EL ANÁLISIS DE CLUSTER: APLICACIÓN, INTERPRETACIÓN Y VALIDACIÓN

Óscar Fernández Santana
Responsable del Área Informático-estadística.
Gabinete de Prospección Sociológica. Presidencia del Gobierno vasco

Resumen

El Análisis de Cluster, junto con el Análisis de Componentes Principales, es un método de gran interés en Sociología dentro de la perspectiva que se puede llamar de descripciones densas. Además su aplicación es más fecunda cuando se realiza complementariamente al Análisis de Componentes. En el artículo se da cuenta del proceso y del diseño de su realización sin entrar en los aspectos matemáticos. Así, se trata en él de la preparación del método, de los procedimientos y criterios de elección del número de clusters y de las formas de validación del mismo. La insistencia en la validación proviene de que aún siendo un método fecundo de agregación, es muy sensible, sin embargo, a las variaciones que puede tener su preparación y a los criterios de selección de grupos.

Resum

L'Anàlisi de Cluster, conjuntament amb la d'Anàlisi de Components Principals, és un mètode de gran interès en Sociologia dins de la perspectiva que pot ser anomenada com a descripcions denses. A més la seva aplicació és més fruitosa quan es realitza de manera complementària a l'Anàlisi de Components. A l'article s'explica el procés i el disseny de la seva realització sense entrar en els aspectes matemàtics. Així doncs, es parla de la preparació del mètode, dels procediments i criteris d'elecció del nombre de clusters i de les formes corresponents de validació. La insistència en la validació prové de que tot i sient un mètode fruitós d'agregació, és també molt sensible a les variacions que pugui tenir la seva preparació i els criteris de selecció de grups.

Abstract

The cluster analysis, together with the Principal Components Analysis constitutes an interesting method for Sociology from what we call the dense description perspective. Its application it is more productive when used as a complement to Principal Components Analysis. The article describes the process and the design of its realization, without mentioning the mathematical aspects. Then, it goes to deal with the preparations of the method, the procedures and

the selection criteria of the number of clusters and their different means of validation. The emphasis on validation is originated in its feature of being responsive to the variations in its preparations and the group selection criteria, althoug it is a productive method.

MARCO GENERAL

El Análisis de Cluster (AC, en adelante) es el nombre genérico otorgado a una gran variedad de técnicas que tienen como *objetivo* primordial la búsqueda de grupos en un conjunto de individuos. En líneas generales, todo método de clasificación parte de un conjunto de elementos singulares que deben ser clasificados en un número reducido de grupos o «clusters», obtenidos por particiones sucesivas del conjunto original y en los que se respete la estructura relacional que en el mismo se matenía. Las leyes matemáticas por las que se rigen estos métodos reciben el nombre de «Taxonomía Numérica»[1].

Este concepto, algo confuso, puede quedar aclarado al delimitar las *propiedades* de los clusters:

1. Densidad: esta primera propiedad define un cluster como un conglomerado espacial de puntos relativamente compacto en comparación con otras áreas de ese espacio que tienen menos o ningún punto.

2. Varianza: grado de dispersión de los puntos de cada conglomerado en el espacio.

3. Forma: configuración espacial de los puntos (redondeada-hiperesfera-alargada, etc.).

4. Separación: grado de solapamiento o de separación entre los clusters.

En los últimos años, el AC ha ido ganando popularidad, en parte porque su objetivo es muy apetecido y fácil de entender, y en parte porque siempre se consigue un resultado interpretable. Es innegable la enorme importancia que tiene el descubrimiento de tipologías en el ámbito de las ciencias sociales y, especialmente, en el marketing. El problema principal de su utilización

1. Sokal, R.R. & Sneath, P.H. *Principles of Numerical Taxonomy*. W.H. Freeman and Co., San Francisco, 1963. En general, la taxonomía numérica es un conjunto de leyes formalizadas que intentan construir clasificaciones basadas en la semejanza observada entre los elementos a clasificar.

radica en el amplio espectro de *controversias* relativas a sus modos de aplicación. En efecto, los postulados teóricos y metodológicos sobre los que se asienta el AC no están tan sólidamente fundamentados como, por ejemplo, los del Análisis Factorial.

Desde una perspectiva eminentemente operativa (nuestro objetivo no es la profundización matemática en el AC), ofrecemos un esbozo de «guía» de las posibilidades más ventajosas en la aplicación, interpretación y validación de los resultados de todo AC.

PRÁCTICA DEL ANÁLISIS DE CLUSTER

Se expone a continuación un esbozo de lo que podría ser una guía para preparar, realizar y evaluar un AC. Desde una perspectiva diacrónica, las tres fases sugeridas son precisamente éstas:

1. Preparación del AC
2. Realización del AC
3. Evaluación del AC

Preparación

En el momento de «diseñar» la aplicación de un AC nos encontramos con tres aspectos que han de ser necesariamente considerados: la matriz de datos, la medida de distancia y el método de clusterización. Las decisiones adoptadas en esta fase determinarán en gran medida las configuraciones resultantes.

Matriz de datos

Dentro de esta matriz inicial se incluyen los individuos a agrupar y la/s variable/s en función de las cuales se llevará a cabo la clasificación.

1. Individuos: En cuanto a la muestra investigada, habrá que señalar su volumen y procedencia, y si se emplean o no todos los sujetos para realizar el AC[2]. Asimismo, se indicará si los casos analizados han sido ponderados y, si lo han sido, en base a qué criterios se ha efectuado la ponderación.

2. Ya veremos más adelante la utilidad de realizar varios AC para diferentes submuestras de una misma población.

2. Variables. Las variables han de ser continuas[3] y su número no muy elevado; el investigador no experimentado suele tender a introducir en el AC una gran cantidad de variables que a la postre no hacen sino perjudicar la interpretabilidad de los resultados.

Además, las variables deben ser comparables entre sí, tanto en su lógica como en sus escalas. Un caso extremo de esta violación sería incluir datos sociodemográficos juntamente con opiniones u otros ítems de naturaleza psicológica. Hemos de tener claro que un conjunto relativamente pequeño de ítems (variables) comparables e igualmente escalados, todos relativos a un mismo tema, proporciona las mejores condiciones para obtener un resultado cluster que tenga sentido.

De antemano hay que valorar la magnitud del peso que se da a los diferentes aspectos dentro de la batería de ítems. Si hay varias variables midiendo la misma o parecida dimensión, entonces este aspecto recibe una oportunidad correspondientemente mayor de tener un efecto en el proceso de agrupamiento que si sólo hubiera un ítem de ese tipo. Aquí radica el peligro de predestinar el resultado.

También hay que plantearse la conveniencia o no de realizar algún tipo de transformación de las variables, tales como la estandarización (normalización) o la factorización (realización de un análisis factorial de las variables analizadas y utilización posterior de las puntuaciones score de los individuos en los factores resultantes).

Medida de similitud

Uno de los factores que ocasionan mayores divergencias en los resultados es la elección de la medida de similitud (destinada a cuantificar la separación entre las unidades de análisis). Las razones para decidir emplear una u otra merecerían cuando menos otro artículo, y rebasarían los objetivos operativos de estas páginas. Las más utilizadas son las medidas de distancia, a pesar de que también existen los coeficientes de correlación, las medidas de asociación y los coeficientes de similitud probabilística[4].

Uno de los aspectos que en última instancia son más importantes en este tipo de decisiones es la disponibilidad de un paquete de programas estadísticos u otro. En efecto, hay una gran diversidad de programas informáticos

3. Algunos métodos de clusterización utilizan variables dicotómicas, pero son los métodos menos populares. Por ello, indicamos que, en principio, las variables serán de intervalo o continuas.

4. Véase Aldenderfer, Mark S.; Blashfield, Roger K. *Cluster analysis.* Sage University Papers. Beverly Hills, 1984.

que realizan análisis de clusters. Desde los paquetes estadísticos generales más conocidos (SPSS, SPADN, BMDP, SAS, etc.) hasta los paquetes específicos para esta técnica (CLUSTAN, BCTRY, CLUS, NTSYS, etc.). La última versión del SPSS/PC (V3.0), por ejemplo, incluye las distancias euclídea, euclídea al cuadrado, Manhattan (City-block), Chebychev, Minkowski y Cosine. Sin embargo, la versión para microordenadores del SPADN tan sólo emplea la distancia euclídea.

Método de clusterización

Lo mismo sucede en el apartado relativo al método elegido para realizar el AC. Su gran proliferación impide detenerse en cada uno de ellos. Valga con señalar que existen siete grandes familias:

— Jerárquicos aglomerativos: distancias mínimas, distancias máximas, distancias entre centroides, distancias ponderadas, Ward
— Jerárquicos divisivos: monotéticos o politéticos
— Partición iterativa: centros móviles, K-means, Hill-Climbing, Isodata
— Búsqueda de densidad: NORMIX, NORMAP
— Análisis factorial «Q»
— Clumping
— Métodos basados en la teoría de los grafos

Las razones de su selección se deben a consideraciones teóricas y prácticas. Entre las segundas podemos señalar nuevamente la concerniente a los programas estadísticos. Así, el SPSS/PC dispone de los siguientes métodos:

— Jerárquicos aglomerativos: distancias mínimas, distancias máximas, distancias entre centroides (interclusters, intraclusters, mediana, centroide) y Ward
— Iterativos: K-means

El SPADN (V1.0, 1985, para microordenadores) emplea un método mixto compuesto por la combinación progresiva de otros dos:

— Jerárquico aglomerativo: Ward (módulo RECIP)
— Iterativo: Centros móviles (módulo PARTI). Tras seleccionar una partición (un número apropiado de clases —grupos, clusters—, al inspeccionar el dendrograma formado en la fase precedente: RECIP), se efectúa una clasificación no jerárquica en la que se somete la partición de base a una serie de iteraciones tendentes a su consolidación, mediante el algoritmo de centros móviles.

Por otra parte, es importante aclarar, para evitar posibles confusiones, que, en Francia, las técnicas de AC son conocidas como Técnicas de Clasificación (o técnicas de clasificación automática) y, más específicamente, los métodos jerárquicos aglomerativos se identifican con el término «Clasificación Ascendente Jerárquica». Además los conglomerados de individuos resultantes son llamados «Clases» en vez de clusters o grupos.

Aplicación

Una vez que hemos justificado en base a qué criterios se ha preparado la matriz de datos y elegido la medida de similitud y el método de clusterización, pasamos a la operacionalización del AC. En esta fase hemos de dar solución a un problema de vital importancia: ¿con cuántos clusters nos quedamos?, esto es, ¿cuál es la partición más adecuada?[5] La respuesta dependerá de uno o varios criterios, que nosotros hemos agrupado en 5 factores:

— % de varianza explicada por cada partición
— Distancias inter e intraclusters
— Distribución de los sujetos en los clusters
— Distribución de los clusters en las variables «activas» y «pasivas»[6].

Varianza explicada por cada partición

Uno de los criterios más decisivos a la hora de establecer cuál es la solución cluster más apropiada a nuestros datos es el de los «saltos de varianza». Dado que cada partición lleva consigo un cierto tanto por ciento de varianza explicada, seleccionaremos aquella en la que su valor sea «significativamente mayor» que la fase con un cluster menos y donde el incremento del número de clusters no proporcione una mejora digna de tener en cuenta[7].

Generalmente, se suelen emplear tácticas gráficas semejantes a la técnica «Scree test» de selección del número de factores en un Análisis Factorial[8]. El procedimiento consiste en situar en el eje vertical de un plano las diversas

5. Una «partición» es cada división de los individuos en un cierto número de grupos. Así, se habla de particiones en 3 clusters, en 5 clusters, etc.

6. Llamaremos variables «activas» a aquellas que intervienen propiamente en el AC para formar los grupos, y variables «pasivas» a aquellas que no cumplen esta condición.

7. En todo caso, algunos autores han desarrollado algunos tests formales, como el «Stopping rule #1», de R. Mojena (1977) y R. Mojena y D. Wishart (1980); el «Likelihood ratio test», de J.H. Wolfe (1971), etc.

8. Véase Óscar Fernández. «Comprensión y manejo del análisis factorial», en *Revista Internacional de Sociología* (aún no publicado).

particiones a elegir y, en el eje horizontal, indicar la cantidad de varianza explicada por cada solución[9]. La línea formada al unir las parejas de puntos resultantes irá ascendiendo progresivamente (cuanto mayor sea el número de clusters mayor será la varianza explicada por la partición). Pues bien, el «truco» está en elegir la partición a partir de la cual la línea tiende a aplanarse.

Antes de pasar a otro punto, conviene aclarar que el «% de varianza explicada por una partición» es igual a la «varianza total» (100%) menos el valor resultante de sumar las «inercias intragrupo», dividir ese sumatorio entre la inercia total y multiplicar esa cantidad por 100, o, lo que es lo mismo, igual a la «inercia intergrupos» dividida por la «inercia total»[10, 11].

Distancias

La consideración de las distancias nos brinda varias posibilidades:

1. Distancia entre los puntos más lejanos de los clusters que están siendo combinados en cada partición[12]: si este mayor es significativamente mayor

9. Ciertos autores, por ejemplo P. Dunn-Ranking, sitúan —al utilizar el método de Ward— las sumas de cuadrados (entre todos los posibles pares de sujetos) en el eje horizontal, en vez del % de varianza explicada, pero no es un criterio muy común. Véase Peter Dunn-Rankin. *Scaling methods.* Lawrence Erlbaum Associates Publishers. Nueva Jersey, 1983, p. 139.

10. Es necesario precisar aún más estos conceptos:
— *Inercia intragrupo*: distancia entre los puntos (individuos) de un grupo y el centro de ese grupo.
— *Inercia intergrupos*: distancia de todos los centros de todos los grupos entre sí.
— *Inercia total*: suma de las inercias intragrupos y la inercia intergrupos.
Partiendo del principio general del AC (minimización de las variaciones internas de los clusters y maximización de la separación entre los mismos), nuestro objetivo siempre será que las inercias intragrupo sean las menores posibles y que la inercia intergrupos sea, por consiguiente, mayor. En una partición en la que hubiera tantos clusters como individuos se daría una «intra» = 0 y una «inter» = 100% de la inercia total; y viceversa en una solución de 1 sólo cluster. Así, a medida que vamos disminuyendo el número de clusters va aumentando la «intra» y reduciéndose la «inter» en un grado proporcional.

11. En el SPADN (V1.0) hemos de realizar las siguientes operaciones para calcular la varianza explicada, según el método empleado:
— Método Ward (RECIP):
Varianza = 100-(100 (Índice/∑Índices))
— Método de Centros Móviles (PARTI):
Varianza = 100-(100 (∑Inercias «intra»/Inercia total))
o, lo que es lo mismo:
Varianza = Inercia «inter»/Inercia total

12. Véase M.J. Norusis. *Advanced Statistics in SPSS/PC+*. SPSS, Inc. Chicago, 1986, p. B-77.

que su inmediato anterior, entonces se están combinando clusters con individuos muy separados. Este es un criterio no muy sólido y que no debe ser, en caso de utilizarse, el único.

2. Distancias intergrupos: si realizamos un análisis multidimensional no métrico (escalograma multidimensional) sobre la matriz de distancias interclusters, podremos obtener la configuración espacial de la partición en cuestión. En palabras de F. Alcantud, «un buen criterio para determinar cuál es la partición más adecuada, sería proseguir las particiones hasta que en el espacio definido por las dos dimensiones del escalograma aparecieran dos clusters muy próximos o incluso superpuestos»[13].

3. Distancias intergrupos y distancias intragrupos: una buena partición sería aquella que, siendo de rango más elevado, cumpliera la condición de que las distancias promedio intraclusters sean en todos los grupos menores que las distancias interclusters.

Distribución de los sujetos en los clusters

Si se examinan detenidamente los individuos de los que están compuestos los clusters resultantes en cada partición, podremos observar que o bien se redistribuyen totalmente en los grupos o bien se subdividen ciertos grupos para dar lugar a otros nuevos grupos (permaneciendo invariables algunos de los clusters anteriores). Aclaremos esto con un ejemplo artificial: si tenemos 200 individuos y en la primera partición en dos clusters se han situado los 100 primeros en el cluster 1 y los 100 últimos en el cluster 2, al realizar una partición en tres clusters, los sujetos podrían «redistribuirse» nuevamente (individuos 1-60 al cluster 1, 61-130 al cluster 2 y 131-200 al cluster 3) o «dividirse» (individuos 1-100 siguen en el cluster 1, 101-140 al cluster 2 y 141-200 al cluster 3).

En principio, aunque no taxativamente, cabe deducir que, si la partición se obtiene por división, los perfiles de los nuevos clusters serán más o menos parecidos, diferenciándose en aspectos muy concretos y quizás difíciles de interpretar. Por lo tanto, y según este criterio —que es discutible—, habría que elegir la partición anterior a aquella en la que los individuos se distribuyen por división y no por redistribución.

Distribución de los clusters en las variables

Un buen criterio para determinar la utilidad práctica de una partición es examinar las distribuciones de cada grupo de individuos resultante en cada

una de las variables utilizadas para su formación[14]. Este examen puede mantenerse en un nivel descriptivo o acudir a tests estadísticos de significación. Entre estos últimos destaca el Análisis de Varianza, que puede ser univariado (ONEWAY; nos dice si las medidas de los grupos en una variable son o no son significativamente diferentes) o multivariado (MANOVA; ídem pero con todas las variables a la vez). Estos tests revisten el peligro de ser casi siempre significativos, incluso en el caso de que no haya clusters «naturales» en los datos (distribución normal de los datos)[15].

Otra posibilidad que adolece del mismo problema consiste en realizar un Análisis Factorial Discriminante utilizando, para cada partición a elegir, la asignación de cada sujeto al cluster como la variable «a precedir» y las variables del AC como variables «predictoras».

Todos estos criterios están encaminados a valorar la bondad de cada partición en base a las variables manejadas en el AC (variables activas). También es posible, e incluso conveniente, tratar de identificar y conocer lo característico de cada cluster de acuerdo con el resto de las variables disponibles (variables pasivas). La confrontación de los grupos obtenidos con alguna otra variable podría afirmar o negar totalmente la utilidad práctica de cualquier partición.

Validación

Por último, después de decidir cuál es la mejor partición de los datos iniciales, nos queda la tarea de validar esta decisión. Esta labor puede satisfacerse por alguno de estos cinco criterios:

— Coeficiente de correlación cophenético
— Coeficiente de pertenencia
— Replicación
— Simulaciones Monte Carlo
— Interpretabilidad teórico-práctica

14. Este criterio suele ser incluido dentro de las estrategias de validación de la partición —analizadas posteriormente—, pero nosotros estimamos más esclarecedor considerarlo bajo este epígrafe.

15. Si, por ejemplo, se efectúa un AC a los miembros de una clase en virtud de sus puntuaciones en un test de inteligencia, pudiera darse el caso de que la muestra analizada se distribuyera normalmente alrededor de la media y no existieran clusters propiamente hablando. A pesar de ello, al realizar el AC podríamos encontrarnos con dos grupos (los que tienen bajas puntuaciones y los que las tienen altas) y, al realizar el análisis de varianza los clusters resultarían ser diferentes con un elevado nivel de confianza en la variable de partida. Sobre este punto, véase Mark S. Aldenderfer y Roger K. Blashfield, op. cit., 1984, p. 65.

Coeficiente de correlación cophenético

Este coeficiente es válido sólo para los métodos jerárquicos aglomerativos. Representa una medida de ajuste entre los datos de partida y la estructura del dendrograma, esto es, el grado en que la partición reproduce el armazón de distancias entre los individuos. En líneas generales, se suele decir que una buena partición debe tener una correlación cophenética de, al menos, .85[16].

Sin embargo, a pesar de ser un indicador frecuentemente manejado de la validez de la partición, también es objeto de críticas de corte matemático y que no vamos a exponer aquí[17].

Coeficiente de pertenencia

Esta es una medida para ver cuán diferentes son los conglomerados, en función de los ítems contenidos en cada uno de ellos. A tal fin se calcula un cociente cuyo numerador expresa la media de la intercorrelación entre los sujetos dentro de un mismo cluster, y cuyo denominador es igual a la media de la intercorrelación de pares de ítems —en donde un ítem en cada par pertenece al grupo de interés.

Si el conglomerado está bien elegido, el numerador será superior a la unidad, sugiriéndose que para un valor del coeficiente igual o superior a 1.3, se puede considerar que un cluster ha sido identificado. Con el manejo de estos coeficientes (también llamados «coeficientes B») no es necesario recurrir a la visualización gráfica de los resultados del análisis, ya que los coeficientes B más elevados representarán los conglomerados más significativos[18].

Replicación

La táctica de la replicación consiste en repetir el AC para diferentes submuestras (dos o más) de la población total, a fin de ver si las particiones resultantes mantienen un cierto nivel de consistencia interna. El fallo de este criterio es que sólo cuando los resultados son muy diferentes cabe sospechar algún problema en la partición original: su «éxito» no es un indicador claro de que la solución cluster establecida sea la más apropiada.

16. J.J. Sánchez Carrión. *Introducción a las técnicas de análisis multivariable aplicadas a las ciencias sociales*. CIS. Madrid, 1984, p. 138.

17. Si se quiere insistir en el tema, se puede consultar la obra de Mark S. Aldenderfer y Roger K. Blashfiels, op. cit., 1984, pp. 63-64.

18. Véase M. García Ferrando. *Socioestadística*. Alianza Universidad. Madrid, 1985, p. 454.

Simulaciones «Monte Carlo»

Esta aproximación es raramente utilizada a causa de su costo y su complejidad. Los procedimientos Monte Carlo se basan en «generadores de números aleatorios» (*random number generators*), para crear una nueva matriz de datos cuyas características generales queden apareadas con las características generales de la matriz original de datos, pero sin contener ningún cluster. Entonces se lleva a cabo un AC idéntico al realizado con los datos originales y se comparan los resultados de los dos AC. Esta comparación, que puede fundamentarse, por ejemplo, en realizar Análisis de Varianza para cada solución (la de la matriz inicial y la de la matriz «duplicada»), tendrá como fruto la validación o invalidación de la partición analizada.

Interpretabilidad teórico-práctica

En último término, todos los criterios de selección de la partición idónea y de su posterior validación carecen de sentido si la clasificación no tiene sentido teórico o es difícilmente interpretable. Es por ello que consideramos este criterio como el que, en última instancia, supera a todos los anteriores.

CONCLUSIONES

Las técnicas de AC tienen como *objetivo* principal la clasificación de un cierto número de individuos[19] en unos cuantos conglomerados o clusters. El principio general que guía la formación de los grupos es la minimización de la variación interna y la maximización de las distancias entre los clusters.

La partición obtenida depende en gran proporción de la medida de similitud empleada para conocer la separación entre los grupos y del método elegido para la formación de tales grupos.

Entre las *medidas de similitud* más utilizadas figuran los coeficientes de correlación, las medidas de distancia, los coeficientes de asociación y los coeficientes de similitud probabilística.

En cuanto a los *métodos de clusterización* cabe destacar siete grandes familias: Jerárquicos aglomerativos, Jerárquicos divisivos, Partición iterativa, Búsqueda de densidad, Análisis factorial de tipo «Q», Clumping y métodos basados en la teoría de los Grafos.

El *diseño e interpretación* de un AC puede llevarse a cabo según un proceso que comprende las siguientes *fases*:

19. También se puede hacer un AC de variables pero no ha sido ese nuestro objetivo de estudio.

1. Preparación: en esta primera etapa se incluyen una serie de puntos relativos a la matriz de datos, la medida de similitud y el método de clusterización.

2. Aplicación: la segunda fase tiene como fin primordial elegir la partición idónea, de acuerdo a alguno de estos siete criterios: % de varianza explicada, distancias, distribución de los sujetos en los clusters y distribución de los clusters en las variables.

3. Validación: evaluación de la solución elegida en virtud de varios indicadores: coeficiente de correlación cophenético, coeficiente de pertenencia, replicación, simulaciones Monte Carlo e interpretabilidad teórico-práctica.

Como conclusión general, podemos finalizar afirmando que el AC es una técnica muy útil en su resultado, la formación de una serie de grupos de individuos con respecto a un conjunto de variables; pero muy peligrosa y sensible a las variaciones en su preparación, aplicación y validación.

BIBLIOGRAFÍA

Aldenderfer, M.S.; Blashfield, Roger K. *Cluster analysis.* Sage University Paper. Beverly Hills, 1984.

Dunn-Rankin, P. *Scaling Methods.* Lawrence Erlbaum Associates Publishers. Nueva Jersey, 1983.

Fernández Santana, J.O. «Comprensión y manejo del análisis factorial», en *Revista Internacional de Sociologia* (aún no publicado).

García Ferrando, M. *Socioestadística.* Alianza Universidad. Madrid, 1985.

Mojena, R. «Hierarchical grouping methods and stopping rules —an evaluation», en *Computer Journal,* 20, 1977.

Mojena, R.; Wishart, D. «Stopping Rules for Ward's Clustering Method», en *Proceedings of COMPSTAT 1980.* Physika-Verlag. Würzburg, RFA, 1980.

Norusis, M.J. *Advanced Statistics in SPSS/PC+.* SPSS, Inc. Chicago, 1986.

Sánchez-Carrión, J.J. *Introducción a las técnicas de análisis multivariable aplicadas a las ciencias sociales.* CIS. Madrid, 1984.

Sokal, R.R.; Sneath, P.H. *Principles of Numerical Taxonomy.* W.H. Freeman and Co., San Francisco, 1963.

Wolfe, J.H. «A Monte Carlo Study of the Sampling Distribution of the Likelihood Ratio for Mixtures of Multinormal Distributions», en *Naval Personnel and Training Research Laboratory Technical Bulletin STB,* 72-2. San Diego, California, 1971.