Tema 6

Análisis cluster

El análisis cluster es un conjunto de técnicas multivariantes cuyo objetivo es agrupar objetos o individuos basándose en las características que estos poseen. El *Análisis Cluster* clasificará a los objetos, de tal forma que cada objeto será muy parecido a los que hay en su grupo. Los grupos resultantes deben mostrar mucha homogeneidad entre los elementos del grupo y un alto grado de heterogeneidad entre los diferentes grupos. A partir de ahora a cada uno de estos grupos los denominaremos cluster.

Este tipo de análisis es ampliamente utilizado en la Psicología, Biología, Sociología, Economía, Ingeniería y negocios y puede recibir diversos nombres según la ciencia en la que se aplique como es el de análisis Q, análisis de clasificación o taxonomía numérica.

El análisis cluster es una herramienta muy útil en diferentes situaciones, por ejemplo, un investigador que ha recogido datos en un cuestionario puede enfrentarse a un número elevado de observaciones que no tendrá sentido a menos que clasifique en grupos manejables. Por lo tanto el Análisis Cluster será una técnica de reducción de datos mediante la reducción de la población en subgrupos más manejables. El análisis cluster es también aplicable a, por ejemplo, la clasificación psicológica o rasgos personales o a la segmentación del mercado. Es un tipo de análisis muy sencillo y aplicable en muchas situaciones.

Ahora bien, el Análisis Cluster tiene algunos problemas, como son el no poder realizar inferencia estadística, utilizándose solamente como una técnica exploratoria. También es importante destacar que la solución no es única y se puede obtener distintas soluciones dependiendo del procedimiento seleccionado.

6.1. Introducción al Análisis Cluster.

Como ya hemos comentado, el análisis cluster se encargará de formar grupos de tal manera que estos sean los mas homogéneos internamente y lo mas heterogéneos entre ellos. El primer paso será la selección de la variables, y la detección de datos atípicos. Posteriormente, tendremos que establecer como se mide la relación entre individuos, para ello tendremos que utilizar una medida de la similitud o relación entre los individuos, de tal manera que uniremos

aquellos individuos que más se parezcan entre si, es decir, la similitud sea máxima. El tercer paso será la elección de una técnica apropiada y finalmente la validación de los resultados.

6.1.1. Elección de las variables

En primer lugar tendremos que estudiar el tipo de variables con la que trabajar. En principio solo nos serán útiles las variables de tipo cuantitativos (numéricas); si tenemos variables cualitativas, como puede ser el nivel de estudios, tendremos que recodificarlas en numéricas.

Otro aspecto muy importante es la elección de las variables a utilizar. Evidentemente, sobre cualquier individuo es posible encontrar un gran número de variables, pero esto no siempre es útil, ya que la inclusión de variables irrelevantes no puede ser contrastada por el análisis cluster y además aumenta la posibilidad de errores en la conclusión final. Por ello se deben de eliminar las variables irrelevantes en base al objetivo de la investigación.

También es interesante el tipificar las variables. Si las variables están medidas en diferentes unidades o escalas, la comparación entre unas variables u otras será difícil. Por ello se suelen tipificar los datos, de tal manera que obtengamos que todas las variables tengan media 0 y desviación típica 1 y además que no existan unidades entre los valores.

Análisis cluster por individuos o variables

Generalmente lo que se pretende agrupar son individuos, pero existe algunas circunstancias en las que es interesante agrupar variables para intentar buscar variables de comportamiento similar. Para ello, la metodología es la misma que para el análisis cluster por individuos y simplemente tendremos que transponer la matriz de datos y aplicar el método general.

6.1.2. Elección de una medida de asociación

Para poder unir los individuos en grupos, hemos de seleccionar una medida de similaridad entre individuos, de tal manera que esta nos marque la relación entre los individuos. Dentro de estas medidas podemos utilizar dos conceptos, el de distancia o el de similaridad.

- Cuando se elige una distancia como medida de asociación los grupos se formarán como aquellos más parecidos, es decir, la distancia sea la mínima, generalmente es usado en aquellos datos que son medibles.
- Cuando se elige una medida de similaridad los grupos se formarán maximizando la similaridad.

De tal manera que existirán muchos tipos diferentes de distancias y similaridades y dependiendo de cada circunstancia se elegirán una u otra. Es usado en variables no medibles.

6.1.3. Elección de la técnica cluster. Métodos jerárquicos y no jerárquicos

Existen dos grandes grupos de técnicas de análisis cluster, que son los métodos jerárquicos y no jerárquicos.

- *Métodos jerárquicos*: son aquellos que para formar un cluster nuevo une o separa alguno ya existente para dar origen a otros dos de forma que se maximice una similaridad o se minimice una distancia. Dentro de estos a su vez se clasifican en:

  1. Asociativos o aglomerativos: se parte de tantos grupos como individuos y se van agrupando hasta llegar a tener todos los individuos en un solo grupo.
  2. Disociativos: se parte de un solo grupo que contenga a todos los individuos y se va separando hasta llegar a formar grupos individuales.

- *Métodos no jerárquicos*: se clasifican los individuos en k grupos, estudiando todas las particiones de individuos en esos k grupos y eligiendo la mejor partición.

La principal ventaja de los métodos jerárquicos es que se puede representar el problema en forma de árbol o dendograma donde se observa muy bien la solución final.

A su vez dentro de cada uno de estos grupos existen muchos métodos entre los que existen diferencias dependiendo de la manera de medir las similitudes no entre individuos, sino entre los grupos, por ejemplo si dos individuos forman un primer grupo, ¿qué medida se toma como representativa de ese grupo, la media, el más cercano, el más lejano...? Dependiendo de esto surgirán muchas posibles técnicas concretas que estudiaremos en el apartado de aplicación. El problema está en que no existe una técnica fiable para determinar cual de estos métodos es mejor.

Otro problema será el número de conglomerados a decidir. Tampoco existe un procedimiento fiable para determinar el número de grupos. Existen algunas técnicas como estudiar las distancias a las que se van uniendo los grupos y parar cuando la distancia llegue a un valor determinado. Generalmente se estudia la solución y nos quedaremos con un número de cluster interesante a nuestro análisis.

6.2. Interpretación de los resultados

Una vez determinados los grupos, corresponderá al investigador de cada campo, psicológico, sociólogo, pedagogo..., investigar los grupos y el por que de su formación y sacar las conclusiones relevantes de este, así como las características en las que se diferencian cada conglomerado.

6.3. Aplicación mediante SPSS

Como aplicación vamos a realizar un Análisis Cluster al fichero de datos **cluster**. Este fichero contiene los datos de veinte variables (nombradas V2,V3,...,V21) de cohesión social de

10 países europeos. El objetivo es realizar un Análisis Cluster que agrupe los países europeos según su comportamiento en términos de su cohesión social.

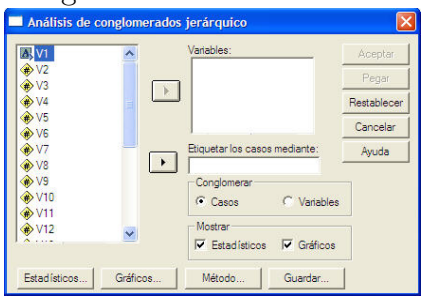
En primer lugar hemos de observar el fichero. Como en cualquier procedimiento estadístico, necesitaremos un análisis descriptivo previo de los datos para observar que no existan datos anómalos. Observamos que además de las 20 variables económicas, disponemos de una variable indicadora del país (V1) que no entrará en el análisis. Los datos están directamente sacados del EUROSTAT (Instituto de Estadística Europeo) y por lo tanto vamos a suponer que no existirán datos anómalos procedentes de la introducción o manipulación de los datos y que si existe algún dato atípico, este será una observación real. Si utilizamos el *menú Explorar* (ver sección 2.4.1) observaremos que en algunas variables existen datos extremos, por ejemplo, Alemania tiene valores extremos superiores para las variables V3 y V5, Italia para las variables V9 y V10 o la V20 tiene un dato extremo superior y otro inferior. Pero como hemos dicho, estos datos son reales y provienen de las distintas economías de cada país y es además según esos factores por los que querremos agrupar.

Una vez estudiados los posibles datos anómalos, no plantearíamos el Análisis Cluster, para ello seleccionaremos los menús:

Analizar
Clasificar
Conglomerados jerárquicos

Obteniendo el siguiente Cuadro:

Figura 6.1: Cuadro cluster



Las opciones disponibles son *Estadísticos, Gráficos, Método* y *Guardar* que pasaremos a describir a continuación.

6.3.1. Botón Método

Antes de realizar el Análisis Cluster, hay que decidir que método vamos a utilizar en el análisis. Las opciones posibles son:

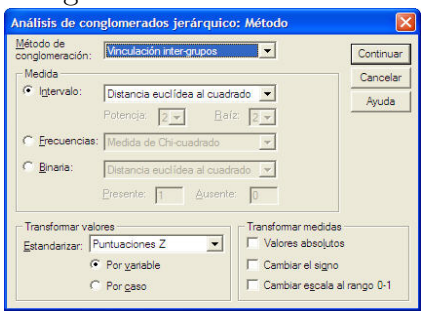
- *Método de conglomeración.* Será la metodología utilizada para calcular las distancias entre clusters. Como vimos, la distancia entre individuos estará basada en medidas matemáticas. El problema estará en como calcular las distancias entre un individuo

y un cluster o entre varios cluster. Dependiendo de como se realicen estas, se obtiene unas metodologías u otras. Las que proporciona SPSS son: vinculación inter-grupos, vinculación intra-grupos, vecino más próximo, vecino más lejano, agrupación de centroides, agrupación de medianas y método de Ward. Generalmente, las más utilizadas son la vinculación inter-grupos, el vecino más próximo y el método de Ward.

- *Medida*. Permite especificar la medida de distancia o similaridad que será empleada en la aglomeración. Dependiendo del tipo de dato que tengamos usaremos una u otra. Los tipos de medidas son:

  1. *Datos de intervalo:* distancia euclídea y distancia euclídea al cuadrado (las más comunes), coseno, correlación de Pearson, Chebychev, bloque, Minkowski y personalizada.
  2. *Datos de frecuencias:* medida de la chi-cuadrado y medida de la phi-cuadrado.
  3. *Datos binarios:* Distancia euclídea, distancia euclídea al cuadrado, diferencia de tamaño, diferencia de configuración, varianza, dispersión, forma, concordancia simple, correlación phi de 4 puntos, Lambda, D de Anderberg, Dice, Hamann, Jaccard, Kulczynski 1 y 2, Lance y Williams, Ochiai, Rogers y Tanimoto, Russel y Rao, Sokal y Sneath 1, 2,3, 4 y5, Y de Yule y Q de Yule.
     NOTA: En la ayuda del programa explica las características de cada medida.

Figura 6.2: Botón Método



- *Transformar valores*. Permite estandarizar los valores de los datos por casos o por variables, antes de calcular las proximidades. Los métodos disponibles de estandarización son: puntuaciones z, rango -1 a 1, rango 0 a 1, magnitud máxima de 1, media de 1 y desviación típica 1. En nuestros estudios utilizaremos la opción de *Puntuaciones Z*, siempre que los datos no tenga la misma escala o unidad.

- *Transformar medidas*. Permite transformar los valores generados por la medida de distancia. Se aplican después de calcular la medida de distancia. Las opciones disponibles son: valores absolutos, cambiar el signo y cambiar la escala al rango 0–1.

Aplicación

Antes de empezar el análisis cluster, es necesario decidir que el método que se va a utilizar, la distancia y si es necesario tipificar los datos.

Con respecto al método, al ser el análisis cluster un método claramente experimental, es conveniente utilizar varios métodos distintos y comparar los resultados finales, ya que no existe un procedimiento para decidir cuál de los métodos es mejor. Para esta cuestión estimaremos varios métodos y nos quedaremos con el más interesante.

Para determinar la distancia, en nuestro caso estamos con datos de *tipo intervalo* y usaremos *la distancia euclídea al cuadrado*, que es la más utilizada.

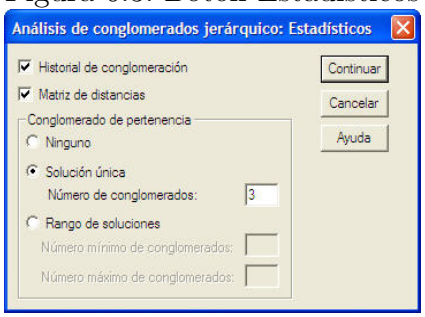
En este ejemplo, los datos están tomados en diferentes escalas de medida, por lo que decidimos transformar los valores en las *puntuaciones Z* para los casos. Con esta opción SPSS calcula la media y desviación de cada variable y tipifica todos los datos (les resta la media y los divide entre la desviación típica).

6.3.2. Botón Estadísticos

Con este botón podremos obtener:

- *Historial de conglomeración*. Muestra los conglomerados en cada etapa, las distancias entre los conglomerados que se combinan, así como el último nivel del proceso de aglomeración en el que cada caso se unió a su conglomerado correspondiente.
- *Matriz de distancias*. Proporciona las distancias o similaridades entre los elementos.
- *Conglomerado de pertenencia*. Muestra el conglomerado al cual se asigna cada caso en una o varias etapas de la combinación de los conglomerados. A su vez las opciones disponibles son Solución única y Rango de soluciones.

Figura 6.3: Botón Estadísticos



Aplicación

Vamos a seleccionar las tres opciones que nos permite SPSS. En la opción conglomerados de pertenencia elegimos solución única para tres conglomerados.

Para realizar el análisis, en la parte de *Variables*, hemos de indicar todas las variables según las cuales clasificaremos (V2-V21) y en *Etiquetar los casos* la variable con las etiquetas de los

países (V1). Evidentemente, como queremos agrupar los países, que son individuos o casos, mantenemos la opción de *Conglomerar por Casos*.

Los resultados que obtenemos son:

- *La matriz de distancias*. Esta matriz señala las distancias entre los individuos según la distancia euclídea al cuadrado. Esta matriz es simétrica y si la observamos veremos que el primer cluster estará formado por aquellos individuos más cercanos (con menor distancia entre ellos) que son Francia y Bélgica.

Figura 6.4: Matriz de distancias

Matriz de distancias

	distancia euclídea al cuadrado									
Caso	1:Belgium	2:Germany	3:Greece	4:Spain	5:France	6:Italy	7:Austria	8:Portugal	9:Finland	10:Bulgaria
1:Belgium	,000	16,609	40,805	40,815	6,281	34,073	28,843	79,060	17,969	28,409
2:Germany	16,609	,000	47,159	46,115	18,265	51,420	40,338	71,150	26,972	55,091
3:Greece	40,805	47,159	,000	20,355	27,219	33,839	38,956	34,497	51,008	46,694
4:Spain	40,815	46,115	20,355	,000	26,404	14,247	40,852	13,627	49,379	54,720
5:France	6,281	18,265	27,219	26,404	,000	28,499	12,400	54,950	8,651	28,259
6:Italy	34,073	51,420	33,839	14,247	28,499	,000	54,214	48,940	54,436	50,584
7:Austria	28,843	40,338	38,956	40,852	12,400	54,214	,000	59,689	6,973	63,729
8:Portugal	79,060	71,150	34,497	13,627	54,950	48,940	59,689	,000	76,584	92,501
9:Finland	17,969	26,972	51,008	49,379	8,651	54,436	6,973	76,584	,000	58,419
10:Bulgaria	28,409	55,091	46,694	54,720	28,259	50,584	63,729	92,501	58,419	,000

- *Historial de la conglomeración*. Nos va indicando el orden de las uniones y la distancia a la que lo hacen. Por ejemplo, los Francia y Bélgica son los primeros que se unen a una distancia de 6.281 formando el cluster 1, luego lo hacen Austria y Finlandia a 6.93 (cluster 2). Posteriormente España y Portugal (cluster 3), luego el cluster 1 ya formado con el cluster 3 (forma el cluster 4). Posteriormente este cluster 4 se le une Alemania. Después al grupo formado por España y Portugal se les une Grecia y luego Italia y así hasta el final.

Figura 6.5: Historial

Historial de conglomeración

	Conglomerado que se combina			Etapa en la que el conglomerado aparece por primera vez		
Etapa	Conglomerado 1	Conglomerado 2	Coeficientes	Conglomerado 1	Conglomerado 2	Próxima etapa
1	1	5	6,281	0	0	4
2	7	9	6,973	0	0	4
3	4	8	13,627	0	0	6
4	1	7	16,966	1	2	5
5	1	2	25,546	4	0	8
6	3	4	27,426	0	3	7
7	3	6	32,342	6	0	9
8	1	10	46,781	5	0	9
9	1	3	50,720	8	7	0

- *Conglomerado de pertenencia*. Si indicamos el número de cluster final que queremos obtener, por ejemplo 3 en este caso, nos indica los cluster finales. Por ejemplo, existe un grupo formado por Bélgica, Alemania, Francia, Austria y Finlandia, otro grupo con Grecia, España, Italia y Portugal y un tercer grupo con Bulgaria.

Figura 6.6: Conglomerado de pertenencia

Conglomerado pertenencia

Caso	3 conglomerados
1:Belgium	1
2:Germany	1
3:Greece	2
4:Spain	2
5:France	1
6:Italy	2
7:Austria	1
8:Portugal	2
9:Finland	1
10:Bulgaria	3

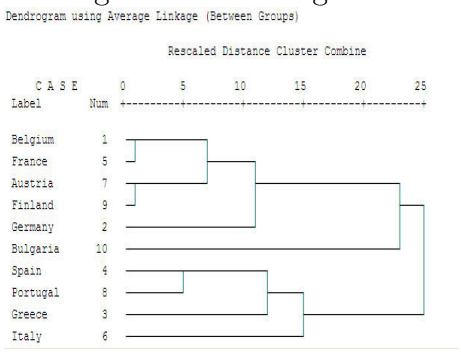
6.3.3. Botón Gráficos

Dentro de la *opción Gráficos* aparece la más interesante del análisis cluster, que es el *dendograma*. El dendograma es la representación gráfica de la formación de los cluster. Es una representación muy intuitiva y que resumen toda la información del análisis que son los cluster que se forman y la distancia a la que lo hacen.

Aplicación

Si seleccionamos la opción Dendograma y lo analizamos:

Figura 6.7: Dendograma



podemos observar que los grupos que se forma son los siguientes:

cluster1	cluster2	cluster3
Belgica	Bulgaria	España
Francia		Portugal
Austria		Grecia
Finlandia		Italia
Alemania		

Por lo que tendríamos los países europeos clasificados en tres grupos según su comportamiento en términos de cohesión social. El objetivos a partir de aquí será utilizar otros métodos de conglomerado (no de distancia) y verificar si se forman los mismos grupos o al menos similares.

6.3.4. Conclusiones

Antes de determinar los subgrupos finales vamos a utilizar otros métodos de conglomerados para ver las diferentes clasificaciones.

1. Método de vinculación intra-grupos:

cluster1	cluster2	cluster3
Belgica	Bulgaria	España
Francia		Portugal
Austria		Grecia
Finlandia		
Alemania		
Italia		

2. Método del vecino más próximo, método del vecino más lejano, método de agrupación de centroides, agrupación de medianas y método de Ward:

cluster1	cluster2	cluster3
Belgica	España	Bulgaria
Francia	Portugal	
Austria	Grecia	
Finlandia	Italia	
Alemania		

Por lo tanto, por casi todos los métodos obtenemos los mismos resultados, por lo que asumiremos que estos países según los datos referentes a la cohesión social un grupo lo formarían Bélgica, Francia, Austria, Alemania y Finlandia; otro grupo con España, Portugal, Grecia e Italia y Bulgaria que no se agruparía con ningún otro país.

6.4. Ejercicios obligatorios

6.4.1. Ejercicio primero

El fichero **Cluster2** contiene los datos de 11 variables de tipo macroeconómico de 24 países europeos. Se quiere encontrar grupos de comportamiento similares según un análisis cluster, utilizando la distancia euclídea al cuadrado con 5 subgrupos y usando las viculaciones intra-grupos e inter-grupos.

6.4.2. Ejercicio segundo

Se desea realizar un estudio de mercado, para ello se selecciona una muestra de 50 individuos a los que se les estudian 7 variables sobre un determinado producto. El objetivo es encontrar patrones de comportamiento en estos individuos para segmentar la población. Para ello se decide realizar un análisis cluster sobre el fichero **Cluster3**, de tal manera que clasifiquemos

a los individuos en grupos según su comportamiento; realizar este análisis cluster mediante los métodos del vecino más cercano y más lejano y con la distancia euclídea al cuadrado. Indicar en cuantos grupos puede ser segmentado la población.