Anuario de Psicología
1992, nº *55*, 63-90

Aplicación y valoración de diferentes algoritmos no-jerárquicos en el análisis *cluster* y su representación gráfica

Rosa Estarelles*
E. Inmaculada de la Fuente**
Pedro Olmedo*
**Universidad de Valencia*
***Universidad de Granada*

*Se ha revisado y comparado el comportamiento de tres algoritmos no-jerárquicos de análisis-*cluster: *K-Means, K-Medoides y el algoritmo propuesto por Dunn y Bezdek, basado en el principio borroso, para la exploración de las principales variables que explican las actitudes hacia el consumo de drogas en adolescentes. Los resultados confirman los hallazgos encontrados en otros estudios y sugieren que las actitudes respecto a las drogas pueden ser entendidas como comportamiento antisocial y capacidad para diferir la recompensa. Existe una relación inversa entre autoestimación y uso de las sustancias. Se destaca el papel exploratorio de la técnica K-Means, como un valioso punto de partida a la hora de generar hipótesis científicas, respecto a la mayor precisión diferencial de técnicas más robustas como K-Medoides, y la mayor aplicación del algoritmo de Dunn y Bezdek en estudios de diagnóstico y seguimiento individual.*
Palabras clave: *Análisis* Cluster, *técnicas no-jerárquicas, K-Means, K-Medoides, técnicas* Cluster *basadas en el principio borroso, gráfico de siluetas, actitud consumo drogas adolescentes.*

We have considered and compared three Non-hierarchical methods, namely K-Means, K-Medoid and the algorithm proposed by Dunn and Bezdek, in order to explore the main variables that explain the teenager's attitudes towards drugs abuse. The results support the current literature suggesting that involvement and attitudes towards drugs can be accounted by Antisocial Behavior and Ability to delay the Reward and an inverse relation exist between Self-Concept and Substances Use. We consider necessary to emphasize the importance of Locus of Control as a key question for prevention and intervention strategies. In current statistical practice both exploratory data analysis and robust and resistant methods have gained

Dirección de los autores: Rosa Estarelles y Pedro Olmedo. Departamento de Metodología, Psicobiología y Psicología Social, Área de Metodología de las Ciencias del Comportamiento, Facultad de Psicología. Avda. de Blasco Ibáñez 21. 46010 Valencia. E. Inmaculada de la Fuente. Departamento de Sociología y Psicología Social, Facultad de Filosofía y Letras. Edificio B, Campus de Cartuja. 18701 Granada.

important roles. These results serve to point out some possible natural groupings as a first step in generating scientific hypotheses for the purpose of further analysis.

Key words: *Cluster Analysis, Non-hierarchical Clustering Methods, K-Means, K-Medoids, Fuzzy Analysis, Silhouette Plot, Silhouette Coefficient, Attitudes towards Drugs in Abuse in Adolescence.*

El análisis *cluster*, dada su gran versatilidad y el importante papel que desempeña como técnica exploratoria, ocupa desde hace más de dos décadas un lugar destacado en muchas líneas y trabajos de investigación. Ello ha motivado una creciente proliferación de algoritmos, junto con gran número de programas de computador que facilitan su aplicación. Estos algoritmos permiten distinguir entre dos grandes grupos de técnicas: a) *Las técnicas no-jerárquicas o de partición*, que clasifican el conjunto de datos en K grupos, siendo el valor K especificado por el usuario, manteniéndose que $K \leq n$; y b) *Las técnicas jerárquicas* que conformarán un conjunto con varios niveles de diferenciación.

La elección de un método u otro dependerá tanto del tipo de datos disponibles como del objetivo de la investigación. Sin embargo, un grave problema que afecta a estos acercamientos es que ninguno de ellos puede ofrecer garantía de haber recuperado la verdadera estructura de los datos. A pesar de ello, muchas técnicas han sido comparadas según su capacidad para recobrar una estructura, supuestamente conocida, de un conjunto de datos, y bastantes las revisiones que se han realizado al respecto (Cormack, 1971; Ball, 1971; Hartigan, 1975; Aldenderfer y Blashfield, 1987; Milligan y Cooper, 1987; Dreger, Fuller y Lemoine, 1988, entre otras).

Lamentablemente no existen trabajos de validación que, hoy en día, permitan ofrecer y sostener una postura determinante al respecto, aunque se destacan los realizados por Blashfield y Morey (1980), Bayne, Beauchamp, Begovich y Kane (1980), y Scheibler y Schneider (1985), llegando a sugerirse la superioridad de ciertas técnicas no-jerárquicas por su robustez (Bayne y cols., 1980; Milligan, 1980), o cuando menos su equiparación con las técnicas jerárquicas (Scheibler y Schneider, 1985), además de la ventaja de que gozan las primeras, al permitir que la ubicación inicial de un objeto en un *cluster* no sea un hecho inamovible, tal como ocurre en muchos de los procedimientos jerárquicos.

No obstante, puede ser acertada la aplicación de algoritmos diferentes, y argumentos a priori pueden no ser suficientes para seleccionar uno solo de entre todos ellos, siendo permisible, en tales situaciones, la aplicación de varios algoritmos sobre un mismo conjunto de datos (aun a sabiendas de que los agrupamientos que se obtengan podrán diferenciarse tanto en número como en contenido), así como el análisis y comparación de las clasificaciones resultantes, para lo cual puede ser de gran ayuda la amplia gama de posibilidades gráficas que estas técnicas permiten.

Ahora bien, previamente a la aplicación de cualquier algoritmo, primero deberá haberse procedido a una selección apropiada de las variables (Edelbrock, 1979; Everitt, 1977; Fowlkes, Gnanadesikan y Kettenring, 1988), así como haber-

se optado por una medida que indique la similitud entre objetos (o distancia), de tal manera que, partiendo de una matriz $n \times p$, en donde filas se corresponden con objetos y columnas con atributos, se llegue a una matriz $n \times n$ de proximidades entre objetos.

Sin profundizar aquí en la amplia literatura existente en torno al concepto de similitud y su definición (véase Hartigan, 1967; Sneath y Sokal, 1973; Clifford y Stephenson, 1975; Tversky, 1977, entre otros), la elección de una medida de similitud, ya sean coeficientes de correlación, medidas de distancia, coeficientes de asociación o medidas de similitud probabilística, estará, en última instancia, determinada por la calidad y propiedades métricas de los datos disponibles, pudiendo plantearse la unificación de las variables seleccionadas en una sola matriz de proximidad, en la línea ya planteada por Ducker, Williams y Lance en 1965, retomada en 1971 por Gower, y ampliada por Kaufman y Rousseeuw en 1989.

El objetivo del presente trabajo ha sido doble: *explorar* las principales variables que explican las actitudes hacia el consumo de drogas en adolescentes, para lo cual se ha *revisado* y *comparado* el comportamiento de tres destacados algoritmos no-jerárquicos, ubicados en los siguientes apartados:

a) *Métodos basados en la construcción de puntos centrales;*
b) Métodos basados en la búsqueda de k objetos representativos;
c) Métodos basados en el principio borroso,

y de los que a continuación hacemos una breve introducción. (Véanse Apéndices para mayor información.)

Respecto a las posibilidades gráficas de estas técnicas, y específicamente en lo que se refiere a los métodos no-jerárquicos, es fácilmente constatable que no son tan copiosas como las existentes para los métodos jerárquicos, aunque la década de los 80 ha sido la más fructífera en este sentido (véase Dunn y Landwehr, 1980; Chambers y Kleiner, 1982; Wainer, 1983; Gale, Halperin y Costanzo, 1984; Wegman, 1985; Edmonston, 1986, y Rousseeuw, 1987).

A lo largo de este trabajo se han utilizado dos tipos de representaciones gráficas: los *perfiles de cluster* propuestos por Bailey y Dubes en 1982, y *los gráficos de siluetas* propuestos por Rousseeuw en 1987. Los primeros fueron introducidos dentro del campo de la verificación de hipótesis, mientras que los segundos podrían considerarse como piezas claves del análisis al ofrecer información sobre cada uno de los objetos agrupados. Los gráficos de siluetas se representan dentro de un mismo bloque, permitiendo observar con gran nitidez la calidad del agrupamiento. Se obtendrán siluetas para cada uno de los *clusters* especificados, siendo su altura proporcional al número de objetos, y su amplitud un índice de la diferenciación entre los distintos *clusters* obtenidos (véase Apéndice II).

Métodos basados en la construcción de puntos centrales

De entre todos los acercamientos no-jerárquicos son posiblemente los basados en la construcción de puntos centrales los que se utilizan con mayor fre-

cuencia, y muy especialmente la técnica K-Means, propuesta por Mac Queen en 1967, y que analizaremos a continuación. No obstante, la década de los 60 fue prolífica en esta línea, siendo muchos los autores que desarrollarían algoritmos ciertamente similares: Forgy, 1965; Ball y Hall, 1965; Jancey, 1966; Friedman y Rubin, 1967; Diday, ya en 1971, algunos de ellos incorporados en el programa CLUSTAN (Wishart, 1978).

El objetivo de estas técnicas será la optimización de un criterio formal predefinido, basado en la minimización de las sumas de cuadrados de las distancias (euclídeas) entre los objetos de un *cluster* y su centroide (punto en un espacio p-dimensional al que se ha llegado tras el promedio de los valores para cada variable seleccionada), por lo que también se conoce a este tipo de acercamiento con el nombre de técnicas de minimización de la varianza (véase Apéndice I).

Una característica de este método es que el número de *clusters* obtenidos es un valor fijo, de tal manera que, si un *cluster* únicamente estuviera compuesto por dos objetos, y uno de ellos pasara a otro *cluster*, el primero quedaría definido por un solo objeto, que, a la vez, sería el centroide de aquel *cluster*. Entre las desventajas de esta técnica se encuentran: a) ser su resolución excesivamente dependiente del orden en el que se introdujeron los datos; b) los centroides no tienen por qué corresponder con objetos del conjunto real de datos, aunque para una investigación aplicada suele ser de gran interés el conocer cuál es el objeto representativo de cada *cluster*; y c) estar excesivamente influida su resolución por valores extremos.

Métodos basados en la búsqueda de *k* objetos representativos

El objetivo de estos métodos, tal como su nombre indica, es la obtención de un conjunto de *k* objetos representativos (considerados como centros, medianas o «medoides»), de tal manera que los *k clusters* especificados por el usuario se construirán mediante la asignación de cada objeto del conjunto de datos al objeto representativo más próximo. Entre los métodos usualmente más utilizados se encuentran el conocido como «K-Centers», los métodos llamados de cobertura, así como la técnica «K-Medoides», que ha sido seleccionada en este trabajo por sus ventajas respecto a las anteriores.

El algoritmo para «K-Medoides» fue propuesto inicialmente por Vinod en 1969, retomado posteriormente por Rao (1971), Church (1978), Massart, Plastria y Kaufman (1983), Klastorin (1985) y Kaufman y Rousseeuw (1987, 1989) entre otros, siendo su cometido el de minimizar la suma de las distancias entre cada objeto y su correspondiente objeto representativo (véase Apéndice II).

Este tipo de formulación ofrece la posibilidad de facilitar la imposición de estructuras previamente especificadas, ya sea limitando el número de objetos en los *clusters*, ya sea evitando que un objeto sea seleccionado como objeto representativo. Pero no es ésta la única preeminencia de esta técnica; Kaufman y Rousseeuw (1989) destacan las ventajas derivadas de la aplicación de este acerca-

miento, y que pueden resumirse en los siguientes extremos: a) ser uno de los métodos más robustos (las técnicas basadas en la minimización de promedios de distancias, o de residuales en valores absolutos —L_1— son más robustas que las basadas en sumas de cuadrados —L_2—; b) ofrecer configuraciones bastante precisas cuando los *clusters* no son excesivamente alargados; c) sus agrupaciones no dependen del orden en que han sido introducidos los objetos, tal como se ha indicado puede suceder con otras técnicas no-jerárquicas.

Métodos basados en el principio borroso

Los métodos de partición borrosos ofrecen una información más detallada de la estructura de los datos que los dos acercamientos anteriormente expuestos, ya que ofrecen todo el abanico de probables ubicaciones para cada uno de los objetos, en los *k clusters*, aunque ello puede conllevar una interpretación más costosa, la implementación de cálculos más complejos que incrementan considerablemente el tiempo de procesamiento, la ausencia de especificación de objetos representativos, así como la necesidad de cuantificar el grado de pertenencia a cada uno de los *clusters*.

Entre los primeros algoritmos de agrupamiento borrosos se encuentra el propuesto como una generalización del método de K-Means por Dunn, 1974 y Bezdek, 1974, independientemente. Posteriormente, en 1978, Roubens, y más tarde, en 1982, Libert y Roubens propondrían una variante del primero (MND2). En este trabajo hemos seleccionado la variante propuesta por Kaufman y Rousseeuw en 1989 (véase Apéndice III), siendo sus ventajas sobre el resto de técnicas basadas en el principio borroso: a) ofrecer mejor conocimiento para *clusters* no esféricos; b) aportar una resolución menos dependiente de posibles valores extremos; y c) no presentar los sesgos sistemáticos que se observan para MND2, tanto al obtener la función del error, como al formular la composición de los *clusters* (como es una tendencia a ofrecer un número de objetos aproximado para cada *cluster*).

Método

Muestra

La muestra total de sujetos estuvo compuesta por 523 adolescentes, todos ellos estudiantes de BUP y FP, seleccionados aleatoriamente de diferentes Centros de la Comunidad Valenciana, con aproximadamente igual número de chicos que de chicas, y de edades comprendidas entre 14 y 17 años, de los cuales el 35 % tenía actitudes positivas hacia el uso de drogas, habiendo consumido algún tipo de sustancias el 68 % de ellos, mostrando el 35 % actitudes negativas hacia dicho consumo.

Del conjunto global de datos se seleccionaron aleatoriamente dos muestras diferentes que, conservando idéntica estructura a la total, según nivel de edad, sexo y confirmación de consumo y actitud hacia las drogas, permitiera la validación de los resultados. Cada una de estas muestras estuvo compuesta por 94 sujetos, 33 pro consumo de drogas, de los cuales 21 eran chicos y 12 chicas, y 61 en contra del uso de sustancias, siendo 30 los varones y 31 las chicas.

Instrumentos

Se han utilizado los seis instrumentos siguientes:

1. El Inventario de Ansiedad Estado-Rasgo (STAI) (Spielberger, Gorsuch y Lushene, 1970).
2. La Escala de Auto-Concepto Tennessee (Fitts, 1964).
3. La Escala de Ajuste y Control del Adolescente (Capafons, 1986).
4. El Inventario de Percepción Familiar (EMBU) (Perris y cols., 1980).
5. El Inventario de Conductas Antisociales (Allsopp y Feldman, 1976).
6. La Escala de Implicación en el Consumo de Sustancias (Estarelles, 1987).

Asimismo se consideraron las siguientes variables categóricas: sexo y consumo de drogas (habitual, ocasional, ausencia de consumo).

Análisis y resultados

Se han aplicado para cada uno de los conjuntos muestrales seleccionados los acercamientos no-jerárquicos anteriormente expuestos, obteniéndose las configuraciones que pasamos a comentar.

En primer lugar se expondrán los resultados finales obtenidos para la primera selección muestral, y tras la aplicación de la técnica K-Means. En el procedimiento de estandarización se ha utilizado la varianza intra-*cluster*, después de verificarse la ausencia de diferencias entre este tipo de estandarización y la obtenida a partir de la usual desviación típica, habiéndose tomado distancias euclídeas, y tras haberse seleccionado diferentes valores para k (2, 3 y 4), teóricamente fundamentados (Estarelles y Ferrer, 1990; Pardeck, 1991). Para la aplicación de las técnicas «K-Medoides» y «*Clusters* Borrosos» también se procedió previamente a la estandarización de los datos, utilizándose igualmente distancias euclídeas.

Dado que uno de los problemas que parece plantear la técnica K-Means es la dependencia del orden establecido en los datos en la configuración de los *clusters* ofrecidos, se han tenido en cuenta cuatro ordenaciones completamente distintas para todos los valores aplicados para k, y tanto para ambos conjuntos muestrales, como para los grupos de chicos y chicas. Se ha comprobado que los resultados ofrecidos, para cada una de las ordenaciones, eran totalmente idénticos. Este procedimiento también se llevó a cabo con las técnicas «K-Medoides» y «*Clusters* Borrosos», no observándose tampoco ninguna disparidad.

A continuación se presentan los resultados más relevantes conseguidos para k=2, según los diferentes acercamientos considerados, así como para k=4, siguiendo un mismo orden. Las variables seleccionadas incorporadas en los sucesivos análisis han sido: *Self-Ético-Moral, Conducta Antisocial, Auto-Control, y Actitud hacia el Consumo de Drogas*. Asimismo, se ha tenido en cuenta el conjunto de la selección muestral, y posteriormente se ha diferenciado entre los grupos de chicos y chicas, tras haberse verificado diferencias significativas entre sexos, para ambas muestras, y para todas las variables consideradas. Por último, no se han apreciado diferencias significativas consecuencia de las distintas edades comprendidas en el estudio, por tanto, en lo que a esta variable se refiere, la muestra se ha considerado como un todo.

Resultados para k=2 (K-Means)

El primer *cluster*, tanto para la muestra total como para los grupos de chicos y chicas, está compuesto por adolescentes con actitudes positivas hacia el consumo de drogas. Estos sujetos además se encontraban en el grupo de jóvenes que habían aceptado consumir ocasional o habitualmente drogas. Todos los casos fueron correctamente clasificados. En cambio, todos los jóvenes del segundo *cluster* habían manifestado actitudes negativas hacia el consumo de sustancias, así como no haber consumido drogas hasta aquel momento, habiéndose clasificado igualmente de forma correcta el 100 % de los sujetos. En la tabla siguiente se incluye la configuración de estos *clusters* para la muestra total y para los grupos de chicos y chicas.

TABLA 1. CONFIGURACIONES *CLUSTERS*: K-MEANS = 2

Concepto	Muestra Total		Varones		Mujeres	
Cluster	1	2	1	2	1	2
N° sujetos	33	61	21	30	12	31
Distancia Promedio	2.188	1.675	2.175	1.620	2.110	1.730

La distancia entre los centroides de ambos *clusters* fue de 12.0679, para la muestra total, y de 11.699 y 11.628 para los grupos de chicos y chicas, respectivamente.

En la Tabla 2 se incorpora la pseudo-razón F ofrecida por el paquete BMDP, que permite la ordenación de las variables, según su importancia relativa, para la configuración de los *clusters*. Aunque esta razón no puede entenderse como una prueba real de significación, ya que su objetivo es la maximización de la ra-

zón *entre* respecto a la *intra*, su inclusión es de gran utilidad para valorar la relevancia de las variables seleccionadas y comparar las diferentes agrupaciones.

TABLA 2. IMPORTANCIA RELATIVA DE LAS VARIABLES SELECCIONADAS

Grupo	Fuente	Variables			
		Self-Et.-M.	Cond. Antis.	Auto-Cont.	Act. Cons.
Muestra Total	Entre	1646.89	9727.56	557.62	3406.05
	Intra	76.05	28.99	12.56	1.25
	Razón F	21.66	335.45	44.41	2717.24
	Valor P	0.00	0.00	0.00	0.00
Varones	Entre	474.12	6240.45	225.38	2042.02
	Intra	65.97	32.45	12.83	1.39
	Razón F	7.19	192.33	17.56	1473.52
	Valor P	0.01	0.00	0.00	0.00
Mujeres	Entre	966.53	3094.86	273.86	1282.88
	Intra	73.50	22.25	10.62	1.07
	Razón F	13.15	139.10	25.79	1201.57
	Valor P	0.00	0.00	0.00	0.00

A continuación se incluyen perfiles de los *clusters* obtenidos para k = 2, estando ordenadas las variables según su importancia relativa a la hora de determinar los mismos. En este tipo de gráfico cada columna describe un *cluster*, el número del *cluster* representa la media de cada variable, los asteriscos representan la línea de la media total y los guiones representan desviaciones típicas.

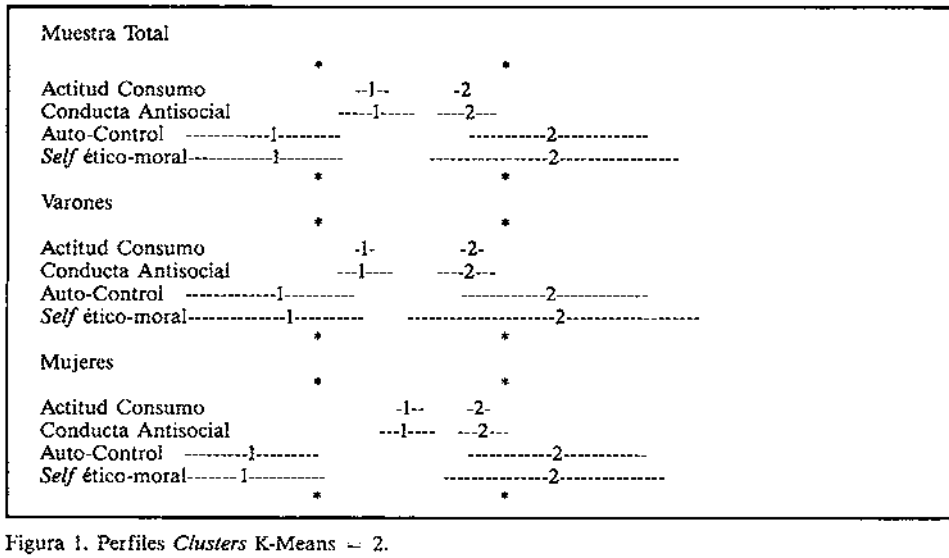

Figura 1. Perfiles *Clusters* K-Means = 2.

Tal como puede apreciarse existe una clara diferenciación entre ambos *clusters*, hecho que también se detecta con gran nitidez tras la observación, en dos y tres dimensiones, de las variables seleccionadas.

Resultados para k=2 (K-Medoides)

Se repite la misma composición y definición de características que para K-Means, pudiendo observarse esta configuración a través de las representaciones gráficas de siluetas para la muestra total (Figuras 2 y 3). El primer grupo se caracteriza por sus actitudes positivas hacia el consumo de drogas, mientras que, para el segundo conjunto se da el fenómeno inverso. No obstante, solamente fueron ubicados en este primer *cluster* treinta sujetos, habiendo sido incorrectamente clasificados en el segundo grupo tres jóvenes. Ahora bien, estos sujetos, tal como puede observarse en la Figura 3, se corresponden con los tres últimos sujetos (varones) incluidos en el segundo *cluster*, con un valor para sus coeficientes de silueta de 0.01, para uno de ellos, lo cual indica que bien podría pertenecer, indistintamente, a cualquiera de ambos *clusters*, y valores negativos para el resto (−0.11 y −0.13) que nos indicaría que estarían mejor ubicados en el primer *cluster*, aunque no con una pertenencia claramente definida (la clasificación de estos sujetos dentro del grupo de chicos es muy similar).

Los adolescentes seleccionados como objetos representativos para ambos *clusters* son claros exponentes de ambos conjuntos. Las características que definen al sujeto representativo para el primer *cluster* son: actitud a favor del consumo de sustancias y consumo más que ocasional de ellas, elevado número de comportamientos antisociales, y muy bajas puntuaciones en auto-control y valores ético-morales. En cambio, el sujeto representativo del segundo *cluster* estaría definido por los valores contrarios.

En la Tabla 3 se detalla la configuración de estos *clusters* según grupos. Idénticas definiciones y características que para la muestra total son válidas para el grupo de chicos y chicas.

En la Tabla 4 se especifican las características de los *clusters* a través de:

a) El diámetro de cada *cluster* (distancia máxima entre dos sujetos pertenecientes al mismo *cluster*).

b) La separación de cada *cluster* (distancia mínima entre cualquier sujeto perteneciente al *cluster* y cualquier otro sujeto del conjunto de datos).

c) La distancia promedio de los sujetos a cada objeto representativo o «medoide».

d) La distancia máxima a cada objeto representativo (máxima distancia entre cualquiera de los sujetos a su objeto representativo).

e) Amplitud promedio del coeficiente de silueta para cada *cluster*.

f) Amplitud promedio del coeficiente de silueta para la configuración total.

Se han tenido en cuenta los valores obtenidos para la muestra total, así como para los grupos de chicos y chicas.

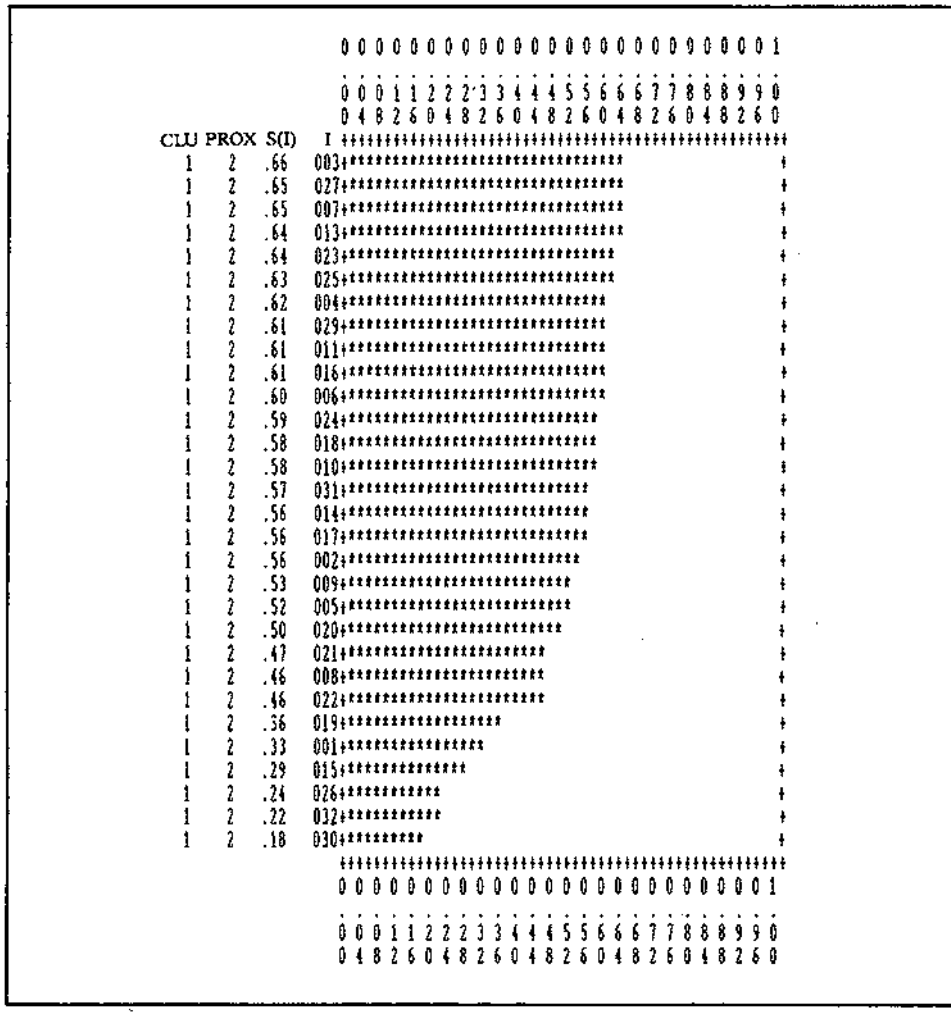

Figura 2. Gráfico de siluetas para el primer *cluster*. Muestra total (K-Medoides=2).

TABLA 3. CONFIGURACIÓN *CLUSTERS*: K-MEDOIDES = 2

Concepto	Muestra total		Varones		Mujeres	
Cluster	1	2	1	2	1	2
N° sujetos	30	64	18	33	12	31
Distancia Promedio	1.50		1.55		1.49	

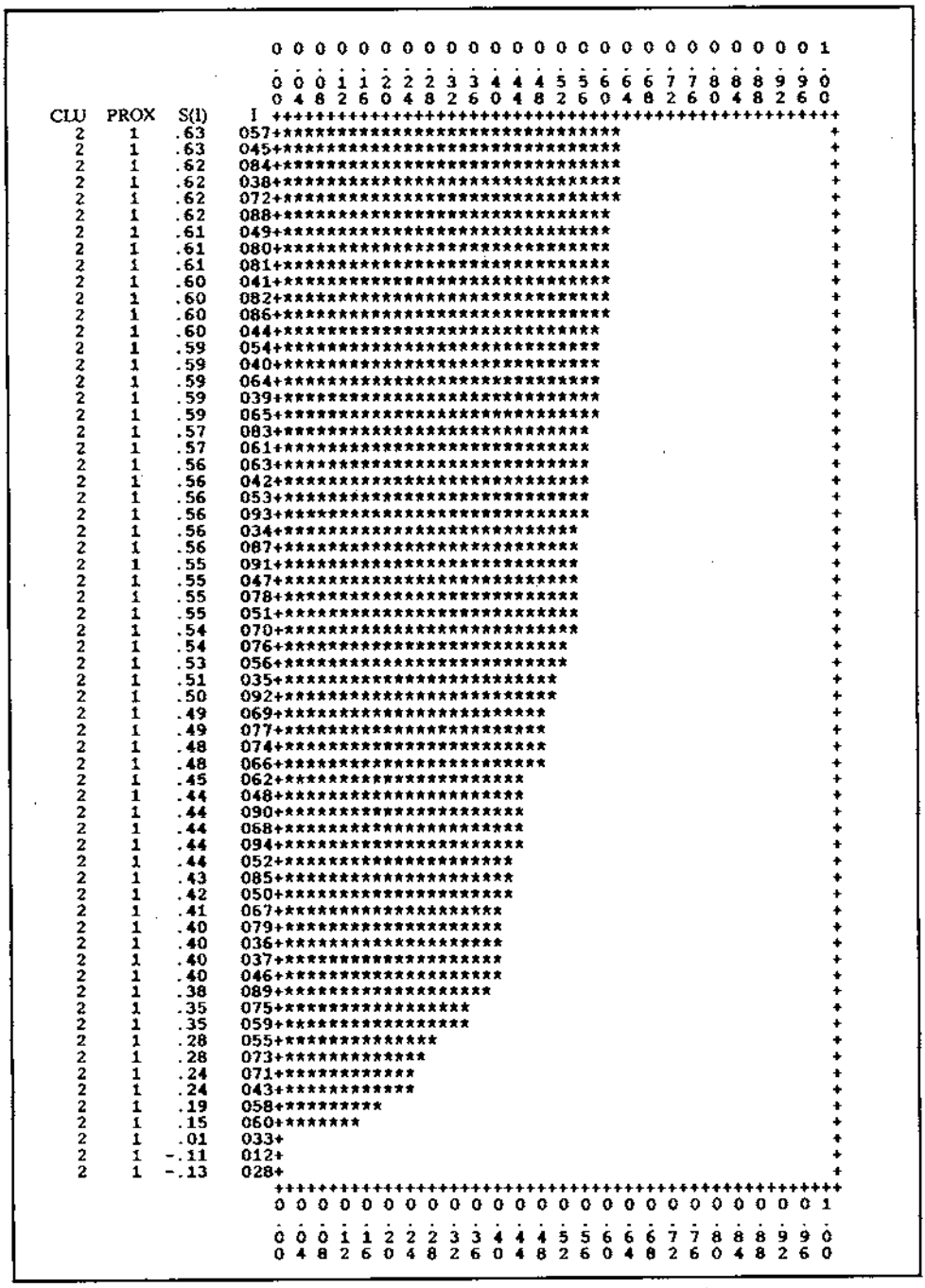

Figura 3. Gráfico de siluetas para el segundo *cluster*. Muestra total (K-Medoides=2).

TABLA 4. CARACTERÍSTICAS DE LOS *CLUSTERS*: K-MEDOIDES = 2

Grupo	Concepto	*Clusters*	
		1	2
Muestra Total	Diámetro *Cluster*	4.36	5.64
	Separación *Cluster*	0.59	0.59
	Distancia Promedio Medoide	1.44	1.53
	Distancia Máxima Medoide	2.95	2.84
	Amplitud Promedio Silueta	0.52	0.47
	Promedio Silueta Total	0.48	
Varones	Diámetro *Cluster*	4.33	5.52
	Separación *Cluster*	1.15	1.15
	Distancia Promedio Medoide	1.25	1.71
	Distancia Máxima Medoide	3.17	3.12
	Amplitud Promedio Silueta	0.54	0.40
	Promedio Silueta Total	0.45	
Mujeres	Diámetro *Cluster*	4.46	4.83
	Separación *Cluster*	2.49	2.49
	Distancia Promedio Medoide	1.61	1.45
	Distancia Máxima Medoide	2.43	3.23
	Amplitud Promedio Silueta	0.49	0.53
	Promedio Silueta Total	0.52	

Resultados para k=2 (*clusters* borrosos)

Tras la aplicación de esta técnica se observa que algunas especificaciones para *k* ofrecen *clusters* más borrosos que otras (el máximo valor de pertenencia corresponde a 1/k). Esta peculiaridad del presente acercamiento se ha presentado a lo largo de las distintas aplicaciones efectuadas. Para k=2 los resultados fueron bastante nítidos, corroborándose la línea de actuación apuntada en los dos métodos antes comentados, tanto para la muestra total como para los grupos de chicos y chicas.

TABLA 5. CONFIGURACIÓN *CLUSTERS* (*CLUSTERS* BORROSOS) K = 2

Concepto	Muestra Total		Varones		Mujeres	
Cluster	1	2	1	2	1	2
N° sujetos	35	59	20	31	15	28

A continuación se incluyen gráficos de siluetas para la muestra total, pudiendo observarse la semejanza con la solución obtenida para K-Medoides, ya que se incluyen erróneamente dos sujetos en el primer *cluster* (un chico y una chica), pero con coeficientes de silueta negativos, es decir, con una posible mejor *ubicación en el segundo cluster* (Figuras 4 y 5). Si observamos estos sujetos en la gráfica anteriormente obtenida, tras aplicar la técnica «K-Medoides», puede comprobarse que obtuvieron los coeficientes de valor inferior para el segundo *cluster*.

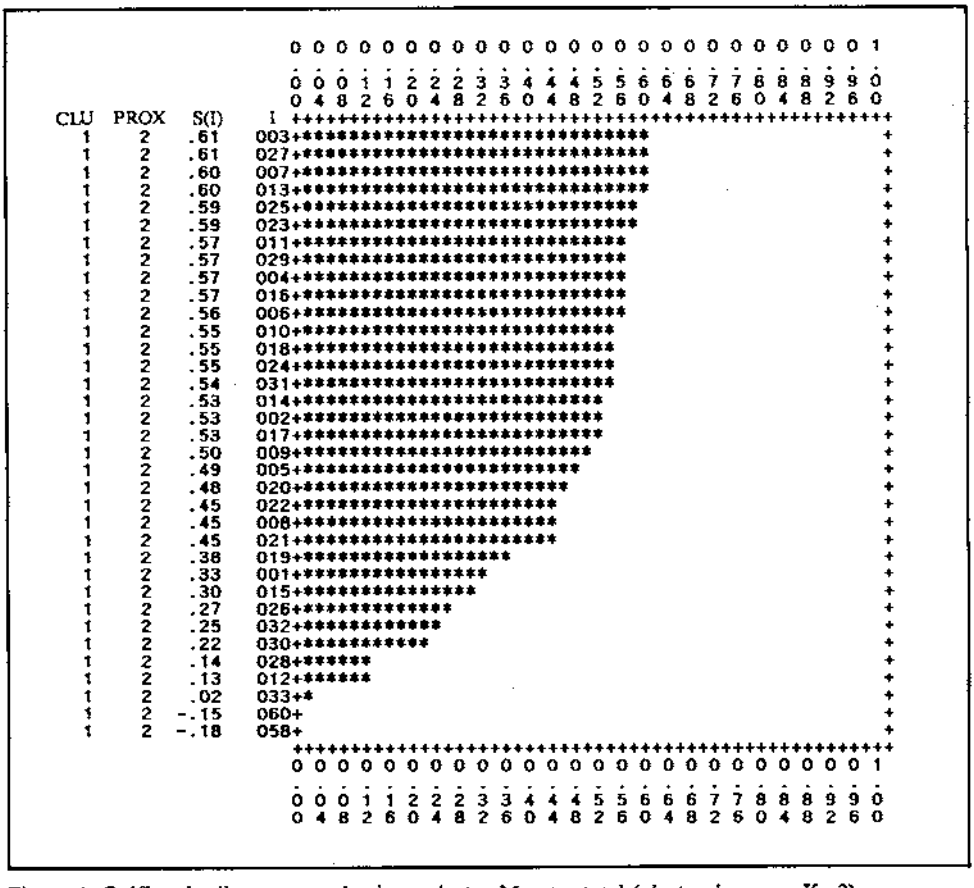

Figura 4. Gráfico de siluetas para el primer *cluster*. Muestra total (*clusters* borrosos K=2).

Para la muestra total, el 73 % de los sujetos obtuvo valores de pertenencia superiores a 0.70, y ninguno inferior a 0.56. El grupo de varones alcanzó índices de pertenencia superiores a 0.70 en el 71 % de los casos, y sólo se observaron tres puntuaciones inferiores a 0.50. En el grupo de chicas el 76 % obtuvo valores por encima de 0.70 y sólo un valor fue inferior a 0.50.

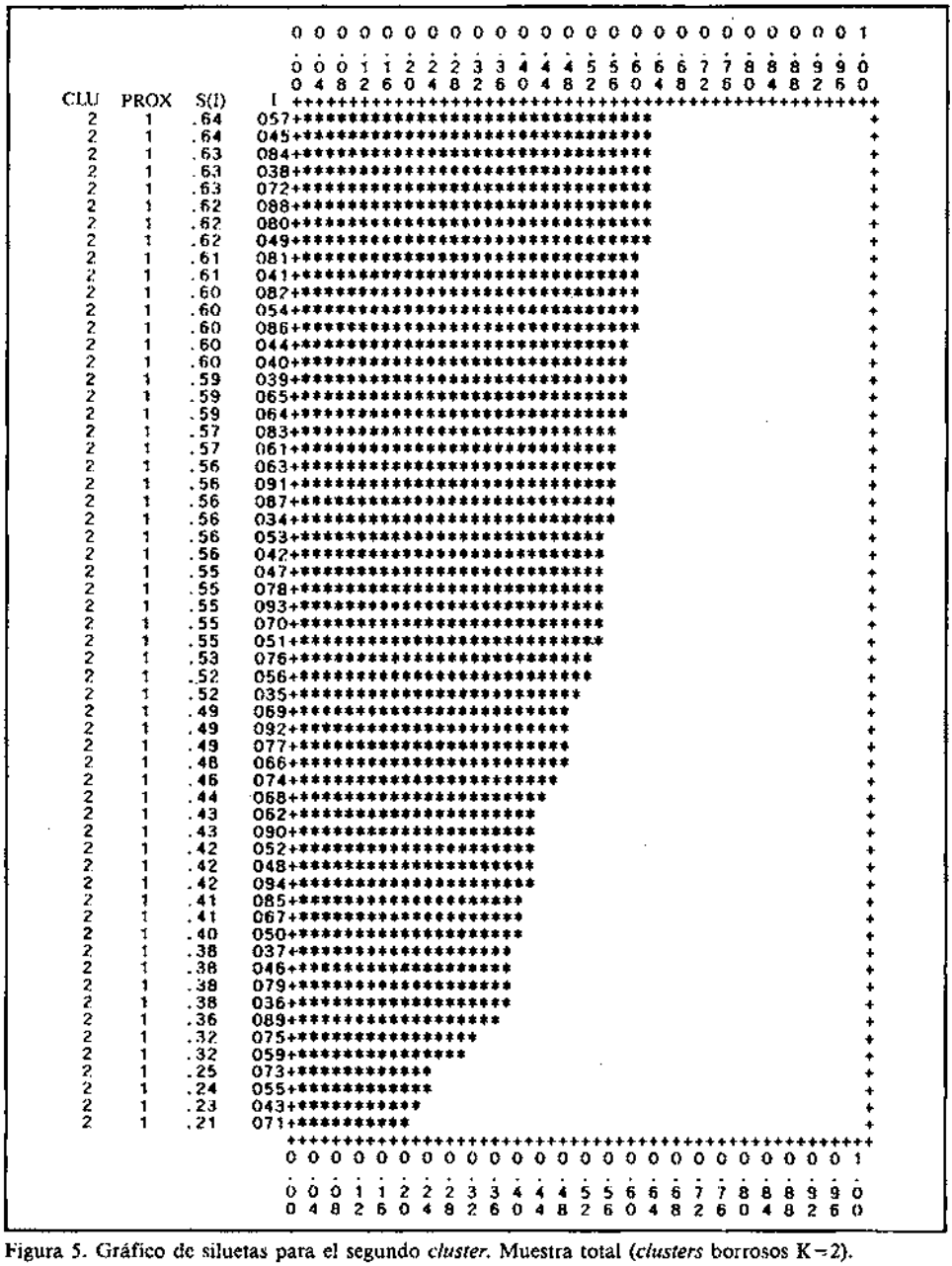

Figura 5. Gráfico de siluetas para el segundo *cluster*. Muestra total (*clusters* borrosos K=2).

Otros resultados de interés, tal como el coeficiente de Dunn y su valor normalizado, así como la amplitud promedio para cada silueta, y para toda la estructura, se incluyen en la Tabla 6.

TABLA 6. CARACTERÍSTICAS *CLUSTERS*: K = 2 (CONJUNTOS BORROSOS)

Grupo	Concepto	*Clusters*	
		1	2
Muestra Total	Promedio Silueta *Cluster*	0.42	0.50
	Promedio Silueta Total Coeficiente Pert. Dunn Coeficiente Normalizado	0.47 0.60 0.20	
Varones	Promedio Silueta *Cluster*	0.48	0.44
	Promedio Silueta Total Coeficiente Pert. Dunn Coeficiente Normalizado	0.45 0.60 0.20	
Mujeres	Promedio Silueta *Cluster*	0.32	0.56
	Promedio Silueta Total Coeficiente Pert. Dunn Coeficiente Normalizado	0.48 0.61 0.22	

La semejanza entre la estructura ofrecida por la presente técnica y K-Medoides para k=2 es evidente, siendo también sus resultados equiparables para los grupos de varones y mujeres.

Resultados para k=4 (K-Means)

El primer *cluster* para la muestra total, así como para los conjuntos de chicos y chicas estuvo compuesto por adolescentes con actitudes muy positivas hacia el consumo de drogas, habiendo manifestado la mayoría de ellos su consumo habitual. El *cluster* número tres, en lo que se refiere a la muestra total y al grupo de varones, corresponde a sujetos con moderada tendencia hacia el consumo de drogas, y que manifestaron consumir sólo ocasionalmente algunas sustancias «blandas». En cambio, para las chicas se observa una agrupación similar a la del primer *cluster* que acabamos de comentar, en este caso representado por el número cuatro, pero en la que se destaca como característica esencial el bajo auto-concepto ético-moral observado para todas las jóvenes. Todos los sujetos del segundo *cluster* manifestaron actitudes totalmente negativas hacia el consumo de sustancias, correspondiendo al *cluster* múmero cuatro de la muestra total y del grupo de varones, y al tres para las chicas, una postura moderada en contra del consumo de drogas, tal como puede observarse en la Tabla 7 y en la Figura 6.

TABLA 7. CONFIGURACIÓN *CLUSTERS* K-MEANS = 4

Concepto	Muestra Total				Varones				Mujeres			
Cluster	1	2	3	4	1	2	3	4	1	2	3	4
Nº sujetos	19	21	14	40	11	10	10	20	8	14	17	4
Dist. Prom.	2.1	1.5	2.3	1.7	1.9	1.5	2.3	1.6	2.4	1.5	1.4	1.8

En la Tabla 8 se puede observar la ordenación de las variables, según su importancia relativa para la configuración de los *clusters*, a través de la pseudo-F.

TABLA 8. IMPORTANCIA RELATIVA DE LAS VARIABLES

Grupo	Fuente	Variables			
		Self-Et.-M.	Cond. Antis.	Auto-Cont.	Act. Cons.
Muestra Total	Entre	1818.01	3625.54	227.93	1136.37
	Intra	35.44	16.88	11.43	1.25
	Razón F	51.30	214.88	19.93	911.09
	Valor P	0.00	0.00	0.00	0.00
Varones	Entre	750.99	2368.42	93.68	681.31
	Intra	30.93	15.43	12.19	1.40
	Razón F	24.28	153.52	7.68	485.17
	Valor P	0.00	0.00	0.00	0.00
Mujeres	Entre	951.04	1105.57	121.44	428.85
	Intra	28.89	17.70	8.84	1.03
	Razón F	32.92	62.46	13.74	416.95
	Valor P	0.00	0.00	0.00	0.00

Por último, a continuación se incluyen perfiles de los cuatro *clusters* para los tres grupos diferenciados:

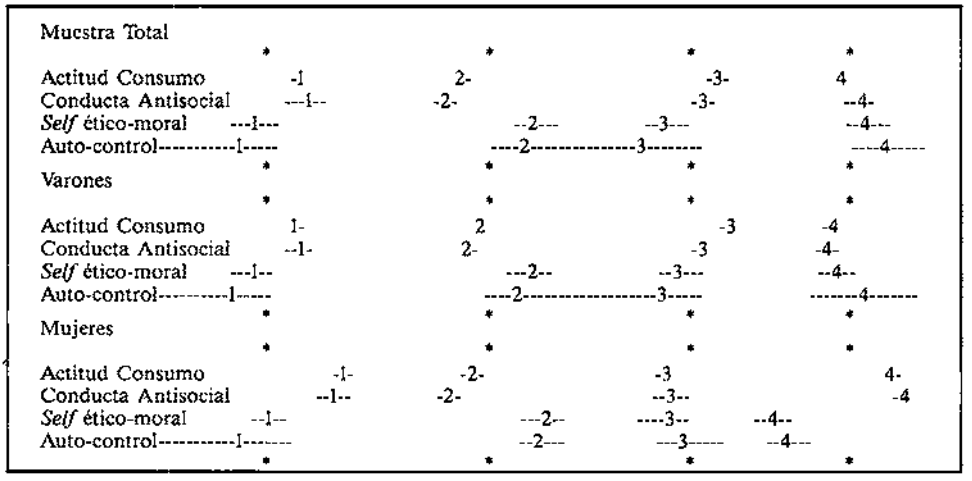

Figura 6. Perfiles *clusters* K-Means = 4.

Resultados para k=4 (K-Medoides)

La solución obtenida para k=4 ubica correctamente a todos los sujetos menos uno. Sin embargo, la configuración para la muestra total es diferente a la obtenida para K-Means, en la medida que son tres los conjuntos diferenciados con actitudes en contra del consumo de drogas y, solamente uno a favor, agrupándose en un mismo *cluster* los sujetos que expresaron su consumo habitual y ocasional de drogas. El sujeto representativo del primer *cluster*, define a este grupo con las mismas características con las que quedó determinado este mismo conjunto a través de la técnica K-Means. Seguidamente se incluye representación gráfica para el *cluster* de adolescentes a favor del consumo de drogas (Figura 7).

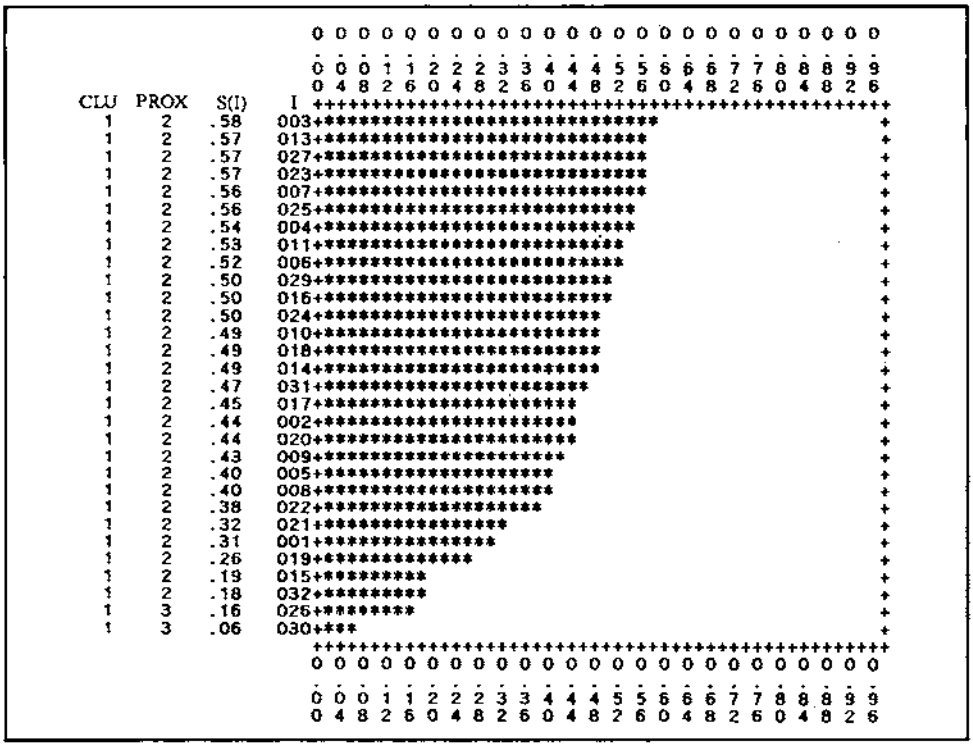

Figura 7. Gráfico de siluetas para el primer *cluster*. Muestra total (K-Medoides=4).

El cuarto *cluster* se caracteriza por el comportamiento opuesto. En cambio, el tercero mantiene sólo posturas moderadas contra las drogas, mientras que el segundo grupo, representado por una chica, al igual que lo son la mayoría de los componentes de este *cluster*, se define por un bajo número de conductas antisociales, bajos valores en auto-concepto y auto-control, así como actitudes moderadas en contra del consumo de drogas. Este *cluster* incorpora dos jóvenes pertenecientes al primer conjunto, aspecto éste que se detecta claramente si observamos el valor de sus coeficientes, −0.07 y −0.14, respectivamente (Figura 8).

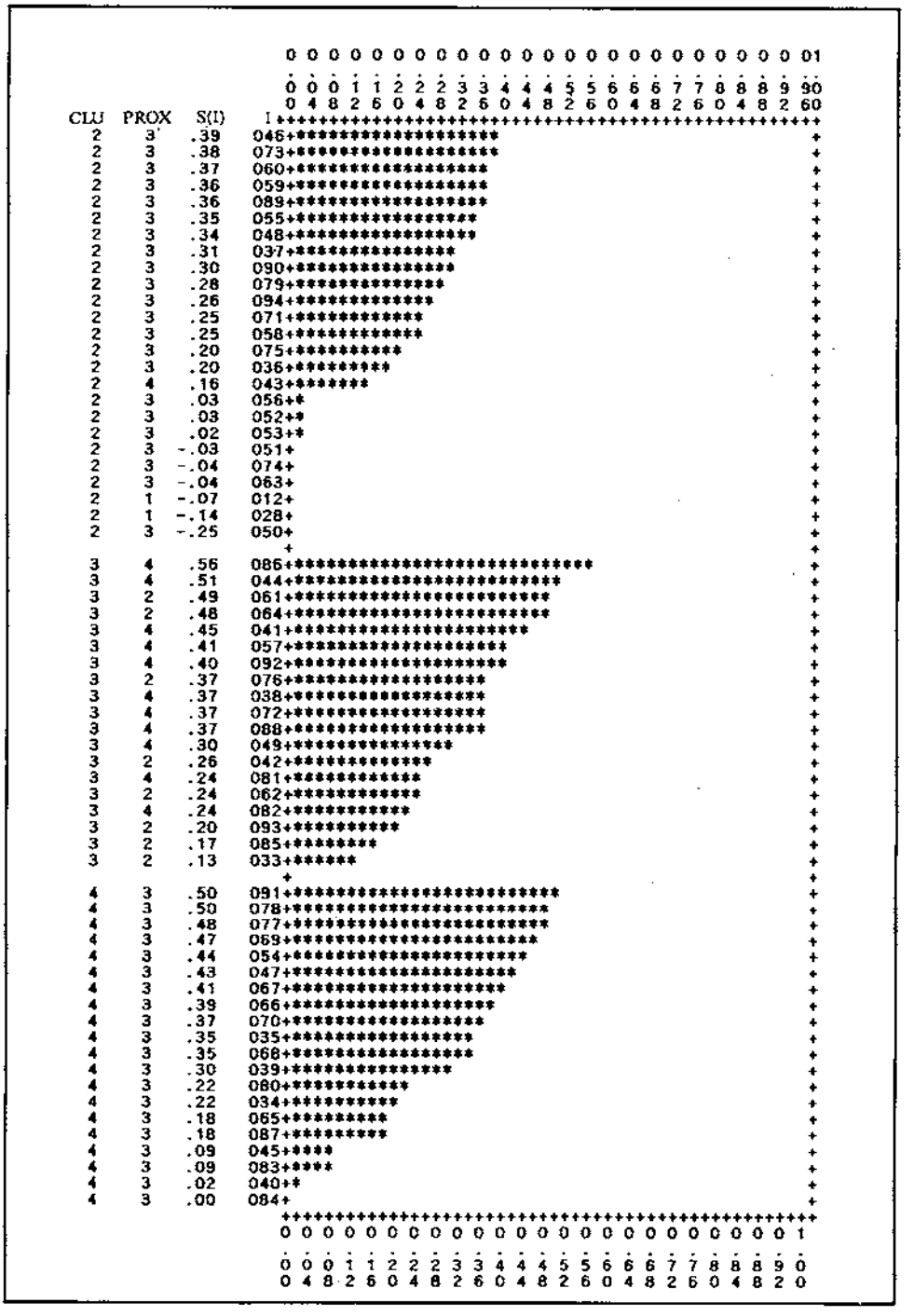

Figura 8. Gráfico de siluetas para los tres *clusters* de adolescentes. Muestra no consumidores. Muestra total (K-Medoides=4).

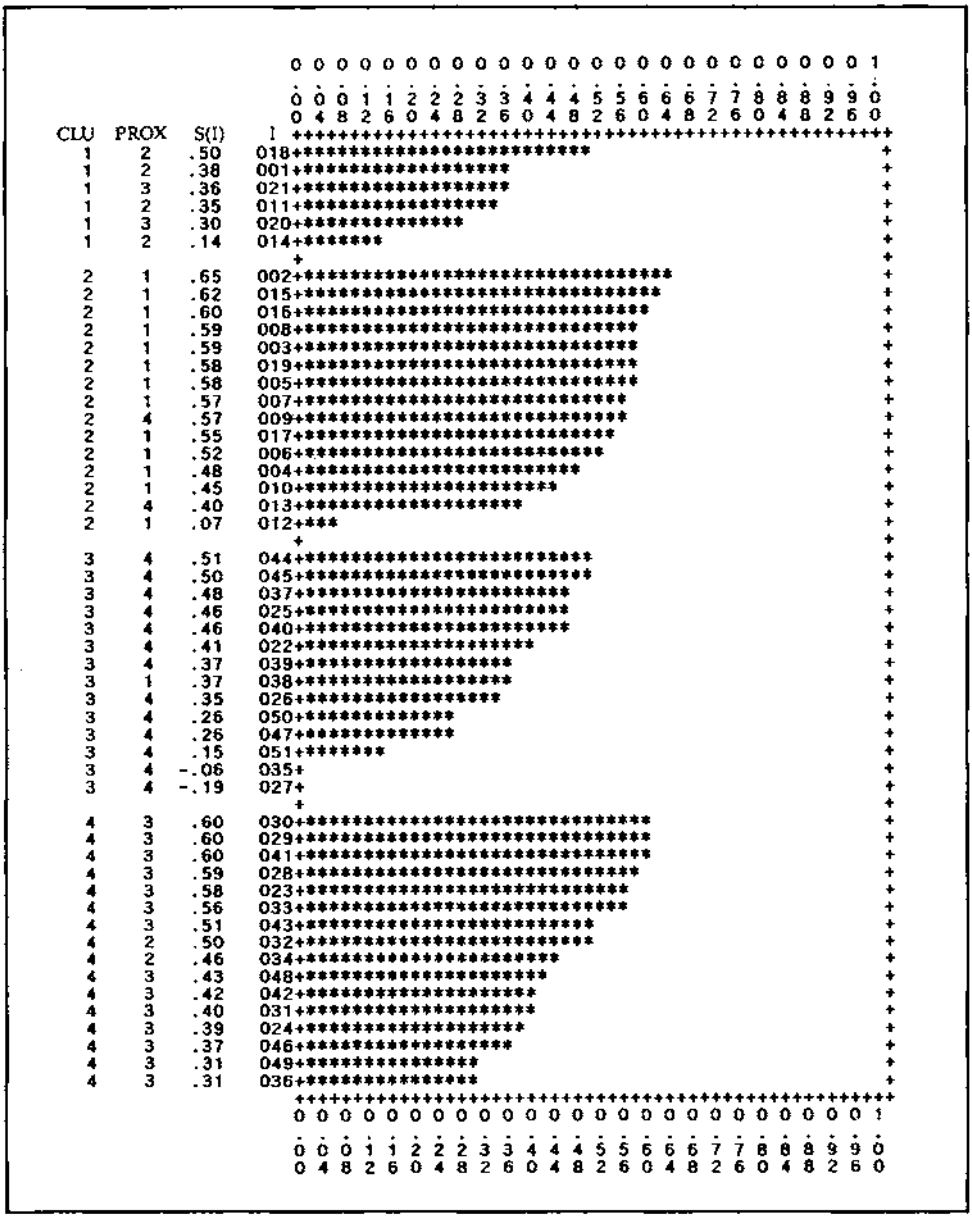

Figura 9. Gráfico de siluetas para varones (K-Medoides=4).

Configuraciones similares se observan para el grupo femenino, integrándose en un mismo conjunto consumidoras habituales y ocasionales, y perfilándose un *cluster* de no consumidoras, con actitudes moderadamente en contra de la droga, pero también con reducido número de conductas antisociales y bajos niveles de auto-control y auto-concepto ético-moral. En cambio, para el conjunto de varones se ofrece el mismo tipo de agrupamientos y definición de características que se observó con la técnica K-Means para este mismo grupo. Como puede comprobarse la aplicación de esta técnica ha permitido precisar mejor las diferencias entre chicos y chicas, aspecto éste de sumo interés para la conformación de un modelo de prevención. Pueden observarse las diferencias entre adolescentes de ambos sexos en las gráficas 9 y 10.

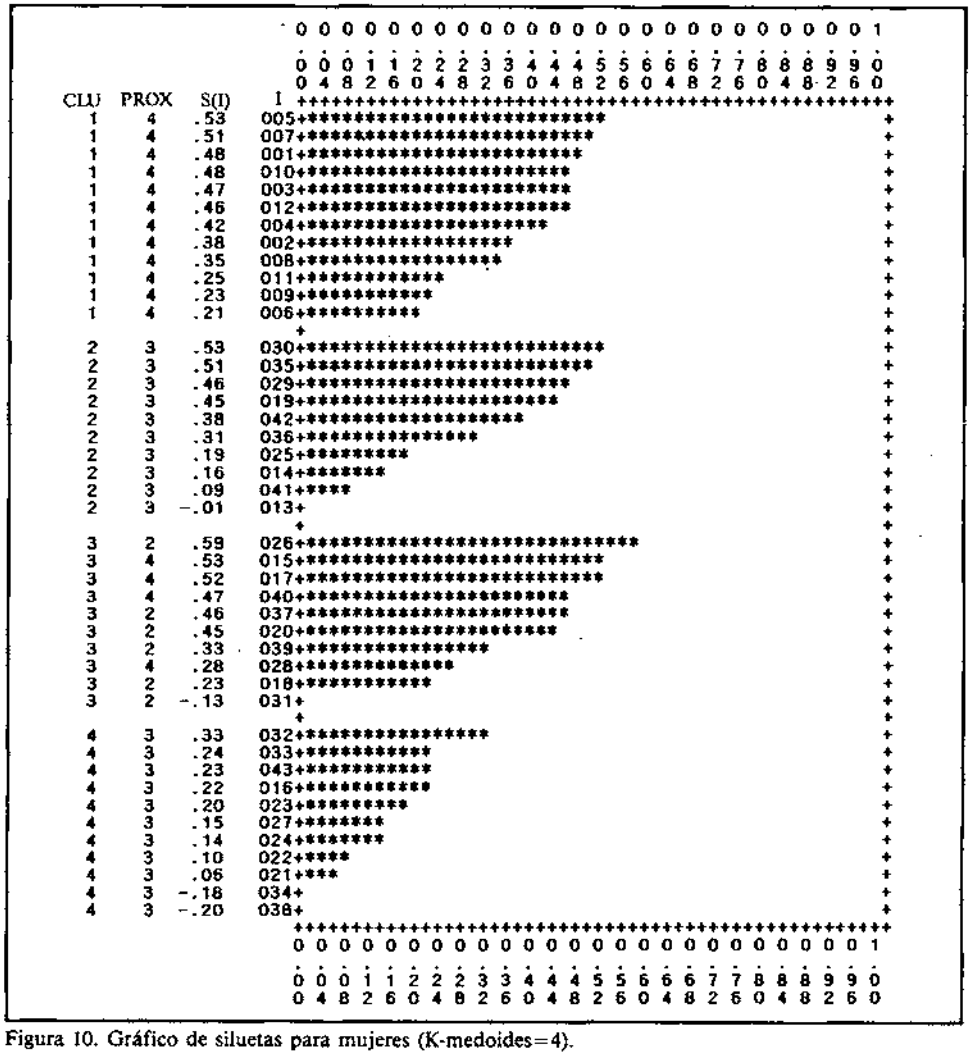

Figura 10. Gráfico de siluetas para mujeres (K-medoides=4).

Las Tablas 9 y 10, que se incluyen seguidamente, muestran la configuración y características de los *clusters* para la muestra total de adolescentes, así como para ambos sexos.

TABLA 9. CONFIGURACIONES *CLUSTERS* K-MEDOIDES = 4

Concepto	Muestra Total				Varones				Mujeres			
Cluster	1	2	3	4	1	2	3	4	1	2	3	4
N° sujetos	30	25	19	20	6	15	14	16	12	10	10	11
Dist. Prom.	1.149				1.096				1.101			

TABLA 10. CARACTERÍSTICAS DE LOS *CLUSTERS*: K-MEDOIDES = 4

Grupo	Concepto	*Clusters*			
		1	2	3	4
Muestra Total	Diámetro *Cluster*	4.36	4.24	2.79	2.84
	Separación *Cluster*	0.59	0.51	0.59	0.51
	Distancia Promedio Medoide	1.44	1.17	0.80	1.01
	Distancia Máxima Medoide	2.95	2.84	2.18	1.69
	Amplitud Promedio Silueta	0.43	0.17	0.35	0.30
	Promedio Silueta Total	0.32			
Varones	Diámetro *Cluster*	2.77	3.34	4.18	2.88
	Separación *Cluster*	1.43	1.43	0.72	0.72
	Distancia Promedio Medoide	1.29	1.00	1.24	0.99
	Distancia Máxima Medoide	2.24	2.11	3.03	1.83
	Amplitud Promedio Silueta	0.34	0.52	0.52	0.48
	Promedio Silueta Total	0.43			
Mujeres	Diámetro *Cluster*	4.46	1.99	2.28	3.34
	Separación *Cluster*	2.49	0.71	0.46	0.46
	Distancia Promedio Medoide	1.61	0.86	0.67	1.15
	Distancia Máxima Medoide	2.43	1.26	1.78	2.35
	Amplitud Promedio Silueta	0.40	0.31	0.37	0.12
	Promedio Silueta Total	0.30			

Resultados para k=4 (conjuntos borrosos)

Tal como se ha indicado, a través de la aplicación de este acercamiento, y según la especificación ofrecida, se obtienen agrupaciones diferencialmente bo-

rrosas. La solución para k=2 ofrecía una estructura claramente definida, los resultados conseguidos para k=3 fueron ligeramente más difusos, con una ubicación correcta para el 56 % de los sujetos (con valores de pertenencia superiores a 0.65), en cambio, para k=4 la solución fue aún más indefinida, con valores de pertenencia superiores a 0.65 sólo en el 38 % de los casos. Las ubicaciones fuertemente definidas correspondieron, tal como era de esperar, a aquellos sujetos con posturas claramente expresadas a favor y en contra del consumo de sustancias.

Obviamente, dentro de la ambigüedad de la resolución de esta técnica se encuentra su propia grandeza. Es de innegable utilidad para estudios de diagnóstico, en los que es preciso obtener un claro pronóstico, así como para la puesta a punto de modelos de prevención. Por todo ello, los resultados alcanzados tras la aplicación de esta técnica han sido de gran interés en el seguimiento de casos individuales, estando en estudio el proceso de evolución de posturas limítrofes, así como sus factores de riesgo.

Resultados de la validación

Por último, y con el fin de validar los resultados obtenidos, se tomó la restante selección muestral, de igual tamaño y composición equivalente, inicialmente identificada. Tras replicar cada uno de los análisis anteriormente expuestos, y para los cuatro tipos de ordenación considerados, los resultados alcanzados corroboraron los hasta aquí comentados, no apreciándose en ningún análisis dependencia en las soluciones obtenidas del orden de los datos.

Conclusiones

Los resultados anteriores confirman la existencia de una estructura bastante definida, con dos conjuntos claramente diferenciados por las variables seleccionadas, hallazgos que se ven refrendados por investigaciones anteriores (Watts y Wright, 1990; Estarelles y Ferrer, 1990; Pardeck, 1991; Thorlindsson y Vilhjalmsson, 1991; Estarelles, de la Fuente y Olmedo, 1992, etc.). Todas las variables incluidas han demostrado ser buenas predictoras de la implicación juvenil en el consumo de sustancias y estudios recientes, orientados hacia la intervención, están ratificando el importante papel desempeñado por el nivel de auto-control alcanzado, así como la relevancia de algunos factores de riesgo, distintos para los grupos de chicos y chicas, ya apuntados aquí tras la obtención de cuatro conjuntos no solapados.

Por otro lado, y desde la práctica estadística actual, se está verificando el importantísimo papel que viene desempeñando el análisis exploratorio, dentro del cual cabe incluir el análisis *cluster*. Es de destacar la relevancia de algunos

métodos robustos, en donde se ubicarían técnicas como K-Medoides y las basadas en el Principio Borroso, y que sin duda nos permiten, tal como hemos podido comprobar, una mayor riqueza de conclusiones, ya que si bien la técnica de K-Means puede constituir un clarificador primer paso en la búsqueda de configuraciones, la técnica K-Medoides aporta una mayor precisión diferencial entre grupos, siendo la aplicación del algoritmo propuesto por Dunn y Bezdek de innegable valor en trabajos de diagnóstico y seguimiento individual.

Finalmente, cabe poner de relieve que el hecho de distinguir entre posibles grupos naturales es un importante, y en algunas situaciones imprescindible, primer paso a la hora de generar hipótesis científicas con el fin de llevar a cabo posteriores análisis.

REFERENCIAS

Aldenderfer, M.S. & Blashfield, R.K. (1987). *Cluster Analysis*. Beverly Hills: Sage University Papers.

Allsopp, J.F. & Feldman, M.P. (1976). Item analysis of questionnaire of personality and antisocial behaviour in schoolboys. *Soc. Behav. Person., 2*, 184-190.

Ball, G.H. (1971). *Classification Analysis*. Stanford Research Institute. SRI Project 5533.

Ball, G.H. & Hall, D.J. (1965). A novel method of data analysis and pattern classification. Technical Report. Stanford Research Institute. California.

Bailey, T.A. & Dubes, R. (1982). Cluster validity profiles. *Pattern Recognition, 15*, 61-83.

Bayne, C.K., Beauchamp, J.J., Begovich, C.L. & Kane, V.E. (1980). Monte Carlo comparisons of selected clustering procedures. *Pattern Recognition, 12*, 51-62.

Bezdek, J.C. (1974). Cluster validity with fuzzy sets. *J. Cybernetics, 3*, 58-72.

Blashfield, R. & Morey, L. (1980). A comparison of four clustering methods using MMPI Monte Carlo data. *Applied Psychological Measurement, 4*, 57-64.

Capafons, A. (1986). *Estudio Psicométrico de un Cuestionario de Auto-control para una población española*. Tesis Doctoral no publicada. Universidad de Valencia.

Chambers, J.M. & Kleiner, B. (1982). Graphical Techniques for multivariate data and for clustering. In *Handbook for Statistics* (pp. 206-244). Volume 2. New York: North-Holland.

Church, R. (1978). Contrasts between facility location approaches and non-hierarchical cluster analysis. Paper presented at ORSA/TIMS Joint National Meeting. Los Angeles, California. Nov. 1978.

Clifford, H., & Stephenson, W. (1975). *An Introduction to Numerical Taxonomy*. New York: Academic Press.

Cormack, R.M. (1971). A review of classification (with discussion). *J. Roy. Statist. Soc., Ser. A., 134*, 321-367.

Diday, E, (1971). Une nouvelle méthode en classification automatique et reconnaisance des formes: La méthode des nuées dynamiques. *Rev. Statist. Appl., 19*, 19-33.

Dreger, R.M., Fuller, J. & Lemoine, R.L. (1988). Clustering Seven Data Sets by Means of Some or All of Seven Clustering Methods. *Multivariate Behavioral Research, 23*, 203-230.

Ducker, S.C., Williams, W.T. & Lance, G.N. (1965). Numerical classifications of the Pacific forms for Chlorodesmis (Chlorophyta). *Austral. J. Botany, 13*, 489-499.

Dunn, J.C. (1974). A fuzzy relative of the ISODATA Process and its use in detecting compact well-separated clusters. *J. Cybernetics, 3*, 32-57.

Dunn, D.M. & Landwehr, J.M. (1980). Analyzing clustering effects across time. *J. Amer. Statist. Assoc., 75*, 8-15.

Edelbrock, C. (1979). Comparing the accuracy of hierarchical clustering algorithms: the problem of classifying everybody. *Multivariate Behavioral Research, 14*, 367-384.

Edmonston, B. (1986, august). *Clustergram: A new graphical display for cluster partitions*. Paper presented at the Annual Meeting of the American Statistical Association, Chicago.

Estarelles, R. (1987). *Clima familiar y Auto-Concepto en la Adolescencia*. Tesis Doctoral no publicada, Universidad de Valencia.

Estarelles, R. & Ferrer, C. (1990). *Design and longitudinal analysis in the prevention strategies in drug abuse: Perceived environment and socialization influences*. Symposium on Statistical Methods for Evaluation of Intervention and Prevention Strategies. Atlanta. Georgia.

Estarelles, R., De la Fuente, I. & Olmedo, P. (1992, marzo). *Clustering three data sets: A comparison of*

hierarchical and non-hierarchical methods. Comunicación presentada al International Symposium on Multivariate Analysis and Its Applications, Hong Kong.
Everitt, B. (1977). *Cluster Analysis*. London: Heinemann Educational Books.
Fitts, W.H. (1964). *Tennessee Self-Concept Scale: Test Booklet*. Nashville, T.N.: Counselor Recordings and Test. Department of Mental Health.
Fowlkes, E.B., Gnanadesikan, R. & Kettenring, J.R. (1988). Variable selection in clustering. *J. Classification, 5*, 205-228.
Forgy, E.W. (1965). Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications (abstract). *Biometrics, 21*, 768-769.
Friedman, H.P. & Rubin, J. (1967). On some invariant criteria for grouping data. *J. Amer. Statistic. Assoc., 62*, 1159-1178.
Gale, N., Halperin, W.C. & Costanzo, C.M. (1984). Unclassed matrix shading and optimal ordering in hierarchical cluster analysis. *J. Classification, 1*, 75-92.
Gower, J.C. (1971). A general coefficient of similarity and some of its properties. *Biometrics, 27*, 857-871.
Hartigan, J. (1967). Representation of similiarity matrices by trees. *Journal of Statistical Computing and Computer Simulation, 4*, 187-213.
Hartigan, J. (1975). *Clustering Algorithms*. New York: Wiley-Interscience.
Jancey, R.C. (1966). Multidimensional group analysis. *Austral J. Botany, 14*, 127-130.
Kaufman, L. & Rousseeuw, P.J. (1987). Clustering by means of medoids. In *Statistical Data Analysis bases on the L_1 Norm* (pp. 405-416). Edited by Y. Dodge, Elsevier/Norht Holland. Amsterdam.
Kaufman, L. & Rousseeuw, P.J. (1989). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley-Interscience.
Klastorin, T.D. (1985). The p-median problem for cluster analysis: A comparative test using the mixture model approach. *Management Sci., 31*, 84-95.
Libert, G. & Roubens, M. (1982). Non-metric fuzzy clustering algorithms and their cluster validity. In *Approximate Reasoning in Decision Analysis*. Edited by M. Gupta and E. Sánchez. Amsterdam: Nort-Holland.
MacQuen, J. (1967). Some methods for classification and analysis of multivariate observations. *5th Berkeley Symp. Math. Statist. Prob.* Volume 1 (pp. 281-297). Edited by L. Le Cam and J. Neyman.
Massart, D.L., Plastria, F. & Kaufman, L. (1983). Non-hierarchical clustering with MASLOC. *Pattern Recognition, 16*, 507-516.
Milligan, G.W. (1980). An examination of the effect of six types of error perturbation of fifteen clustering algorithms. *Psychometrika, 45*, 325-342.
Milligan, G.W. & Cooper, M.C. (1987). Methodology Review: Clustering Methods. *Applied Psychological Measurement, 4*, 329-354.
Pardeck, J.T. (1991). A multiple regression analysis of family factors affecting the potential for alcoholism in college students. *Adolescence, 102*, 341-347.
Perris, C., Jacobsson, L., Linsdtrom, H., Von Knorring, L. & Perris, H. (1980). Development of a new inventory for assessing memories of parental rearing behaviour. *Acta Psychiat. Scand., 61*, 265-274.
Rao, M.R. (1971). Cluster Analysis and mathematical programming. *J. Amer. Statist. Assoc., 66*, 622-626.
Roubens, M. (1978). Pattern classification problems and fuzzy sets. *Fuzzy Sets and Systems, 1*, 239-253.
Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math., 20*, 53-65.
Scheibler, D. & Schneider, W. (1985). Monte Carlo Tests of the Accuracy of Cluster Analysis Algorithms: A comparison of Hierarchical and Non-hierarchical Methods. *Multivariate Behavioral Research, 20*, 283-304.
Sneath, P.R.A. & Sokal, R.R. (1973). *The principles and practice of numerical classification*. San Francisco: Freeman.
Spielberger, C., Gorsuch, R.L. & Lushene, R.E. (1970). *STAI Manual for the State-Trait Anxiety Inventory (Self-Evaluation Questionnaire)*. Palo Alto, California: Consulting Psycologists Press.
Thorlindsson, Th. & Vilhjalmsson, R. (1991). Factors related to cigarrette smoking and alcohol use among adolescents. *Adolescence, 102*, 399-418.
Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327-352.
Vinod, H. (1969). Integer programming and the theory of grouping. *J. Amer. Statist. Assoc., 64*, 506-517.
Wainer, H. (1983). On multivariate display. In *Recent Advances in Statistics* (pp. 469-508). Edited by M.H. Rizzi, J.S. Rustagi and D. Siegmund. New York: Academic Press.
Watts, W. & Wright, L.S. (1990). The relationship of alcohol, tobacco, marijuana, and other illegal drug use to delinquency among Mexican-American, Black and White adolescents males. *Adolescence, 97*, 171-181.
Wegman, E.J. (1985). *Hyperdimensional data analysis and the estimation of manifolds*. Centennial Meeting of the ISI. Amsterdam, 12-22 August. Contributed papers volume (pp. 205-206).
Wishart, D. (1978). *CLUSTAN User Manual, 3rd. ed. Program Library Unit. University of Edinburgh.*

APÉNDICE I

MÉTODOS BASADOS EN LA CONSTRUCCIÓN DE PUNTOS CENTRALES (K-MEANS)

En la técnica de K-Means el punto de partida es la división de los objetos en K subconjuntos, no vacíos, ya sea a través de la elección de los *k* primeros objetos de la muestra, ya sea teniendo en cuenta el conocimiento a priori del usuario, seleccionando los *k* objetos que estén mutuamente más alejados, o bien, partiendo de un *cluster* conjunto con todos los datos, y tras la selección de la variable con mayor dispersión, se obtendrán los dos primeros *clusters* tomando como punto de división su rango medio (éste es el procedimiento por defecto que utiliza el paquete BMDP KM).

En un segundo paso se obtendrá la media de la variable *j* en el *cluster v*, y el número de objetos que pertenezcan a ese *cluster n(v)*, calculándose a continuación la distancia entre el objeto *i* y el *cluster v* a partir de:

$$D(i,v) = [\sum_{j=1}^{p} (x(i,j) - \bar{x}(v,j)^2]^{1/2}$$

definiéndose el error para cada agrupamiento como:

$$E[P(n,K)] = \sum_{i=1}^{n} D(i,v_{(i)}]^2$$

En los pasos siguientes se procederá a la comparación entre los *clusters* obtenidos y otros posibles *clusters*, teniendo en cuenta la reducción en el error que ello pudiera conllevar, obteniéndose:

$$R_{v(i),v} = \frac{n(v)\ D(i,v)^2}{n(v) + 1} - \frac{n(v(i)\ D(I,v(i))^2}{n(v(I)) - 1}$$

Cuando el valor para todo $R_{v(i),v}$ sea positivo, el proceso se detendrá, ya que ello implicará que ningún cambio que se realice en la configuración podrá ofrecer un error más reducido.

APÉNDICE II

MÉTODOS BASADOS EN LA BÚSQUEDA DE K OBJETOS REPRESENTATIVOS (K-MEDOIDES)

Partiendo de que la distancia entre los objetos *i* y *j* sea d(i,j), la resolución del algoritmo, tal como fue propuesto por Kaufman y Rousseeuw (1989) vendrá determinado por:

a) *La elección de los k objetos representativos.* Para ello se definirá y_{ij} como una variable 0-1, siendo $y_{ij}=1$ sólo cuando el objeto *i* se haya seleccionado como objeto representativo.

b) *La ubicación de cada objeto j junto a uno de los k objetos representativos seleccionados.* Se definirá z_{ij} como una variable 0-1, siendo $z_{ij}=1$ sólo cuando el objeto *j* haya sido ubicado en aquel *cluster* cuyo objeto representativo sea *i*, teniendo un valor igual a cero para cualquier otra circunstancia. A partir de aquí la resolución tendrá que:

$$\text{Minimizar} \sum_{i=1}^{n} \sum_{j=1}^{n} d(i,j)z_{ij}$$

teniendo en cuenta que:

$$\sum_{i=1}^{n} z_{ij} = 1, \qquad j = 1, 2, \ldots, n$$

$$z_{ij} \leq y_i, \qquad i,j = 1, 2, \ldots, n$$

$$\sum_{i=1}^{n} y_i = k, \qquad k = n^{o} \text{ de } \textit{clusters}$$

$$y_i, z_{ij} \in (0,1), \qquad i, j = 1, 2, \ldots, n$$

de tal manera que la distancia entre un objeto *j* y su objeto representativo *i* será:

$$\sum_{i=1}^{n} d(i,j)z_{ij}$$

siendo, por tanto, la función a minimizar la distancia total representada al inicio.

Por tanto, el algoritmo, tal como se ha señalado, funciona en dos etapas. En la primera fase tendrá lugar la selección, paso a paso, de cada objeto representativo de los *k clusters* especificados, siendo el primer objeto elegido aquél para el cual la suma de las distancias al resto de objetos sea mínima. Una vez escogido un objeto representativo *i*, aún no seleccionado, la identificación de cada objeto implicará cuatro movimientos, que se sucederán en el orden siguiente:

a) Para cada objeto *j* se calculará la diferencia entre su distancia D_j con el objeto más similar a él, previamente seleccionado, y su distancia con el objeto *i* recientemente incluido como objeto representativo.

b) Si esta diferencia es positiva, el objeto *j* estará contribuyendo a la selección del objeto *i*, calculándose entonces:

$$C_{ij} = \max (D_j - d(j,i),0)$$

c) La mejora total debida a la elección del objeto *i* se obtendrá de:

$$\sum_{j} C_{ji}$$

d) Por último, se seleccionará el objeto *i* que maximice el sumatorio anterior.

En la segunda fase se intentará mejorar el conjunto de objetos representativos, para ello se tendrán en cuenta todos los pares de objetos (i,h) para los que el objeto *i* ha sido seleccionado pero el objeto *h* no. Se observará la modificación que tendría lugar si se eligiera el objeto *h*, pero no el objeto *i*. Para verificar el efecto de este cambio se llevarán a cabo los dos pasos siguientes:

1. Se tomará un objeto no seleccionado *j* y se calculará la contribución de C_{jih} al cambio:

a) Si el objeto *j* está más alejado de *i* y de *h* que de cualquiera de los otros objetos representativos, C_{jih} tendrá un valor igual a cero.

b) Si el objeto *j* no está más distanciado de *i* que de cualquier otro objeto representativo se tendrá en cuenta si:

— El objeto *j* está más próximo de *h* que del segundo objeto representativo más cercano,

$$d(j,h) < E_j$$

siendo entonces la contribución de *j* al cambio entre los objetos *i* y *h* la siguiente:

$$C_{jih} = d(j,h) - d(j,i)$$

En esta situación la contribución de C_{jih} podría ser tanto positiva como negativa, dependiendo del lugar ocupado por los objetos *j*, *h* e *i*. Únicamente si el objeto *j* está más próximo de *i* que de *h* la contribución será positiva, en cuyo caso el objeto *j* no estaría favoreciendo el cambio.

— El objeto *j* se encuentra al menos tan alejado de *h* como del segundo objeto representativo más próximo:

$$d(j,h) \geq E_j$$

en tal caso, la contribución del objeto *j* al cambio sería la siguiente:

$$C_{jih} = E_j - D_j$$

Ahora, la contribución de C_{jih} será siempre positiva.

c) Por último, si el objeto *j* se encuentra más alejado del objeto *i* que al menos uno de los objetos representativos, pero más próximo de *h* que de cualquier otro objeto representativo, la contribución de *j* al cambio sería:

$$C_{jih} = d(j,h) - D_j$$

2. Se calculará el cambio total sumando cada una de las contribuciones de C_{jih}

$$T_{ih} = \sum_j C_{jih}$$

y se pasará a decidir si se lleva a cabo dicha modificación, para ello:

3. Se seleccionará el par (i,h) con el fin de:

$$\underset{i,h}{\text{minimizar}}\ T_{ih}$$

4. Si el valor más reducido para T_{ih} es negativo, el cambio tendrá lugar, y se volverá al primer paso. En cambio, si el valor más pequeño para T_{ih} es positivo o cero el algoritmo se detendrá.

Dentro de las opciones gráficas de esta técnica se encuentran las representaciones de siluetas, las cuales incluyen un estadístico para valorar el acierto en la asignación de un objeto a un *cluster*. Este estadístico *s(i)* puede interpretarse de forma similar a como lo haríamos con un coeficiente de correlación. Este estadístico se define de tal manera que, partiendo de que *A* sea el *cluster* donde ha sido ubicado el objeto *i*, y definiendo *a(i)* como la distancia promedio de *i* respecto al resto de objetos en el *cluster A* (asumiendo que este *cluster* no está compuesto por un único objeto), habrá de verificarse que cualquier otro *cluster C*, distinto de *A*, es menos válido para la ubicación de *i*, teniendo en cuenta para ello la distancia *d(i,C)*, que será igual a la distancia promedio de *i* respecto a todos los objetos de *C*, y calculando a continuación este mismo valor para el resto de *clusters* diferentes de *A*. Por último, se tomará el valor más pequeño obtenido de manera que:

$$b(i) = \min_{C \neq A} d(i,C)$$

siendo entonces *B* el segundo *cluster* más próximo al objeto *i*, después de *A*. El valor de *s(i)* se calculará a partir de:

$$s(i) = \frac{b(i) - a(i)}{\max [a(i),b(i)]}$$

de lo que se deduce que $-1 \leq s(i) \leq 1$.

Los valores de *s(i)* para cada silueta se ofrecen en orden descendente, especificándose en el gráfico resultante: a) el número de *cluster* al que pertenece el objeto *i*; b) el número del *cluster* más próximo a *i*, distinto del *cluster A* donde ha sido ubicado; c) el valor de *s(i)* para cada objeto; d) la amplitud promedio de *s(i)* para cada *cluster*; y e) la amplitud promedio de *s(i)* para el conjunto total de datos, que se denominará como *s(k)*, y cuyo valor bien puede utilizarse como criterio para seleccionar el número óptimo de *clusters* a especificar.

APÉNDICE III

MÉTODOS BASADOS EN EL PRINCIPIO BORROSO (ALGORITMO DE DUNN Y BEZDEK, 1974)

El algoritmo buscará la minimización de la siguiente función objetiva:

$$C = \sum_{v=1}^{k} \frac{\sum_{i,j=1}^{n} u_{iv}^2 u_{jv}^2 d(i,j)}{2\sum_{j=1}^{n} u_{jv}^2}$$

en la que *d(i,j)* representa la distancia entre los objetos *i* y *j*, mientras que u_{iv} se corresponde con la pertenencia del objeto *i* al *cluster v*. Ahora bien, las funciones de pertenencia están sujetas a las siguientes limitaciones:

$$u_{iv} \geq 0 \qquad \text{para } i = 1,...,n;\ v = 1,...,k$$
$$\sum_{v} u_{iv} = 1 \qquad \text{para } i = 1,...,n$$

lo cual indica que la pertenencia nunca puede ser negativa, así como que la pertenencia total de cada objeto ha de ser constante (está normalizada a 1), aunque se distribuirá a lo largo de los diferentes *clusters*. El coeficiente de partición de Dunn nos ofrece una medida de la consistencia del agrupamiento, obteniéndose a partir de:

$$F_k = \sum_{i=1}^{n} \sum_{v=1}^{k} u_{iv}^2/n$$

Su valor mínimo será 1/k, siendo éste el valor que corresponderá a un agrupamiento completamente borroso, aunque valores extremos (1 o 0) podrán surgir de una partición. La versión normalizada de este coeficiente, que siempre oscilará entre 0 y 1, independientemente del valor que se haya tomado para *k*, se calculará a partir de:

$$F'_k = \frac{F_k - (1/k)}{1 - (1/k)} = \frac{kF_k - 1}{k - 1}$$

Entre las opciones gráficas más usuales para este acercamiento se encuentra la representación a través de siluetas comentada anteriormente.