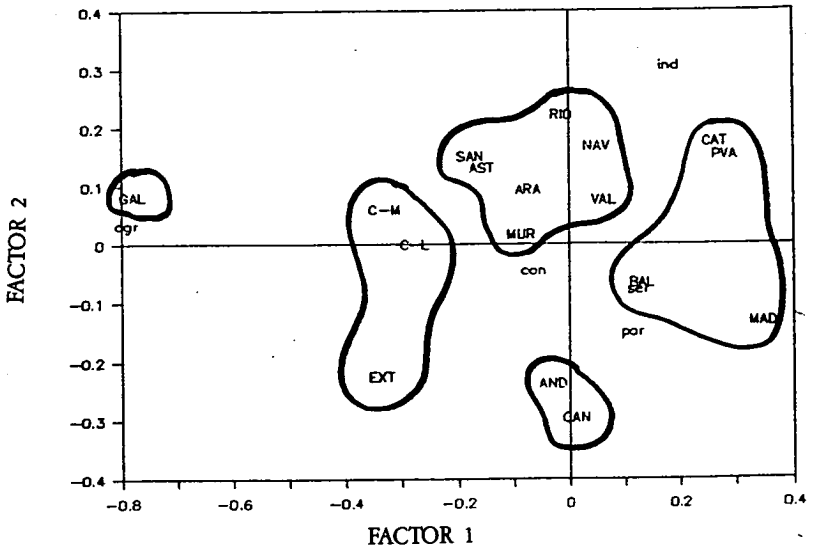


Dendrograma que representa las sucesivas agregaciones de las clases

Adicionalmente, estas clases pueden caracterizarse comparando los valores medios que toman las p variables dentro de cada grupo con el valor medio general.

FIGURA 4

Plano Factorial 1-2



Como se puede observar fácilmente, el resultado del AC previo ya sugiere y facilita el razonar sobre las proximidades que se revelan entre las comunidades. El análisis de clasificación sistematiza este proceso.

9. CONCLUSIONES

El AC resulta un instrumento sumamente potente para el análisis de la interrelación entre variables categóricas siempre que la matriz de datos cumpla los requisitos de homogeneidad y exhaustividad mencionados. Dado que el problema de la medida en las ciencias sociales y del comportamiento adquiere especial dificultad, estos métodos, que relajan los requisitos exigibles a los datos, abren un amplio espectro de posibilidades para el estudio de estos fenómenos.

La simplicidad de la interpretación de los resultados, incluso para no expertos, junto con la disponibilidad de paquetes estadísticos de fácil manejo, hacen de esta técnica un instrumento imprescindible para cualquier investigador.

Los métodos de clasificación tienen inherentes una serie de problemas, entre los que se pueden citar los siguientes:

1. No existe acuerdo universal acerca de lo que constituye un *cluster.*
2. Como técnica exploratoria, algunos de los algoritmos se han desarrollado en áreas específicas para resolver problemas concretos, aplicándose posteriormente en áreas diversas.
3. Si los datos no están de acuerdo con los supuestos del algoritmo de clasificación, éste puede imponer más que descubrir una estructura en los mismos. De hecho, distintos algoritmos pueden generar distintas clasificaciones para un mismo conjunto de datos.

No obstante, a pesar de estos problemas, estas técnicas dan respuesta válida, si se aplican de forma no automática, al problema de definición de tipologías. Por otro lado, su campo de aplicación no se reduce a complementar las técnicas de reducción de datos, como en el caso presentado, sino que tienen entidad propia y son susceptibles de ser aplicados a cualquier matriz de individuos/variables.

Anexo 1

EL METODO ESTADISTICO

Al establecer el objetivo del AC como el estudio de las relaciones existentes intra y entre filas y columnas de la tabla de contingencia, se explicitó la necesidad de obtener un espacio de dimensión reducida que ajustara de forma óptima la nube de puntos original de forma que pudieran observarse con claridad las relaciones existentes.

Sin duda esta aproximación al método recuerda otras técnicas de reducción de datos, tales como ACP y MDS. Sin embargo, la métrica aquí utilizada y definidá en (1) no es una suma de cuadrados, por lo que no coincide con la euclídea propia de aquellas técnicas de determinación de la dimensionalidad latente. En estas condiciones, la obtención del espacio de dimensión reducida no puede efectuarse somentiendo directamente a ACP las tablas de perfiles 2a y 2b, pero si estas tablas se transforman de forma que el punto-fila i-ésimo de R^J (o j-ésimo de R^I) sea el:

$$z_{ij} = \frac{1}{\sqrt{f_{\cdot j}}} \quad \frac{f_{ij}}{f_{i\cdot}} \qquad (2)$$

el problema se reduce a una aplicación del ACP sobre la matriz Z de perfiles transformados según (2), puesto que la distancia euclídea habitual entre perfiles transformados coincide con la distancia X^2 entre perfiles definida en (1).

$$d^2_{\times 2}(i, i') = d^2(z_i, z_{i'}) = \sum_{j=1}^{J} \left(\frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}} - \frac{f_{i'j}}{f_{i'\cdot}\sqrt{f_{\cdot j}}} \right) \qquad (1 \text{ bis})$$

De este modo, obtener los ejes principales se reduce a diagonalizar la matriz de varianzas-covarianzas S entre perfiles transformados, de idéntica forma que se hace en ACP, siendo

$$S = Z'D_I Z - MM' \qquad (3)$$

y $M = (\sqrt{f_{\cdot 1}}, \sqrt{f_{\cdot 2}}, \ldots, \sqrt{f_{\cdot j}})'$ el individuo (autonomía) medio.

$$D_I = \text{diag}\,(f_1., f_2., \ldots, f_I)$$

Análogamente se resolvería el problema dual consistente en obtener los ejes principales de R^I. Sin embargo, las relaciones, conocidas como baricéntricas, que se observan entre los valores y vectores propios de ambos procesos de diagonalización hacen innecesario repetir el proceso en R^I (Lebart, Morineau y Warwirk, 1984). Debido a ello es suficiente resolver el problema de la representación en uno de los dos espacios, el de menor dimensión (filas o columnas), y determinar mediante las relaciones baricéntricas las proyecciones de la otra nube de puntos.

A pesar de lo que en común comparten las técnicas de ACP y de AC, pues ambas representan perspectivas distintas del análisis estadístico-matemático cuyo objetivo es determinar los ejes principales —de máxima variabilidad— para representar los datos de la matriz de partida en términos de un número de dimensiones reducidas, el AC se interesa en igual medida por los elementos de las filas como por los de las columnas de la matriz de datos, mientras que el objetivo del ACP se centra fundamentalmente en las relaciones entre las columnas-variables de la matriz de datos. Además, por otro lado, su utilización obedece a circunstancias bien distintas; así, se someterán a ACP aquellas matrices de datos cuyas columnas incluyan variables medidas en escalas de intervalo o de razón; por el contrario, el AC es adecuado para tablas de contingencia en las que se cruzan variables de tipo nominal.

Anexo 2

ANALISIS DE CORRESPONDENCIAS MULTIPLES

Si en lugar de considerar la doble partición efectuada en la población, propia del AC, en base a sólo dos criterios cualitativos (sector ocupacional y comunidad autónoma), se consideraran p caracteres, la metodología anterior es susceptible de generalizarse para el estudio de la relación inter e intra estas p variables.

Un caso particular importante de aplicación del ACM es el análisis de encuestas o entrevistas mediante cuestionario (véase J. Palacios, en este mismo número), donde cada pregunta se refiere a un carácter. Las respuestas a las preguntas de un cuestionario se presentan en forma completamente disyuntiva, ya que los individuos pueden situarse en una sola categoría de respuesta, lo cual las hace mutuamente exclusivas. Este proceso supone que cada pregunta efectúa una partición en la población (encuestados) en tantos grupos como categorías se incluyan. Si a cada pregunta le asociamos un conjunto de variables indicativas, una para cada modalidad, la tabla de datos, X, que será disjunta, constará de N filas y tantas columnas como categorías en total incluyan las p variables pregunta.

$$X = \begin{matrix} \\ (i) \end{matrix} \left[\begin{array}{c|c|c|c|c} 0\,1\,0 & 0\,0\,0\,1 & \cdots & \cdots & 0\,0\,1\,0\,0 \\ X_1 & X_2 & & & X_p \end{array}\right]$$

Pregunta X_1 tiene 3 modalidades de respuesta.
Pregunta X_2 tiene 4 modalidades de respuesta.

El método de AC que se ha presentado para el estudio conjunto de dos variables puede extenderse a este caso con p caracteres mediante la generalización del análisis canónico a p variables indicatrices (Carrol, 1968), ya que el AC —diagonalización— efectuado sobre la matriz X es equivalente al que se efectúa sobre la tabla B de Burt generada por ésta:

$$B = X' X$$

Una encuesta sobre los conocimientos de los padres acerca del desarrollo y educación de sus hijos (véase J. Palacios, en este mismo número) ha ilustrado algunos de los conceptos del ACM; otros serían repetición de los ya considerados en el AC simple, por lo cual se hace caso omiso de ellos. Sin embargo, en este caso, conviene poner el énfasis en la distinción entre elementos activos e ilustrativos o suplementarios, aspecto tratado sólo marginalmente en el AC. En muchas ocasiones conviene considerar activas aquellas variables que describen objetivamente al individuo (por ejemplo, las sociodemográficas) y declarar como ilustrativas —sin intervenir, por lo tanto, en el cálculo de los ejes y situadas posteriormente como «baricentro» de los individuos que caracterizan- aquellas que constituyen la parte substantiva de la encuesta y que se desea relacionar con las primeras pero no necesariamente entre ellas. Este proceso redunda en grandes ventajas sobre el análisis de tablas cruzadas y optimiza en gran medida el tiempo de cálculo.

Por último, mención aparte merece el caso especial en el que todas las preguntas del cuestionario tienen sólo dos categorías. En esta situación el análisis se simplifica, bien reduciéndolo a un ACP de las preguntas caracterizadas por una sola de sus categorías, bien analizando sólo una submatriz de la tabla de Burt (Lebart *et al.*, 1984).

Notas

[1] Contrariamente a la creencia popular, los métodos multivariantes fueron sugeridos hace mucho tiempo (Hotteling, 1933, Pearson, 1901, Spearman, 1904). De hecho, desde R. Fisher, se ha escrito relativamente poco verdaderamente novedoso. Análogamente a lo que ha sucedido en Medicina con la cirugía, en donde la sensación de avance espectacular se debe principalmente al avance tecnológico, en Estadística ha sido la utilización de ordenadores, al eliminar la dificultad de cálculo, quien ha permitido operativizar y desarrollar aquellas ideas pioneras.

[2] La recodificación de variables continuas está justificada cuando coexistan en el análisis otras de naturaleza no continua, lo que impide la utilización de la técnica adecuada de Análisis en Componentes Principales.

[3] Los datos se han obtenido del Anuario EL PAIS 1986.

[4] Se entiende por modalidad cada uno de los valores o categorías que puede tomar una variable. En nuestro ejemplo la variable Comunidad Autónoma incluye las modalidades: Andalucía, Aragón, Baleares, etc.; mientras que la variable Sector de Ocupación adopta las de Agricultura, Industria, Construcción, Servicios y Paro.

[5] El término de marginales se refiere aquí al clásico concepto estadístico de frecuencia absoluta marginal, es decir, el número total de observaciones de cada modalidad de una variable. En nuestro ejemplo, los marginales de las Comunidades representan el total de población activa de cada una de ellas. Así, un marginal elevado de una comunidad indica un alto nivel de población activa.

[6] Esta distancia recibe el nombre de CHI-CUADRADO debido a que su expresión coincide con la de la prueba del mismo nombre que tradicionalmente se ha utilizado para comprobar la dependencia estocástica entre variables cualitativas.

[7] De hecho los paquetes de programas estadísticos multivariantes desarrollados por la escuela francesa (SPAD, etc.) permiten, en cualquier técnica de reducción de datos, la simultaneidad de ambas representaciones. Por el contrario, los desarrollados por la escuela anglosajona, principalmente SPSS y BMDP, más preocupados por reducir el espacio de las variables, sólo recientemente han incorporado en sus programas la posibilidad de representar directamente en el espacio de los componentes o factores también a los individuos. Estos últimos presuponen que los individuos constituirán generalmente muestras aleatorias, sin interés por sí mismos.

[8] Debido a la naturaleza de la matriz base del análisis (matriz de perfiles), donde cada fila corresponde a la idea de distribución condicionada de frecuencias y por tanto su suma es constante, existirá sistemáticamente una combinación lineal que hace que el número máximo de valores propios no nulos sea el menor de (I-1) y (J-1).

[9] En la actualidad pueden distinguirse siete grandes familias que representan distintas perspectivas de clasificación: métodos jerárquicos —aglomerativos y divisivos—, métodos de participaciones iterativas, métodos de búsqueda de zonas densas, analítico-factoriales, métodos «clumping» y métodos basados en la teoría de grafos (Gordon, 1971).

Resumen

Este artículo introduce en la utilización de dos procedimientos de análisis de los datos (análisis de correspondencias y técnicas de clasificación), que, aunque tienen sentido por sí mismos, pueden ser utilizados combinadamente en la investigación en ciencias sociales y del comportamiento. Se discuten los orígenes de estas técnicas, su lógica interna y su ámbito de aplicación, presentándose toda su argumentación alrededor de un ejemplo concreto.

Summary

This paper is an introduction to the use of two forms of data analysis (correspondence analysis and cluster analysis), that, although useful in themselves, can be fruitfully used together in research both in social and behavioral sciences. The origins of these procedures, their internal rationals and their domains of application are discussed. Discussion has been referred to a specific example.

Referencias

ALDENDERFER, M. S. y BLASHFIELD, R. K. (1984). *Cluster analysis.* Beverly Hills: Sage.

BATISTA, J. M. y ESTIVILL, X. (1983). Aplicació de l'analisi de components principals a nivell de dades municipals: definició de zones homogenees y estudi de la jerarquía urbana a Catalunya. Barcelona: *Actas del First International Symposium on Statistics.*

BENZECRI, J. P. (1973). *L'analyse des données. Tome I: La Taxonomie. Tome II: L'analyse des correspondences.* París: Dunod.

BENZECRI, J. P. (1977). Histoire et prehistoire de l'analyse des données. *Les cahiers de l'analise des données, 2,* 9-53.

CARROLL, J. D. (1968). Generalization of canonical correlation to three or more sets of variables. *Proc. Amer. Psych. Assoc.,* 227-228.

DALING, J. y TAMURA, H. T. (1970). Use of ortogonal factors for selection of variables in a regresion equation. *J. R. S. S.,* 260-268.

EVERITT, B. (1980). *Cluster analysis.* Aldershot: Gower.

EVERITT, B., GOURLAY, A. J. y KENDALL, R. E. (1971). An attemptat validation of traditional psychiatric syndromes by cluster analysis. *British Journal of Psychiatry, 119,* 299-412.

GORDON, A. D. (1981). *Classification.* London: Chapman and Hall.

HOTTELING, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24,* 417-441y 498-520.

JORESKOG, K. G. (1969). A general approach to confirmatory factor analysis. *Psychometrika, 34,* 183-202.

LEBART, L., MORINEAU, A. y WARWICK, K. M. (1984). *Multivariate descriptive statistical analysis.* New York: Wiley.

MOSER, C. A. y WOLF, S. (1961). *Bristish towns: a statistical study of their social and economical differences.* Center for Urban Studies (Oliver and Boy, Ltd).

PEARSON, K. (1901). On lines and planes of closest fit to a system of points in space. *Phil. Mag., 2,* 151-180.

SÁNCHEZ CARRIÓN, J. J. (Ed.). *Introducción al análisis multivariante aplicado a las ciencias sociales.* Madrid: CIS-Siglo XXI.

SPEARMAN, C. H. (1904). General intelligence objetively determinen and measured. *American Journal Psychology, 15,* 201-293.

UPTON, C. J. G. (1978).*The analysis of cross-tabulated data.* New York: Wiley.