Análisis de Conglomerados o Cluster

1.- PLANTEAMIENTO DEL PROBLEMA

Sean $X_1,...,X_p$, p variables cuantitativas observadas en n objetos o individuos.

Sea x_{ij} = la realización de la variable X_j en el i-ésimo objeto i=1,...,n; j=1,...,p.

- El objetivo del Análisis Cluster es obtener grupos de objetos de forma que, por un lado, los objetos pertenecientes a un mismo grupo sean muy semejantes entre sí y, por el otro, los objetos pertenecientes a grupos diferentes tengan un comportamiento distinto con respecto a las variables analizadas.

- Es una técnica exploratoria puesto que la mayor parte de las veces no utiliza ningún tipo de modelo estadístico para llevar a cabo el proceso de clasificación.

- Conviene estar siempre alerta ante el peligro de obtener, como resultado del análisis, no una *clasificación* de los datos sino una *disección* de los mismos en distintos grupos que sólo existen en la memoria del computador. ***El conocimiento que el analista tenga acerca del problema decidirá cuáles de grupos obtenidos son significativos y cuáles no***.

<u>Ejemplo 1</u>: Clasificación de países de la UE con datos binarios

Se tiene la situación de 6 países europeos en 1996 con respecto a los 4 criterios exigidos por la UE para entrar en la Unión Monetaria: Inflación, Interés, Déficit Público y Deuda Pública y vienen dados en la tabla siguiente:

País	Inflación	Interés	Déficit	Deuda
Alemania	1	1	1	0
España	1	1	1	0
Francia	1	1	1	1
Grecia	0	0	0	0
Italia	1	1	0	0
Reino Unido	1	1	0	1

Este es un ejemplo en el que todas las variables son binarias de forma que, este caso 1 significa que el país sí satisfacía el criterio exigido y 0 que no lo satisfacía.

<u>Ejemplo 2</u>: Características socioeconómicas en diversos países

Se tiene datos sobre diversas variables económicas, sanitarias y demográficas correspondientes a 102 países del mundo en el año 1995.

Variable	Significado
POB	Logaritmo de la Población
DENS	Logaritmo de la Densidad
ESPF	Logaritmo de 83-Esperanza de vida Femenina
ESPM	Logaritmo de 78 - Esperanza de vida masculina
ALF	Logaritmo de 101-Tasa de Alfabetización
MINF	Logaritmo de la Tasa de Mortalidad Infantil
PIBCA	Logaritmo del PIB per cápita
NACDEF	Logaritmo de Nacimientos/Defunciones
FERT	Logaritmo del número medio de hijos por mujer

2. MEDIDAS DE PROXIMIDAD Y DE DISTANCIA

Una vez establecidas las variables y los objetos a clasificar el siguiente paso consiste en establecer una medida de proximidad o de distancia entre ellos que cuantifique el grado de similaridad entre cada par de objetos.

- Las **medidas de proximidad, similitud o semejanza** miden el grado de semejanza entre dos objetos de forma que, cuanto mayor (resp. menor) es su valor, mayor (resp. menor) es el grado de similaridad existente entre ellos y con más (resp. menos) probabilidad los métodos de clasificación tenderán a ponerlos en el mismo grupo.

- Las **medidas de disimilitud, desemejanza o distancia** miden la distancia entre dos objetos de forma que, cuanto mayor (resp. menor) sea su valor, más (resp. menos) diferentes son los objetos y menor (resp. mayor) la probabilidad de que los métodos de clasificación los pongan en el mismo grupo.

En la literatura existen multitud de medidas de semejanza y de distancia dependiendo del tipo de variables y datos considerados. Veremos algunas de las más utilizadas.

Siguiendo el manual de SPSS podemos distinguir, esencialmente, los siguientes tipos de datos:

2.1 Tipos de datos

1) **De intervalo**: se trata de una matriz objetosxvariables en donde todas las variables son cuantitativas, medidas en escala intervalo o razón.

2) **Frecuencias:** las variables analizadas son categóricas de forma que, por filas, tenemos objetos o categorías de objetos y, por columnas, las variables con sus diferentes categorías. En el interior de la tabla aparecen frecuencias.

3) **Datos binarios:** se trata de una matriz objetosxvariables pero en la que las variables analizadas son binarias de forma que 0 indica la ausencia de una característica y 1 su presencia.

2.2 Medidas de proximidad

a) Medidas para variables cuantitativas

1) **Coeficiente de congruencia**

$$c_{rs} = \frac{\sum_{j=1}^{p} x_{rj} x_{sj}}{\sqrt{\sum_{j=1}^{p} x_{rj}^2} \sqrt{\sum_{j=1}^{p} x_{sj}^2}}$$

que es el coseno del ángulo que forman los vectores $(x_{r1},...,x_{rp})'$ y $(x_{s1},...,x_{sp})'$.

2) **Coeficiente de correlación**

$$r_{rs} = \frac{\sum_{j=1}^{p}\left(x_{rj}-\overline{x}_r\right)\left(x_{sj}-\overline{x}_s\right)}{\sqrt{\sum_{j=1}^{p}\left(x_{rj}-\overline{x}_r\right)^2}\sqrt{\sum_{j=1}^{p}\left(x_{sj}-\overline{x}_s\right)^2}}$$

donde $\overline{x}_r = \frac{\sum_{j=1}^{p} x_{rj}}{p}$ y $\overline{x}_s = \frac{\sum_{j=1}^{p} x_{sj}}{p}$.

Si los objetos r y s son variables, r_{rs} mide el grado de asociación lineal existente entre ambas.

- Conviene observar, además, que tanto c_{rs} como r_{rs} toman valores comprendidos entre -1 y 1 pudiendo tomar, por lo tanto, valores negativos. Dado que en algunos casos (por ejemplo, si los objetos a clasificar son variables) los valores negativos cercanos a -1 pueden implicar fuerte semejanza entre los objetos clasificados, conviene utilizar como medida de semejanza sus valores absolutos.

b) Medidas para datos binarios

En este caso se construyen, para cada par de objetos r y s, tablas de contingencia de la forma:

Objeto s \Objeto r	0	1
0	a	b
1	c	d

donde a = número de variables en las que los objetos r y s toman el valor 0, etc y p = a+b+c+d.

Utilizando dichas tablas algunas de las medidas de semejanza más utilizadas son:

- ✓ **Coeficiente de Jacard**: $\frac{d}{b+c+d}$
- ✓ **Coeficiente de acuerdo simple**: $\frac{a+d}{p}$

Ambas toman valores entre 0 y 1 y miden, en tanto por uno, el porcentaje de acuerdo en los valores tomados en las p variables, existente entre los dos objetos. Difieren en el papel dado a los acuerdos en 0. El coeficiente de Jacard no los tiene en cuenta y el de acuerdo simple sí.

c) Medidas para datos nominales y ordinales

Una generalización de las medidas anteriores viene dada por la expresión:

$$s_{rs} = \sum_{k=1}^{p} S_{rsk}$$

donde s_{rsk} es la contribución de la variable k-ésima a la semejanza total. Dicha contribución suele ser de la forma $1-d_{rsk}$ donde d_{rsk} es una distancia que suele tener la forma $\delta_{k\ell m}$ siendo ℓ el valor del estado de la variable X_k en el r-ésimo objeto y m el del s-ésimo objeto.

En variables nominales suele utilizarse $\delta_{k\ell m} = 1$ si $\ell = m$ y 0 en caso contrario. En variables ordinales suele utilizarse medidas de la forma $|\ell - m|^r$ con $r>0$.

2.3 Medidas de distancia

a) Medidas para variables cuantitativas

1) **Distancia euclídea y distancia euclídea al cuadrado**

$$\sqrt{\sum_{j=1}^{p}\left(x_{rj}-x_{sj}\right)^2} \text{ y } \sum_{j=1}^{p}\left(x_{rj}-x_{sj}\right)^2$$

2) **Distancia métrica de Chebychev**: $\max_i |x_{ri}-x_{si}|$

3) **Distancia de Manhattan**: $\sum_{i=1}^{p}\left|x_{ri}-x_{si}\right|$

4) **Distancia de Minkowski**: $\sqrt[q]{\sum_{i=1}^{p}\left(x_{ri}-x_{si}\right)^q}$ con $q \in \mathbf{N}$.

- Las tres primeras medidas son variaciones de la distancia de Minkowski con $q=2,\infty$ y 1, respectivamente. Cuanto mayor es q más énfasis se le da a las diferencias en cada variable.
- Todas estas distancias no son invariantes a cambios de escala por lo que se aconseja estandarizar los datos si las unidades de medida de las variables no son comparables.

- Además no tienen en cuenta las relaciones existentes entre las variables. Si se quieren tomar en cuenta se aconseja utilizar la **distancia de Mahalanobis** que viene dada por la forma cuadrática:

$$\left(x_r - x_s\right)' S^{-1} \left(x_r - x_s\right)$$

donde $x_r = (x_{r1},...,x_{rp})'$ y $x_s = (x_{s1},...,x_{sp})'$

b) Medidas para tablas de frecuencias:

Generalmente están basadas en la χ^2 de Pearson. Las más utilizadas son:

$$\chi^2 = \sqrt{\sum_{i=1}^{p} \frac{(x_{ri} - E(x_{ri}))^2}{E(x_{ri})} + \sum_{i=1}^{p} \frac{(x_{si} - E(x_{si}))^2}{E(x_{si})}}$$

$$\varphi^2 = \sqrt{\frac{\sum_{i=1}^{p} \frac{(x_{ri} - E(x_{ri}))^2}{E(x_{ri})} + \sum_{i=1}^{p} \frac{(x_{si} - E(x_{si}))^2}{E(x_{si})}}{N}}$$

donde $E(x_{ri}) = \frac{x_{r.} x_{.i}}{N}$ con $x_{r.} = \sum_{i=1}^{p} x_{ri}$ y $x_{.i} = x_{ri} + x_{si}$ es el valor esperado de la frecuencia x_{ri} si hay independencia entre los individuos r y s, y las categorías 1,...,p de las variables y $N = x_{r.} + x_{s.}$ es el total de observaciones.

- La diferencia entre ambas medidas radica en la división por N en el caso de φ^2 para paliar la dependencia que tiene la χ^2 de Pearson respecto a N.

c) Medidas para datos binarios

Las más utilizadas son:

- ✓ **Distancia euclídea al cuadrado:** b+c

- ✓ **Lance y Williams:** $\frac{b+c}{2d+b+c}$

Esta última ignora los acuerdos en 0.

d) Medidas para datos de tipo mixto

- Si en la base de datos existen diferentes tipos de variables: binarias, categóricas, ordinales, cuantitativas no existe una solución universal al problema de cómo combinarlas para construir una medida de distancia.

- Algunas soluciones sugeridas son:

  - Expresar todas las variables en una escala común, habitualmente binaria, transformando el problema en uno de los ya contemplados anteriormente. Esto tiene costos en términos de pérdida de información.
  - Combinar medidas con pesos de ponderación mediante expresiones de la forma:

$$d_{ij} = \frac{\sum_{k=1}^{p} w_{ijk} d_{ijk}}{\sum_{k=1}^{p} w_{ijk}}$$

    donde d_{ijk} es la distancia entre los objetos i y j en la k-ésima variable y w_{ijk} = 0 ó 1 dependiendo de si la comparación entre i y j es válida en la k-ésima variable.
  - Realizar análisis por separado utilizando variables del mismo tipo y utilizar el resto de las variables como instrumentos para interpretar los resultados obtenidos.

Ejemplo 1

En este caso todas las variables son binarias simétricas y podemos utilizar como medida de distancia la distancia euclídea al cuadrado. La matriz de distancias obtenida viene dada por:

	Al	Es	Fr	Gr	It	RU
Al	0	0	1	3	1	2
Es		0	1	3	1	2
Fr			0	4	2	1
Gr				0	2	3
It					0	1
RU						0

Así, por ejemplo, la distancia entre España y Francia es 1 puesto que solamente difieren en un criterio: el de la deuda pública que Francia satisfacía y España no.

Ejemplo 2

En este caso todas las variables son cuantitativas pero medidas en diferentes unidades. Por esta razón utilizaremos la distancia euclídea pero con los datos estandarizados previamente.

3. MÉTODOS DE CLASIFICACIÓN

Entre los muchos tipos de métodos que existen en la literatura cabe destacar los siguientes:

- **Jerárquicos**: en cada paso del algoritmo sólo un objeto cambia de grupo y los grupos están anidados en los de pasos anteriores. Si un objeto ha sido asignado a un grupo ya no cambia más de grupo.

- **Repartición:** tienen un número de grupos, g fijado de antemano, como objetivo y agrupa los objetos para obtener los g grupos. Comienzan con una solución inicial y los objetos se reagrupan de acuerdo con algún criterio de optimalidad.

- **Métodos tipo Q**: son similares al análisis factorial y utilizan como información la matriz **XX'** utilizando las variables como objetos y los objetos como variables.

- **Procedimientos de localización de modas:** agrupan los objetos en torno a modas con el fin de obtener zonas de gran densidad de objetos separadas unas de otras por zonas de poca densidad.

- **Métodos que permiten solapamiento:** permiten que los grupos tengan elementos en común.

.

3.1 Métodos jerárquicos

- Se caracterizan porque en cada paso del algoritmo sólo un objeto cambia de grupo y los grupos están anidados en los de pasos anteriores. Si un objeto ha sido asignado a un grupo ya no cambia más de grupo.

- Pueden ser, a su vez de dos tipos: aglomerativos y divisivos:

    - *Los métodos aglomerativos* comienzan con n clusters de un objeto cada uno. En cada paso del algoritmo se recalculan las distancias entre los grupos existentes y se unen los 2 grupos más similares o menos disimilares. El algoritmo acaba con 1 cluster conteniendo todos los elementos.

    - *Los métodos divisivos* comienzan con 1 cluster que engloba a todos los elementos. En cada paso del algoritmo se divide el grupo más heterogéneo. El algoritmo acaba con n clusters de un elemento cada uno.

- Para determinar qué grupos se unen o dividen se utiliza una función objetivo o criterio que recibe el nombre de *enlace.*

3.1.1 Tipos de enlace

Se utilizan con los métodos aglomerativos y proporcionan diversos criterios para determinar, en cada paso del algoritmo, qué grupos se deben unir.

- Enlace simple o vecino más próximo

Mide la proximidad entre dos grupos calculando la distancia entre sus objetos más próximos o la similitud entre sus objetos más semejantes.

- Enlace completo o vecino más alejado

Mide la proximidad entre dos grupos calculando la distancia entre sus objetos más lejanos o la similitud entre sus objetos menos semejantes

- ***Enlace medio entre grupos***

Mide la proximidad entre dos grupos calculando la media de las distancias entre objetos de ambos grupos o la media de las similitudes entre objetos de ambos grupos. Así, por ejemplo, si se utilizan distancias, la distancia entre los grupos r y s vendría dada por:

$$\frac{1}{n_r n_s} \sum_{j \in r} \sum_{k \in s} d(j,k)$$

donde d(j,k) = distancia entre los objetos j y k y n_r, n_s son los tamaños de los grupos r y s, respectivamente.

- ***Enlace medio dentro de los grupos***

Mide la proximidad entre dos grupos con la distancia media existente entre los miembros del grupo unión de los dos grupos. Así, por ejemplo, si se trata de distancias, la distancia entre los grupos r y s vendría dada por:

$$\frac{1}{C^2_{n_r+n_s}} \sum_{(j,k) \in r \cup s} d(j,k)$$

3.1.2 Métodos del centroide y de la mediana

Ambos métodos miden la proximidad entre dos grupos calculando la distancia entre sus centroides

$$d_{rs}^2 = \sum_{j=1}^{p} \left(\overline{x}_{rj} - \overline{x}_{sj} \right)^2$$

donde $\overline{x}_{rj}$ y $\overline{x}_{sj}$ son las medias de la variable X_j en los grupos r y s, respectivamente.

Los dos métodos difieren en la forma de calcular los centroides:

- El método del centroide utiliza las medias de todas las variables de forma que las coordenadas del centroide del grupo r = s $\cup$ t vendrán dadas por:

$$\overline{x}_{rj} = \frac{1}{n_r} \sum_{m=1}^{n_r} x_{rjm} = \frac{n_s}{n_s + n_t} \overline{x}_{sj} + \frac{n_t}{n_s + n_t} \overline{x}_{tj} \quad j = 1,\ldots,p$$

- En el método de la mediana el nuevo centroide es la media de los centroides de los grupos que se unen

$$\overline{x}_{rj} = \frac{1}{2} \overline{x}_{sj} + \frac{1}{2} \overline{x}_{tj}$$

3.1.3 Método de Ward

El método busca minimizar $\sum_r SSW_r$ donde SSW_r corresponde a las sumas de cuadrados intragrupo para cada grupo r, que viene dada por:

$$SSW_r = \sum_{m=1}^{n_r} \sum_{j=1}^{p} \left(x_{rjm} - \overline{x}_{rj} \right)^2$$

donde x_{rjm} denota el valor de la variable X_j en el m-ésimo elemento del grupo r.

En cada paso del algoritmo une los grupos r y s que minimizan:

$$SSW_t - SSW_r - SSW_s = \frac{n_r n_s}{n_r + n_s} d_{rs}^2$$

con $t = r \cup s$ y d_{rs}^2 la distancia entre los centroides de r y s.

3.1.4 Comparación de los diversos métodos aglomerativos

1) El enlace simple conduce a clusters encadenados

2) El enlace completo conduce a clusters compactos

3) El enlace completo es menos sensible a outliers que el enlace simple

4) El método de Ward y el método del enlace medio son los menos sensibles a outliers

5) El método de Ward tiene tendencia a formar clusters más compactos y de igual tamaño y forma en comparación con el enlace medio

6) Todos los métodos salvo el método del centroide satisfacen la desigualdad ultramétrica:

$$d_{ut} \leq \min \{d_{ur}, d_{us}\} \qquad t = r \cup s$$

3.1.5 Elección del número de grupos

- Existen diversos métodos de determinación del número de grupos: algunos están basados en intentan reconstruir la matriz de distancias original, otros en los coeficientes de concordancia de Kendall y otros realizan análisis de la varianza entre los grupos obtenidos. No existe un criterio universalmente aceptado.

- Dado que la mayor parte de los paquetes estadísticos proporciona las distancias de aglomeración, es decir, las distancias a las que se forma cada grupo, una forma de determinar el número de grupos consiste en localizar en qué iteraciones del método utilizado dichas distancias pegan grandes saltos.

- Utilizando dichas distancias se pueden utilizar criterios como el *criterio de Mojena* que determina el primer $s \in \mathbf{N}$ tal que $\alpha_{s+1} > \overline{\alpha} + ks_{\alpha}$ si se utilizan distancias y < si son similitudes donde $\{\alpha_j; < j=1,...,n-1\}$ son las distancias de aglomeración, $\overline{\alpha}$, s_{α} su media y su desviación típica respectivamente y k una constante entre 2.5 y 3.5.

Ejemplo 1

Los resultados de aplicar un método jerárquico aglomerativo con enlace completo utilizando el paquete estadístico SPSS se muestran a continuación:

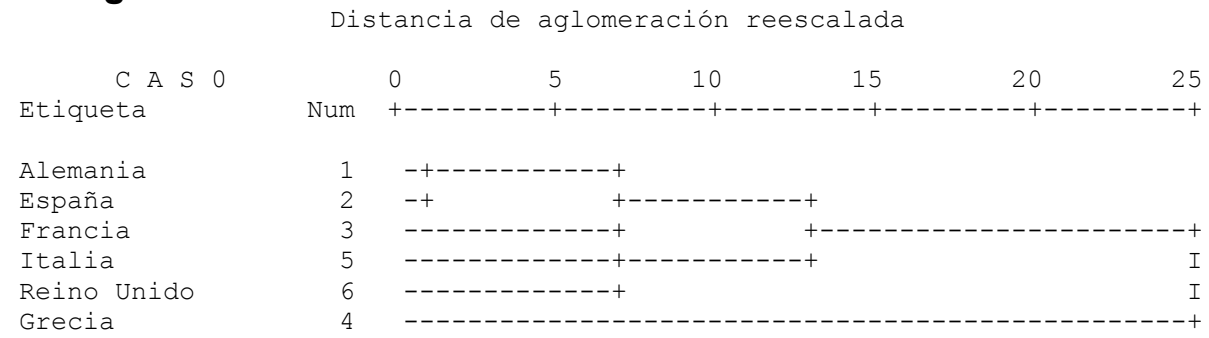
Historial de conglomeración

Etapa	Conglomerado que se combina		Coeficientes	Etapa en la que el conglomerado aparece por primera vez		Próxima etapa
	Conglomerado 1	Conglomerado 2		Conglomerado 1	Conglomerado 2	
1	1	2	0	0	0	3
2	5	6	1	0	0	4
3	1	3	1	1	0	4
4	1	5	2	3	2	5
5	1	4	4	4	0	0

Diagrama de témpanos vertical

Número de conglomerados	Caso										
	4:Grecia		6:Reino Unido		5:Italia		3:Francia		2:España		1:Alemania
1	X	X	X	X	X	X	X	X	X	X	X
2	X		X	X	X	X	X	X	X	X	X
3	X		X	X	X		X	X	X	X	X
4	X		X	X	X		X		X	X	X
5	X		X		X		X		X	X	X

Dendograma



El historial de aglomeración muestra las distancias de aglomeración y los grupos que se han ido formando al aplicar el algoritmo.

El diagrama de témpanos y el dendograma dan dicha información de forma gráfica. Así, en el primer paso del algoritmo se unieron Alemania y España a una distancia de aglomeración igual a 0. Posteriormente, a dicho grupo, se unió Francia e Italia y Reino Unido formaron otro grupo, todo ello a una distancia de aglomeración igual a 1. Estos dos grupos se unieron formando un único grupo a una distancia de aglomeración igual a 2. Finalmente Grecia se unió a todos los demás países a una distancia de aglomeración igual a 4, la máxima posible.

Si tomamos como punto de corte 1 nos quedaríamos con 3 grupos: {España, Alemania y Francia}, {Italia, Reino Unido} y {Grecia}. Estos grupos están formados por países que difieren entre sí en a lo más un criterio.

Ejemplo 2

En el gráfico 1 se muestran las distancias de aglomeración del algoritmo jerárquico aglomerativo tomando como función de enlace, el enlace intergrupos y utilizando el paquete estadístico SPSS

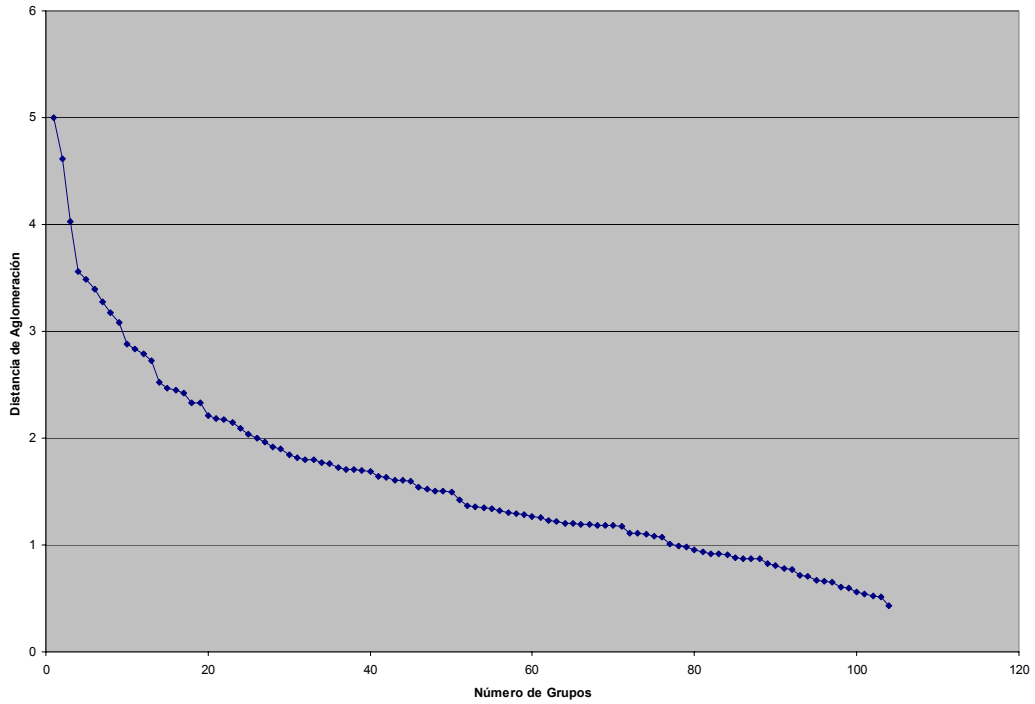


Gráfico 1: Distancias de aglomeración

Se observa que los mayores saltos se dan cuando el algoritmo pasa de 4 a 3, 3 a 2 y 2 a 1 grupo. El criterio de Mojena aplicado con k=2.5 da una distancia de corte igual a 3.83 y selecciona un número de grupos igual a 4. Por todas estas razones tomamos como número de grupos 4.

3.2 Método de las k-medias

- Este tipo de método es conveniente utilizarlo cuando los datos a clasificar son muchos y/o para refinar una clasificación obtenida utilizando un método jerárquico. Supone que el número de grupos es conocido a priori.

- Existen varias formas de implementarlo pero todas ellas siguen, básicamente, los siguientes pasos:
    1) Se seleccionan k centroides o semillas donde k es el número de grupos deseado
    2) Se asigna cada observación al grupo cuya semilla es la más cercana
    3) Se calculan los puntos semillas o centroides de cada grupo
    4) Se iteran los pasos 2) y 3) hasta que se satisfaga un criterio de parada como, por ejemplo, los puntos semillas apenas cambian o los grupos obtenidos en dos iteraciones consecutivas son los mismos.

- El método suele ser muy sensible a la solución inicial dada por lo que es conveniente utilizar una que sea buena. Una forma de construirla es mediante una clasificación obtenida por un algoritmo jerárquico

Ejemplo 2

Los resultados de aplicar el algoritmo de las k-medias implementado en SPSS, con un número de grupos igual a 4 y tomando como punto de partida los centroides de los grupos obtenidos anteriormente vienen dados por las siguientes tablas y gráficos. El algoritmo converge en 10 iteraciones y obtiene 4 grupos de tamaños 24, 39, 1 y 41 países respectivamente.

Historial de iteraciones[a]

	Cambio en los centros de los conglomerados			
Iteración	1	2	3	4
1	,592	,109	1,036E-07	,172
2	,487	6,262E-02	,000	,125
3	,214	,000	,000	4,648E-02
4	,231	,000	,000	5,287E-02
5	,225	6,193E-02	,000	3,981E-02
6	,306	5,276E-02	,000	9,411E-02
7	,235	,000	,000	9,347E-02
8	,250	6,932E-02	,000	,115
9	,227	7,083E-02	,000	,121
10	,305	,174	,000	5,141E-02

a. Las iteraciones se han detenido porque se ha llevado a cabo el número máximo de iteraciones. Las iteraciones no han convergido. La distancia máxima en la que han cambiado los centros es ,172. La iteración actual es 10. La distancia mínima entre los centros iniciales es 3,007.

Grupo obtenidos

PAIS	GRUPO	DISTANCIA
Venezuela	1	1,10992
Ecuador	1	1,17341
Malasia	1	1,19941
Panamá	1	1,24843
Acerbaján	1	1,27096
Colombia	1	1,31659
Armenia	1	1,33676
Chile	1	1,36857
Rep. Dominicana	1	1,49939
Turquía	1	1,57329
Uzbekistán	1	1,65333
Líbano	1	1,67326
México	1	1,69396
Tailandia	1	1,81748
El Salvador	1	1,81842
Corea del Norte	1	1,82812
Paraguay	1	1,88032
Jordania	1	1,90393
Argentina	1	2,05071
Emiratos Árabes	1	2,26097
Corea del Sur	1	2,28927
Costa Rica	1	2,56727
Kuwait	1	2,5803
Bahrein	1	2,78161
Austria	2	0,84751
Irlanda	2	1,02262
Dinamarca	2	1,03776
Croacia	2	1,17118
Bélgica	2	1,25977
Finlandia	2	1,29839
Grecia	2	1,39139
Polonia	2	1,39569
España	2	1,41288
Lituania	2	1,42745
Hungía	2	1,43235

Portugal	2	1,45946
Bielorusia	2	1,47973
Gran Bretaña	2	1,53294
Bulgaria	2	1,53866
Georgia	2	1,62389
Nueva Zelanda	2	1,68732
Suecia	2	1,69381
Rumanía	2	1,69529
Italia	2	1,71363
Alemania	2	1,71408
Países Bajos	2	1,77523
Noruega	2	1,83862
Uruguay	2	1,93886
Cuba	2	1,94022
Francia	2	1,98214
Estonia	2	2,01381
Letonia	2	2,02654
Suiza	2	2,04078
Ucrania	2	2,19731
Estados Unidos	2	2,30185
Canadá	2	2,60291
Australia	2	2,69585
Israel	2	2,71955
Rusia	2	2,89912
Japón	2	3,11629
Barbados	2	3,15042
Singapur	2	3,48935
Hong Kong	2	3,75342
Islandia	3	0,0000
Camerún	4	0,57933
Senegal	4	0,72504
Kenia	4	0,81205
Egipto	4	1,01448
Guatemala	4	1,07179
Camboya	4	1,17287
Marruecos	4	1,34473
Burkina Faso	4	1,35581
Nicaragua	4	1,40744

Tanzania	4	1,44743
Irán	4	1,45366
Nigeria	4	1,47222
Iraq	4	1,50176
Sudáfrica	4	1,51414
Perú	4	1,53181
Liberia	4	1,54648
Bolivia	4	1,56759
Uganda	4	1,57074
Honduras	4	1,58019
Zambia	4	1,58128
Etiopía	4	1,68095
Pakistán	4	1,69868
Afganistán	4	1,73597
Somalia	4	1,78696
Siria	4	1,86294
Haití	4	1,86689
Burundi	4	1,99972
Filipinas	4	2,03681
Indonesia	4	2,12085
Ruanda	4	2,13195
Vietnam	4	2,14496
Gambia	4	2,31622
Brasil	4	2,31901
Rep. C. Africana	4	2,41386
Arabia Saudí	4	2,4842
Bangladesh	4	2,5958
Libia	4	2,77066
Gabón	4	2,94421
India	4	2,96665
Botswana	4	2,96857
China	4	3,63459

Distancias entre los centros de los conglomerados finales

Conglomerado	1	2	3	4
1		3,038	5,466	2,594
2	3,038		4,233	4,967
3	5,466	4,233		7,460
4	2,594	4,967	7,460	

4. INTERPRETACION DE LOS RESULTADOS

Interpretar la clasificación obtenida por un Análisis Cluster requiere, en primer lugar, un conocimiento suficiente del problema analizado. Hay que estar abierto a la posibilidad de que no todos los grupos obtenidos tienen por qué ser significativos. Algunas ideas que pueden ser útiles en la interpretación de los resultados son las siguientes:

- ✓ Realizar ANOVAS y MANOVAS para ver qué grupos son significativamente distintos y en qué variables lo son.
- ✓ Realizar Análisis Discriminantes.
- ✓ Realizar un Análisis Factorial o de Componentes Principales para representar, gráficamente los grupos obtenidos y observar las diferencias existentes entre ellos.
- ✓ Calcular perfiles medios por grupos y compararlos.

Ejemplo 2

En la tabla siguiente se muestran los resultados de aplicar un ANOVA para cada una de las variables analizadas. Se observa que existen diferencias significativas en todas las variables al 1 y al 5% con excepción de las variables POB y DENS en las que solamente existen diferencias al 5%.

ANOVA

	Conglomerado		Error			
	Media cuadrática	gl	Media cuadrática	gl	F	Sig.
Puntua(POB)	3,563	3	,941	101	3,786	,013
Puntua(DENS)	3,086	3	,929	101	3,321	,023
Puntua(ESPF)	26,744	3	,263	101	101,702	,000
Puntua(ESPM)	23,077	3	,375	101	61,560	,000
Puntua(ALF)	26,760	3	,254	101	105,301	,000
Puntua(MINF)	28,487	3	,180	101	157,954	,000
Puntua(PIBCA)	23,533	3	,355	101	66,341	,000
Puntua(NACDE)	23,794	3	,303	101	78,483	,000
Puntua(FERT)	26,351	3	,238	101	110,829	,000

Las pruebas F sólo se deben utilizar con una finalidad descriptiva puesto que los conglomerados han sido elegidos para maximizar las diferencias entre los casos en diferentes conglomerados. Los niveles críticos no son corregidos, por lo que no pueden interpretarse como pruebas de la hipótesis de que los centros de los conglomerados son iguales.

Los dos gráficos siguientes muestran los perfiles medio de cada grupo y los diagramas de cajas de las variables analizadas para cada uno de los grupos. Se observa que los países de los grupos 1 y 4 poseen una menor renta per cápita y peores indicadores los índices de alfabetización, mortalidad y esperanza de vida así como una mayor fertilidad y natalidad que la de los países de los grupos 2 y 3 siendo estas diferencias más acusadas en los países del grupo 4 que la de los grupo 1. También queda de manifiesto el carácter atípico de Islandia debido a su baja natalidad, mortalidad infantil, población y densidad y su alta alfabetización, esperanza de vida.

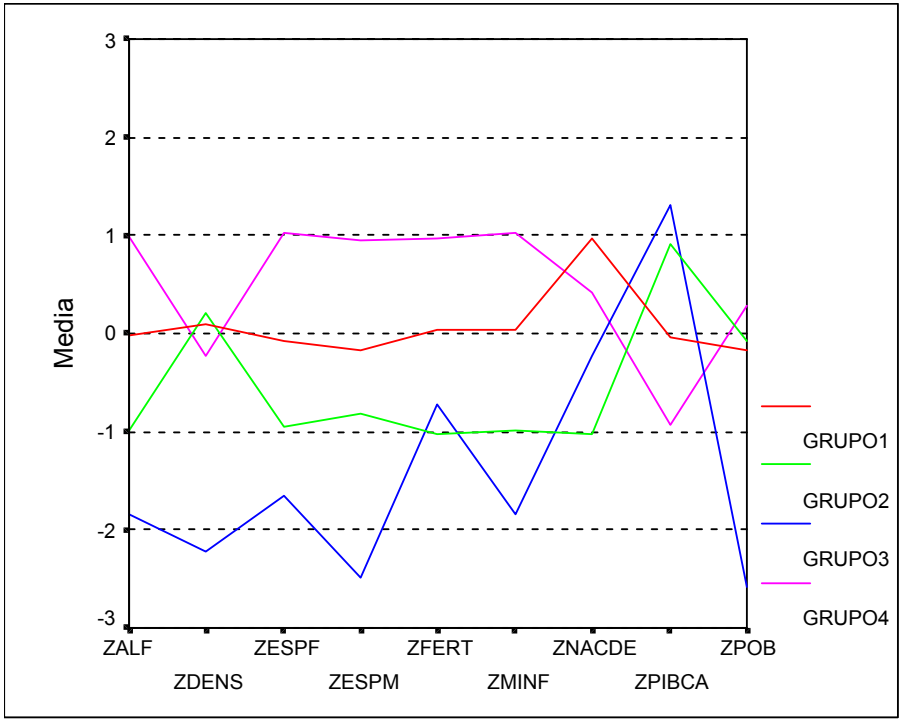


Gráfico 1: Perfiles medios de cada grupo

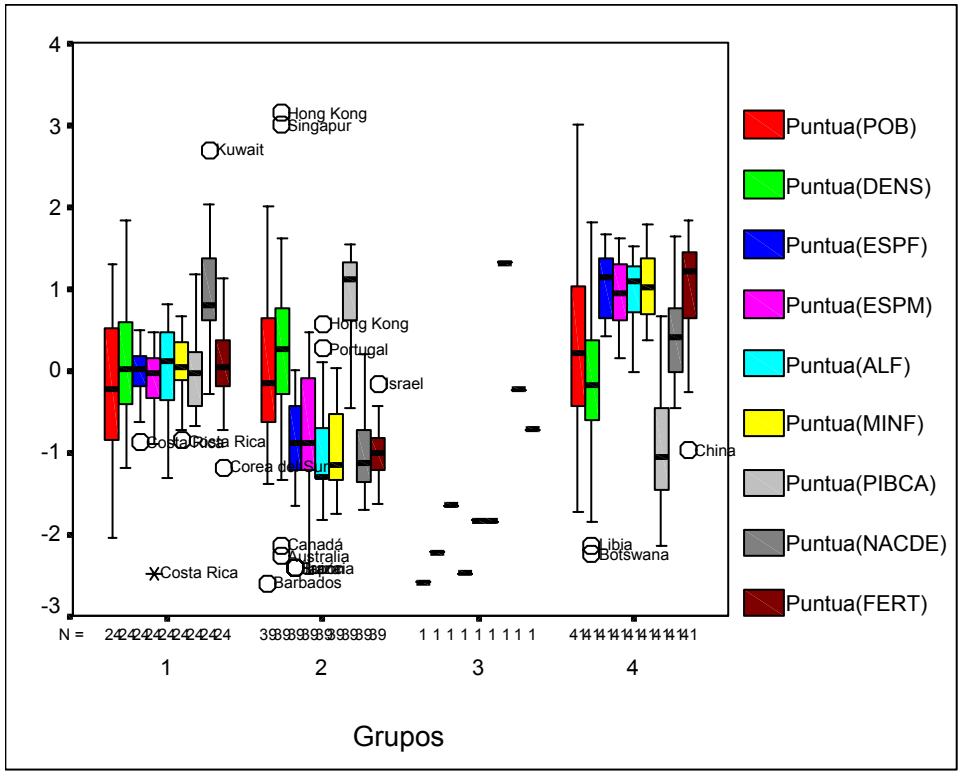


Gráfico 2: Diagrama de cajas correspondiente a cada grupo

5. VALIDACIÓN DE LA SOLUCIÓN

Una vez obtenidos los grupos e interpretado los resultados conviene, siempre que sea posible, proceder a la validación de los mismos con el fin de averiguar, por un lado, hasta qué punto los resultados obtenidos son extrapolables a la población de la que vienen los objetos seleccionados y, por el otro, por qué han aparecido dichos grupos. Esta validación se puede realizar de forma externa o interna.

5.1 Validez interna

Se puede establecer utilizando procedimientos de validación cruzada. Para ello se dividen los datos en dos grupos y se aplica el algoritmo de clasificación a cada grupo comparando los resultados obtenidos en cada grupo. Por ejemplo, si el método utilizado es el de las k-medias se asignaría cada objeto de uno de los grupos al cluster más cercano obtenido al clasificar los datos el otro grupo y se mediría el grado de acuerdo entre las clasificaciones obtenidas utilizando los dos métodos

5.2 Validez externa

Se puede realizar comparando los resultados obtenidos con un criterio externo (por ejemplo, clasificaciones obtenidas por evaluadores independientes o analizando en los grupos obtenidos, el comportamiento de variables no utilizadas en el proceso de clasificación) o realizando un Análisis Cluster con una muestra diferente de la realizada.

Ejempo 2

En los 3 gráficos siguientes se muestra la composición de cada grupo por religión mayoritaria, región económica y clima predominante. Se observa que la mayor parte de los países cristianos pertenecen al grupo 2 siendo esta diferencia más clara en los cristianos ortodoxos y protestantes. Por otro lado, los países musulmanes y los que practican otras religiones están en los grupos 1 y 4. Los países budistas se distribuyen equitativamente en los 3 grupos

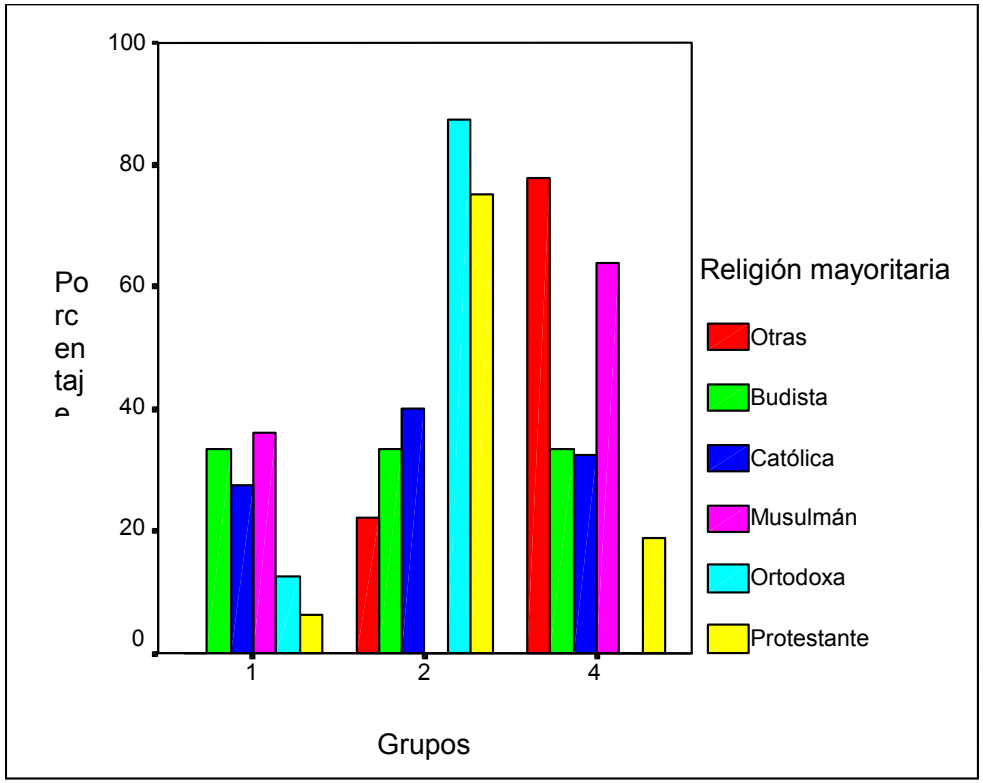


Gráfico 3: Composición de los grupos por religión

Por regiones económicas, los países del primer y segundo mundos (OCDE y Europa Oriental) pertenecen todos al segundo grupo, los países de América Latina y Oriente Medio tiende a estar en el grupo 1 mientras que todos los países africanos y la mayor parte de los países de Asia están incluidos en el grupo 4. Los grupos reflejan, por lo tanto, las diferencias existentes entre las diversas regiones económicas del mundo.

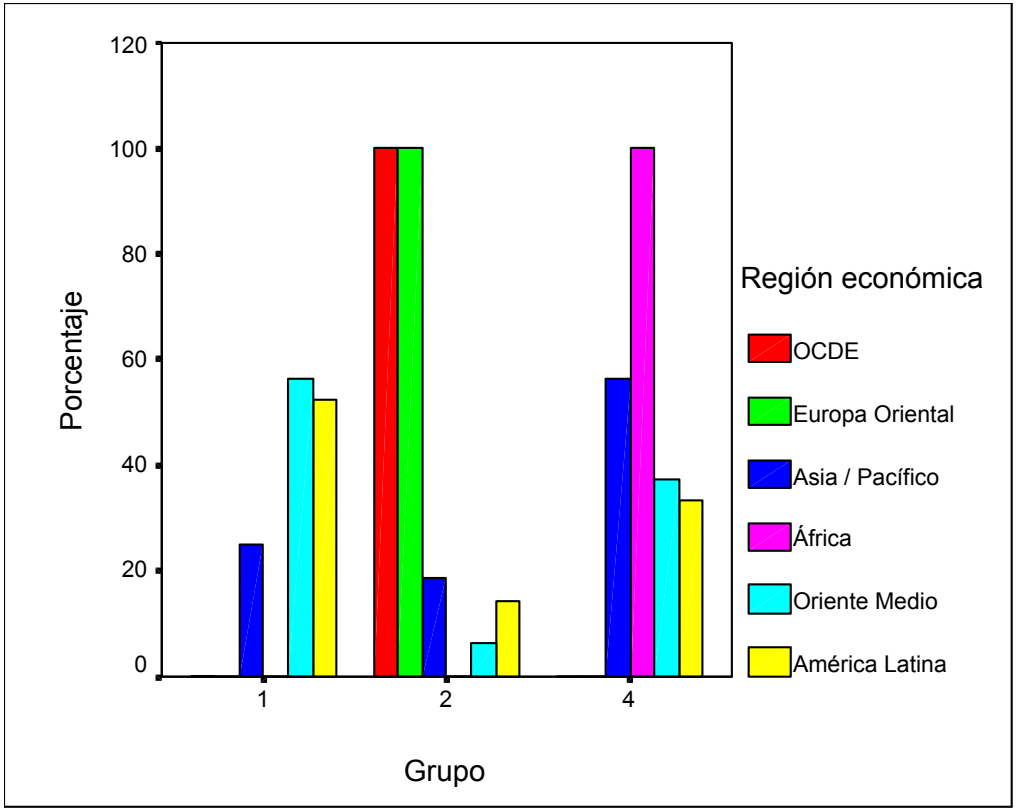


Gráfico 4: Composición de los grupos por región económica

El gráfico 5 pone de manifiesto la influencia del clima en la composición de los grupos. La mayor parte de los países con climas templados y frío pertenecen al grupo 2 mientras que los países con clima desértico, ecuatorial y tropical tienden a estar en el grupo 4 y los de clima árido en el grupo 1.

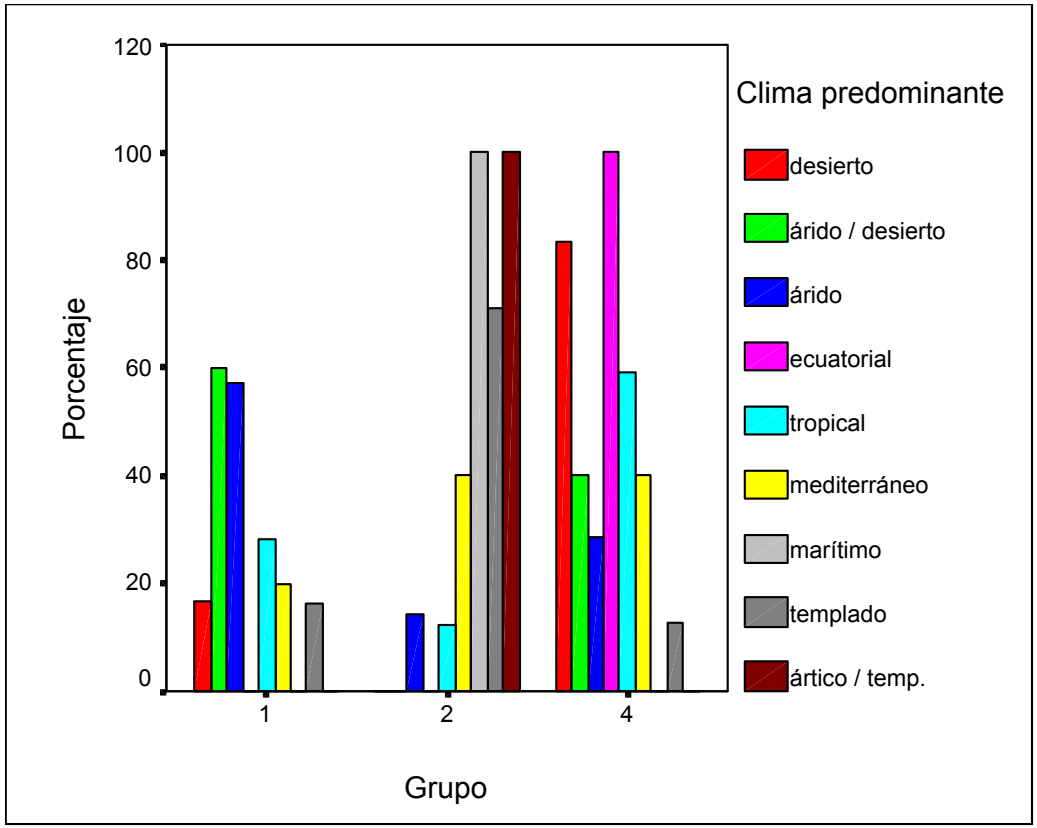


Gráfico 5: Composición de los grupos por clima predominante

Resumen

- El Análisis Cluster, también conocido como Análisis de Conglomerados, Taxonomía Numérica o Reconocimiento de Patrones, es una técnica estadística multivariada cuya finalidad es dividir un conjunto de objetos en grupos (*cluster* en inglés) de forma que los perfiles de los objetos en un mismo grupo sean muy similares entre sí (cohesión interna del grupo) y los de los objetos de clusters diferentes sean distintos (aislamiento externo del grupo).

  Para llevar a cabo un análisis de este tipo se deben los siguientes pasos:

  1. Plantear el problema a resolver por un Análisis Cluster
  2. Establecer medidas de semejanza y de distancia entre los objetos a clasificar en función del tipo de datos analizado
  3. Analizar algunos de los métodos de clasificación propuestos en la literatura haciendo especial énfasis en los métodos jerárquicos aglomerativos y en el algoritmo de las k-medias, y determinar el número de grupos.
  4. Interpretar los resultados obtenidos
  5. Analizar la validez de la clasificación obtenida

- Conviene hacer notar, finalmente, que es una técnica eminentemente exploratoria cuya finalidad es sugerir ideas al analista a la hora de elaborar hipótesis y modelos que expliquen el comportamiento de las variables analizadas identificando grupos homogéneos de objetos. Los resultados del análisis deberían tomarse como punto de partida en la elaboración de teorías que expliquen dicho comportamiento.

SÍNTESIS PRINCIPALES MÉTODOS EN SPSS:

Análisis de Conglomerados Jerárquico

Este procedimiento intenta identificar grupos relativamente homogéneos de casos (o de variables) basándose en las características seleccionadas, mediante un algoritmo que comienza con cada caso (o cada variable) en un conglomerado diferente y combina los conglomerados hasta que sólo queda uno. Las medidas de distancia o similaridad a utilizar se escogen de acuerdo al tipo de datos considerados en el análisis.

Los datos pueden ser cuantitativos, binarios o datos de recuento (frecuencias). El escalamiento de variables es un aspecto importante, ya que las diferencias en escalamiento pueden afectar las soluciones de conglomeración. Si las variable muestran grandes diferencias en el escalamiento (por ejemplo una se mide en pesos y la otra en años), puede ser útil considerar su estandarización.

Para aplicar este análisis, se supone que las medidas de distancia empleadas son adecuadas para los datos analizados (ver medidas de proximidad o distancia). Asimismo, se debe incluir todas las variables relevantes en el análisis. Si se omiten variables de interés la solución obtenida puede ser equívoca.

El análisis de conglomerados jerárquico es un método exploratorio, por lo tanto los resultados deben considerarse provisorios hasta que sean confirmados mediante otra muestra independiente.

Ejemplo. ¿Existen grupos identificables de programas televisivos que atraigan a audiencias similares dentro de cada grupo? Con el análisis de conglomerados jerárquico, podría agrupar los programas de TV (los casos) en grupos homogéneos basados en las características del espectador. Esto se puede utilizar para identificar segmentos de mercado. También puede agrupar ciudades (los casos) en grupos homogéneos, de manera que se puedan seleccionar ciudades comparables para probar diversas estrategias de marketing.

Análisis de conglomerados de K-medias

Este procedimiento intenta identificar grupos de casos relativamente homogéneos basándose en las características seleccionadas y utilizando un algoritmo que puede gestionar un gran número de casos. Sin embargo, el algoritmo requiere que el usuario especifique el número de conglomerados. Puede especificar los centros iniciales de los conglomerados si conoce de antemano dicha información. Aunque estos estadísticos son oportunistas (ya que el procedimiento trata de formar grupos que de hecho difieran), el tamaño relativo de los estadísticos F de los análisis de varianza, proporciona información acerca de la contribución de cada variable a la separación de los grupos.

Las variables a considerar en este análisis deben ser cuantitativas. Si los datos son binarios o de frecuencias, es preferible el análisis de conglomerados jerárquico.

La distancia utilizada en este análisis es la distancia euclidiana simple, de manera que el escalamiento de las variables es un aspecto importante, ya que las diferencias en escalamiento pueden afectar las soluciones de conglomeración. Si las variable muestran grandes diferencias en el escalamiento (por ejemplo una se mide en pesos y la otra en años). En tal caso, se debe considerar la estandarización de las variables.

El procedimiento supone que se ha seleccionado el número apropiado de conglomerados y que se ha incorporado a todas las variables relevantes.