# Capítulo 9

*Análisis de Cluster*

*o Análisis de Conglomerados*

# Capítulo 9

## *Análisis de Cluster o Análisis de Conglomerados*

### *1. Introducción*

El **análisis de *cluster***, o de **conglomerados**, es una técnica de análisis que se centra, más que en las variables, en las unidades de análisis. Su finalidad fundamental es encontrar la agrupación implícita que subyace en las unidades de análisis en relación con un determinado conjunto de variables. El análisis de cluster nos permite **clasificar** las unidades de análisis en grupos homogéneos de tal manera que las unidades pertenecientes a uno de los grupos o conglomerados serán lo más parecidas entre sí aunque muy diferentes respecto a los otros grupos.

Tal y como ya sucediera con el análisis factorial, esta técnica no es unitaria, pudiendo identificar una gran variedad de métodos de análisis de cluster. De las dos grandes categorías en las que se suelen estructurar estos métodos (jerárquicos y no jerárquicos), el procedimiento que se va a seguir en esta exposición es el **taxonómico aglomerativo jerárquico**:

- **Taxonómico** pues la finalidad perseguida es clasificar a las unidades estadísticas en grupos lo más homogéneos posibles;

- Esta taxonomía es **aglomerativa** porque a partir de las observaciones éstas se van agrupando de forma progresiva en grupos o *clusters* cada vez mayores. El análisis empieza con tantos conglomerados como casos y finaliza con la agrupación en un solo conglomerado de todas las unidades de análisis;

- Por último, los grupos que se obtienen son **jerárquicos** porque cada nueva fusión se va ampliando conforme decrece la similitud entre los mismos. Esto es, a partir de la matriz de datos originales se van formando nuevos conglomerados de forma ascendente agrupando, en cada etapa, a las unidades de los conglomerados más próximos. Las técnicas jerárquicas se aplican fundamen-

talmente cuando el número de las unidades de análisis es menor a 120.

Antes de empezar el desarrollo del análisis de *cluster* debemos tener en cuenta las siguientes consideraciones relativas a la selección de las variables.

- Esta técnica de análisis se puede aplicar sobre variables con cualquier nivel de medida aunque, es conveniente, que éste sea el mismo para todas ellas.

- Del mismo modo, las variables seleccionadas deben tener la misma escala pues, de lo contrario, los resultados pueden verse afectados. Si las variables no tienen la misma escala aquella que tenga mayor rango tendrá mayor peso en la medida de semejanza y/o distancia que seleccionemos para establecer los grupos. En estos casos procederíamos a la estandarización previa de las variables.

- Los resultados pueden verse alterados cuando el número de variables con las que trabajamos es muy elevado. Cuantas más variables seleccionadas, no solo más complicado es el proceso sino que también más posibilidades de que hayan variables que estén relacionadas (que midan la misma característica o una muy semejante). En estos casos conviene aplicar, como paso previo, una técnica reduccionista (tal es el caso de un análisis factorial si trabajamos con variables cuantitativas), simplificando el número de variables a sólo aquellas que estén incorrelacionadas. La clasificación definitiva se efectuará en base a los factores arrojados.

- Por último, no debemos olvidar que los grupos que obtengamos, y la clasificación asociada a los mismos, se realiza a partir de aquellas variables que hemos considerado como identificativas de los grupos y que, en consecuencia, las áreas sociales homogéneas que obtengamos serán significativas desde un punto de vista sociológico siempre y cuando los indicadores seleccionados sean los más relevantes.

En este capítulo exponemos las dos decisiones sobre las que se apoya esta técnica de análisis, a saber:

1. Elección de una medida de proximidad entre los individuos.
2. Elección de un criterio a partir del cual agrupar a los individuos o unidades de análisis (secciones censales, países, ciudades,...) en los conglomerados.

## *2. Elección de una medida de proximidad*

Una vez determinada la ***matriz inicial de datos*** (constituida por un conjunto mínimo de variables significativas, no correlacionadas y medidas en la misma unidad y escala) podemos cuantificar la *proximidad* o *similitud* que presentan las unidades de análisis respecto a los valores que en cada una de ellas toman dichas variables (o factores).

En esta fase del análisis deberemos elegir, de entre todos los posibles, un **criterio de similitud**. Dado el número elevado de casos y con la finalidad de agilizar los cálculos se ha aplicado el **cuadrado de la distancia euclídea** ( ésta es la opción que aparece por defecto en el SPSS cuando las variables tienen un nivel de medida interval).

A partir de la ***matriz de similaridad*** o ***distancias*** que se genera deberemos elegir, de entre todos los posibles, un algoritmo con el que poder formar los grupos entre las secciones censales. Es más, la diferencia entre los distintos métodos jerárquicos aglomerativos reside en el tipo de distancia elegida para medir la proximidad entre grupos. Dentro de las técnicas jerárquicas el **algoritmo de clasificación** que hemos aplicado ha sido el **método del promedio entre grupos** (este método calcula la proximidad entre dos grupos como el promedio de las distancias entre todos los pares de casos de tal manera que cada componente del par pertenece a un conglomerado distinto).

Una vez concluido el proceso de formación de conglomerados o grupos este método ofrece la posibilidad de seguir las distintas etapas de formación. La información detallada de lo que ha sucedido en cada una de ellas queda recogida en el **historial de conglomeración**. A partir de la solución obtenida en cada una de

las etapas se decidirá el número de conglomerados que se forman.

No obstante, para facilitar esta decisión contamos con la posibilidad de ilustrar de forma gráfica la información contenida en el historial de conglomeración. Los dos gráficos que se identifican con esta técnica son:

- El **dendograma**, gráfico más representativo de este tipo de análisis, asume la forma de un árbol de clasificación en el que es posible observar con toda claridad la forma y el número de los grupos que se van formando. En este gráfico es el eje de ordenadas el que adquiere verdadero protagonismo pues representa los distintos niveles de similidad en torno al cual se han ido agrupando las unidades de análisis en función de la medida elegida. Por su parte, en el eje de abcisas únicamente se identifican los casos u observaciones. El problema de esta representación gráfica, es que sólo se puede emplear cuando el número de casos es reducido ($n < 200$). Constituye un resumen de la información original presente en la matriz de distancias o similaridades y la información que presenta será más útil cuanto más agrupado sean los datos que represente.

- Otro gráfico muy representativo e identificativo de estos análisis es el **témpanos** o gráfico de **carámbanos**. En este gráfico aparece estructurado en columnas y filas en las que se sitúan las unidades de análisis o casos y el número de clusters correspondiente para cada paso del análisis respectivamente. La lectura del gráfico debe efectuarse de abajo a arriba. Este tipo de gráfico presenta la ventaja de que en cada paso podemos ver fácilmente el número de clusters correspondiente, (margen izquierdo del gráfico), de este modo si se desea una solución con 6 clusters basta con observar en la columna izquierda del gráfico el número 6 y a partir de ello ver como se aglutinan los diferentes casos en los cluster que hasta el paso correspondiente se han formado.

## *3. Elección de una criterio para decidir el número de conglomerados óptimo*

El principal inconveniente que podemos identificar en los métodos jerárquicos es que no hay criterio único que nos permita determinar cuántos grupos quedan una vez obtenidos los resultados. Nuevamente la decisión del número de conglomerados, grupos o áreas homogéneas, queda en manos del investigador. Éste, y considerando el objeto de estudio y el marco teórico desde el que ha sido formulada la investigación, decidirá con cuántos de los grupos se queda.

Para tal fin, Martínez enumera dos criterios en los que nos podemos apoyar. Ambos parten de la observación del dendograma.

- Determinar un punto en la escala de distancias a partir del cual hemos decidido que se consigue el equilibrio entre el número de subáreas y el grado de homogeneidad interna requerido.

- O bien, escoger los grupos que mantienen fuerte homogeneidad interna y cuya unión con otros grupos supone un gran salto en la escala de distancias y con ello una pérdida sensible de información específica del grupo.

Una vez expuestos los aspectos metodológicos básicos a considerar cuando es la técnica de *cluster* la aplicada en nuestros análisis, pasamos a relacionar la secuencia y pasos a seguir cuando el paquete estadístico SPSS es el que disponemos. El objetivo perseguido con el ejemplo, para el que recogemos los resultados en la última sección del capítulo, es el de clasificar un total de 11 países africanos atendiendo a su situación sanitaria. Los indicadores sanitarios son considerados indicadores indirectos de desarrollo, progreso y bienestar.

## *4. Cuadro de Diálogo del Análisis de Cluster*

**1er paso**: El análisis empieza seleccionado el el menú analizar la técnica en cuestión. La secuencia es la que sigue: **Analizar: Clasificar: Conglomerados Jerárquico** *(figura 1)*. En nuestro ejemplo, partimos de 11 unidades de análisis, por lo que inicialmente realizaremos este tipo de análisis de *cluster*.

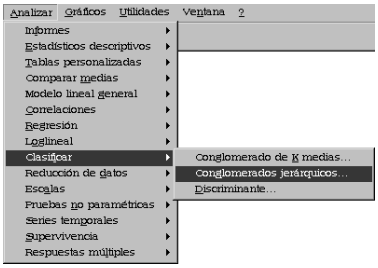

Figura 1

Figura 2

Figura 3

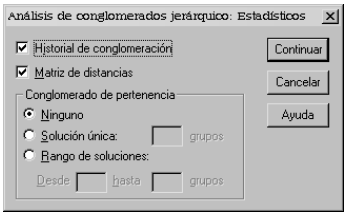

Figura 4

**2º paso**: Una vez dentro del cuadro de diálogo del análisis de *cluster*, seleccionamos las variable y las unidades de análisis *(figura 2)*. Donde pone **Variables**, introducimos las variables que nos van a permitir la ordenación y clasificación de los países seleccionados. Donde pone **Etiquetar los casos mediante**, introducimos la variable a clasificar que en este caso son las unidades de análisis, osea, los países.

**3er paso**: Una vez seleccionadas las variables y las unidades solicitamos que el proceso de **Conglomerar** se realicen por **Casos**.

Una vez señalados los requisitos básicos deberemos ir introduciendo las restricciones y peticiones que nos parezcan las más indicadas para el tipo de análisis que estamos desarrollando. Para ello deberemos acceder a los subcuadros de diálogos correspondientes.

**4º paso**: Lo primero que deberemos concretar es la medida de distancia que vamos a aplicar y el método de conglomeración. Cliqueando sobre el botón de comando **Método**, situado en la parte inferior del cuadro de diálogo principal, accedemos al subcuadro de diálogo Método (*figura 3*). Allí:

- Y dado que las variables seleccionadas son de **Intervalo** seleccionaremos de la ventana desplegable la opción que viene por defecto, esto es, **Distancia euclídea al cuadrado**.

- En el mismo subcuadro seleccionamos el **Método** o algoritmo de clasificación y que en nuestro caso es el de **Vinculación promedio inter-grupos** (estas dos últimas especificaciones son las que aparecen por defecto).

- Por último, como las variables seleccionadas no tienen la misma escala debemos **Transformar los valores** en **Puntuaciones Z**.

**5ºpaso**: Cliqueando en el botón de comando **Estadísticos**, situado en la parte inferior del cuadro de diálogo principal, accedemos al subcuadro que nos permite seleccionar *(figura 4)*:

- La **Matriz de distancias**. Esta matriz define la distancia entre dos conglomerados como el promedio de las dis-

tancias entre todos los pares de individuos en los cuáles un miembro del par pertenece a cada uno de los clusters formados. Este método, a diferencia de otros, utiliza información de todas las distancias entre pares de individuos y no solo de los más alejados o de los más próximos. Por esta razón el método de promedio entre grupos, el aplicado en el ejemplo que recogemos, es uno de los más utilizados.

- El **Historial de conglomeración**. El historial de conglomeración recoge las etapas seguidas en la construcción de los conglomerados. Por ejemplo, en la etapa número uno, los países Iraq (4) y Jordania (5) se unen para formar el primer conglomerado.

Por último, nos quedaría representar de una forma más visual los resultados obtenidos. Las técnicas jerárquicas, tanto en su versión aglomerativa como divisiva, permiten diferentes formas de representación gráfica que son de gran utilidad para la comprensión e interpretación del proceso de obtención de los distintos *clusters* que se han ido creando en relación con las diferentes medidas de similaridad y algoritmos de clasificación elegidos. A este respecto, entre las representaciones gráficas más comunes destacamos el gráfico de carámbanos o icicle plot y el dendograma.

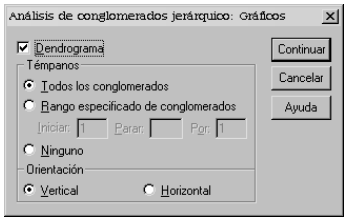

Figura 5

**6°paso**: Dentro del subcuadro de diálogo de **Gráficos** *(figura 5),* seleccionamos la opción de **dendograma** con una **Orientación** en **Vertical** y de esa forma obtendremos los gráficos representativos del análisis de *cluster*. El gráfico de **Témpanos** aparece seleccionado por defecto.

La lectura del gráfico de **Témpanos** debe efectuarse de abajo a arriba. En nuestro ejemplo, en el primer paso tenemos 10 *clusters*, ya que aunque partimos de 11 casos en este primer momento se unen el 5 y el 4; en el siguiente paso se forma otro grupo que aglutina los casos (5, 4) y 3, así sucesivamente vamos subiendo hasta establecer todos los *clusters*.

## *5. Bibliografía Comentada*

• Ayuntamiento de Alcobendas (1992): *Vivir en Alcobendas. Estructura social y conflicto*. Madrid, Ayuntamiento de Madrid, 165 págs.

> *En esta publicación, y con la finalidad de establecer las claves del cambio social por las que pasa el municipio, se recoge el análisis de su estructura social en su doble vertiente; esto es, en su forma tanto dinámica como estática. En el capítulo dedicado al estudio de su estructura social al análisis de la desigualdad (se aplica la técnica factorial de componentes principales para determinar los factores o dimensiones de diferenciación social) y diferenciación espacial (las puntuaciones factoriales han sido aplicadas en un análisis de cluster obteniendo las áreas de segregación social homogéneas del ámbito de estudio).*

• Basulto, Jesús y Arias, Carlos (1989): "Un estudio sobre la diferenciación residencial en el espacio urbano de Sevilla", *Ciudad y Territorio*, nº 79-1, pp. 85-92.

> *Esta investigación recoge un doble objetivo: en primer lugar se pretende identificar los grupos de población residente homogéneos (se aplica el análisis de cluster); y en segundo lugar, replicar aplicando el análisis factorial de componentes principales el modelo de Burguess. Se parte de las 27 áreas morfológicas establecidas por la Gerencia de Urbanismo y de 11 indicadores sociodemográficos.*

## *6. Resultados*

Considerando las especificaciones relacionadas en la exposición del cuadro de diálogo del análisis de *clusters* los resultados que se obtienen son:

- En primer lugar, la **matriz de distancias**. Ésta se obtiene aplicando a la matriz da datos originales la medida de distancia seleccionada que en nuestro caso corresponde a la vinculación promedio inter-grupos. Es a partir de esta

matriz desde la que se desencadenan el resto de tablas y de información.

- Una vez cuantificada la proximidad o similitud que presentan las unidades de análisis respecto a los valores que en cada una de ellas toman las variables y aplicando un algoritmo de clasificación, en nuestro caso el de promedio entre grupos, obtenemos el **historial de conglomeración**, tabla a partir de la cual podemos seguir las etapas de formación de conglomerados.

- Por último, aparecen los **gráficos de carámbanos** o **diagrama de témpanos** y el **dendograma**. Estas representaciones, junto con el historial de conglomeración, nos ayudan a decidir el número de grupos definitivos.

### *6.1. Datos para el Análisis de Cluster. Situación Sanitaria de algunos paises de África*

| Pais | Medent | Farmac | Enferm | Camas | Grasas | Feculas | Espvida |
|---|---|---|---|---|---|---|---|
| ***Argelia*** | 129 | 23 | 350 | 3392 | 21 | 57 | 35 |
| ***Egipto*** | 483 | 131 | 454 | 2225 | 15 | 73 | 54 |
| ***Iran*** | 329 | 107 | 290 | 1113 | 24 | 60 | 51 |
| ***Iraq*** | 241 | 81 | 235 | 1898 | 28 | 57 | 54 |
| ***Jordania*** | 284 | 96 | 241 | 1712 | 25 | 49 | 52 |
| ***Libano*** | 933 | 192 | 564 | 4071 | 35 | 50 | 60 |
| ***Libia*** | 338 | 41 | 612 | 3215 | 24 | 55 | 57 |
| ***Marruecos*** | 94 | 26 | 233 | 1516 | 21 | 57 | 53 |
| ***Siria*** | 254 | 70 | 140 | 1163 | 13 | 69 | 52 |
| ***Tunez*** | 114 | 39 | 248 | 2967 | 21 | 57 | 53 |
| ***Turquia*** | 412 | 57 | 306 | 1738 | 16 | 71 | 55 |

*Medent: Médicos y Dentistas, Farmac: Farmaceuticos, Enferm: Enfermeras, Camas: Camas en Hospitales, Grasas: Grasas en la dieta, Feculas: Feculas en la dieta, Aspvida: Esperanza de Vida.*

### *6.2. Resumen de los Datos*

| Casos | | | | | |
|---|---|---|---|---|---|
| Valid | | Perdidos | | Total | |
| N | Porcentaje | N | Porcentaje | N | Porcentaje |
| *11* | *100,0%* | *0* | *,0%* | *11* | *100,0%* |

*a. Distancia euclídea al cuadrado usada*

### *6.3. Matriz de Distancias*

| Caso | distancia euclídea al cuadrado | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1:Argelia | 2:Egipto | 3:Iran | 4:Iraq | 5:Jordania | 6:Libano | 7:Libia | 8:Marrueco | 9:Siria | 10:Tunez | 11:Turquia |
| **1:Argelia** | | 22,582 | 15,742 | 14,760 | 14,584 | 46,715 | 16,547 | 12,425 | 19,357 | 8,928 | 18,525 |
| **2:Egipto** | 22,582 | | 8,030 | 12,553 | 15,025 | 28,404 | 12,896 | 14,567 | 8,437 | 13,071 | 3,535 |
| **3:Iran** | 15,742 | 8,030 | | 1,943 | 2,471 | 28,415 | 12,181 | 4,296 | 6,011 | 6,683 | 5,395 |
| **4:Iraq** | 14,760 | 12,553 | 1,943 | | 1,470 | 26,039 | 9,684 | 2,979 | 9,081 | 3,426 | 7,725 |
| **5:Jordania** | 14,584 | 15,025 | 2,471 | 1,470 | | 25,790 | 10,962 | 3,984 | 10,875 | 4,798 | 10,813 |
| **6:Libano** | 46,715 | 28,404 | 28,415 | 26,039 | 25,790 | | 19,707 | 41,951 | 50,391 | 33,877 | 37,130 |
| **7:Libia** | 16,547 | 12,896 | 12,181 | 9,684 | 10,962 | 19,707 | | 11,313 | 21,577 | 7,657 | 12,339 |
| **8:Marrueco** | 12,425 | 14,567 | 4,296 | 2,979 | 3,984 | 41,951 | 11,313 | | 5,599 | 2,242 | 6,243 |
| **9:Siria** | 19,357 | 8,437 | 6,011 | 9,081 | 10,875 | 50,391 | 21,577 | 5,599 | | 8,464 | 2,620 |
| **10:Tunez** | 8,928 | 13,071 | 6,683 | 3,426 | 4,798 | 33,877 | 7,657 | 2,242 | 8,464 | | 7,186 |
| **11:Turquia** | 18,525 | 3,535 | 5,395 | 7,725 | 10,813 | 37,130 | 12,339 | 6,243 | 2,620 | 7,186 | |

*Esta es una matriz de disimilaridades*

### *6.4. 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** | 4 | 5 | 1,470 | 0 | 0 | 2 |
| **2** | 3 | 4 | 2,207 | 0 | 1 | 5 |
| **3** | 8 | 10 | 2,242 | 0 | 0 | 5 |
| **4** | 9 | 11 | 2,620 | 0 | 0 | 6 |
| **5** | 3 | 8 | 4,361 | 2 | 3 | 7 |
| **6** | 2 | 9 | 5,986 | 0 | 4 | 7 |
| **7** | 2 | 3 | 9,376 | 6 | 5 | 8 |
| **8** | 2 | 7 | 12,326 | 7 | 0 | 9 |
| **9** | 1 | 2 | 15,939 | 0 | 8 | 10 |
| **10** | 1 | 6 | 33,842 | 9 | 0 | 0 |

### *6.5. Gráfico de Carambanos (Icicle plot)*

**Diagrama de témpanos vertical**

| Número de conglomerados | Caso | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 6:Libano | | 7:Libia | | 10:Tunez | | 8:Marrueco | | 5:Jordania | | 4:Iraq | | 3:Iran | | 11:Turquia | | 9:Siria | | 2:Egipto | | 1:Argelia |
| **1** | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X |
| **2** | X | | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X |
| **3** | X | | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | | X |
| **4** | X | | X | | X | X | X | X | X | X | X | X | X | X | X | X | X | X | | X |
| **5** | X | | X | | X | X | X | X | X | X | X | X | X | | X | X | X | X | | X |
| **6** | X | | X | | X | X | X | X | X | X | X | X | X | | X | X | X | | | X |
| **7** | X | | X | | X | X | X | | X | X | X | X | X | | X | X | X | | | X |
| **8** | X | | X | | X | X | X | | X | X | X | X | X | | X | | X | | | X |
| **9** | X | | X | | X | | X | | X | X | X | X | X | | X | | X | | | X |
| **10** | X | | X | | X | | X | | X | X | X | | X | | X | | X | | | X |

### *6.6. Gráfico Dendograma*

```
* * * H I E R A R C H I C A L  C L U S T E R  A N A L Y S I S * * *

 Dendrogram using Single Linkage

                          Rescaled Distance Cluster Combine

    C A S E          0         5        10        15        20        25
 Label      Num  +---------+---------+---------+---------+---------+
 Iraq         4   -+
 Jordania     5   -+---+
 Iran         3   -+   +-----+
 Marruecos    8   ---+-+     I
 Tunez       10   ---+       +-----+
 Siria        9   ---+-+     I     I
 Turquia     11   ---+ +-----+     +---+
 Egipto       2   -----+           I   +---------------------------+
 Libia        7   -----------------+   I                           I
 Argelia      1   ---------------------+                           I
 Libano       6   -----------------------------------------------+
```