Capítulo 22

Análisis de conglomerados (II): El procedimiento *Conglomerados jerárquicos*

Análisis de conglomerados jerárquico

A diferencia de lo que ocurre con el procedimiento *Análisis de conglomerados de K medias*, el procedimiento *Análisis de conglomerados jerárquico* permite aglomerar tanto *casos* como *variables* y elegir entre una gran variedad de métodos de aglomeración y medidas de distancia. Pero la diferencia fundamental entre ambos procedimientos está en que en el segundo de ellos se procede de forma *jerárquica*.

El análisis de conglomerados *jerárquico* comienza con el cálculo de la *matriz de distancias* entre los elementos de la muestra (casos o variables). Esa matriz contiene las distancias existentes entre cada elemento y todos los restantes de la muestra. A continuación se buscan los dos elementos más próximos (es decir, los dos más similares en términos de *distancia*) y se agrupan en un conglomerado. El conglomerado resultante es indivisible a partir de ese momento: de ahí el nombre de *jerárquico* asignado al procedimiento. De esta manera, se van agrupando los elementos en conglomerados cada vez más grandes y más heterogéneos hasta llegar al último paso, en el que todos los elementos muestrales quedan agrupados en un único conglomerado global. En cada paso del proceso pueden agruparse casos individuales, conglomerados previamente formados o un caso individual con un conglomerado previamente formado. El análisis de conglomerados *jerárquico* es, por tanto, una técnica *aglomerativa*: partiendo de los elementos muestrales individualmente considerados, va creando grupos hasta llegar a la formación de un único grupo o conglomerado constituido por todos los elementos de la muestra.

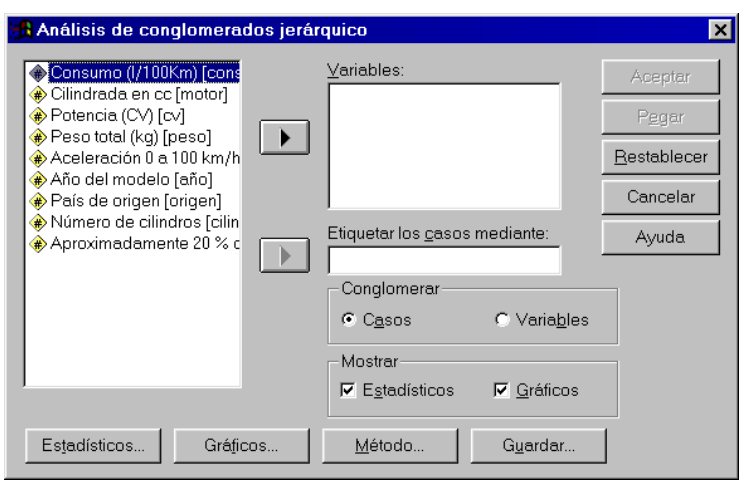
El procedimiento *Conglomerados jerárquicos* del SPSS informa de todos los pasos realizados en el análisis, por lo que resulta fácil apreciar qué elementos o conglomerados se han fundido en cada paso y a qué distancia se encontraban cuando se han fundido. Esto permite valorar la heterogeneidad de los conglomerados que se van fundiendo en cada etapa del análisis y decidir en cuál de ellas la fusión de elementos incrementa excesivamente la heterogeneidad de los conglomerados. Aunque el análisis termina cuando se ha conseguido agrupar a todos los casos en un único conglomerado, el objetivo del analista será el de descubrir la existencia de grupos homogéneos "naturales" que puedan existir en el archivo de datos.

La versatilidad del análisis de conglomerados jerárquico radica en la posibilidad de utilizar distintos tipos de medidas para estimar la distancia existente entre los casos o las variables, la posibilidad de transformar la métrica original de las variables y la posibilidad de seleccionar de entre una gran variedad de métodos de aglomeración. Pero no existe ninguna combinación de estas posibilidades que optimice la solución obtenida. En general, será conveniente valorar distintas soluciones para elegir la más consistente.

Para realizar un análisis de conglomerados jerárquico:

- Seleccionar la opción **Clasificar** > **Conglomerados jerárquicos** del menú **Analizar** para acceder al cuadro *Análisis de conglomerados jerárquico* que muestra la figura 22.1.

Figura 22.1. Cuadro de diálogo *Análisis de conglomerados jerárquico*.



La lista de variables de del archivo de datos contiene todas las variables del archivo, incluidas las variables de cadena (si bien estas últimas sólo pueden utilizarse para etiquetar los casos).

Para obtener un análisis de conglomerados jerárquico:

- ▶ Seleccionar las variables numéricas que se desea utilizar para diferenciar los casos y formar los conglomerados, y trasladarlas a la lista **Variables**.
- ▶ Opcionalmente, seleccionar una variable para identificar los casos en las tablas de resultados y en los gráficos y trasladarla al cuadro **Etiquetar los casos mediante**.

Conglomerar. Las opciones de este apartado permiten decidir qué elementos del archivo de datos se desea agrupar:

- ○ **Casos**. Se agrupan los casos a partir de sus puntuaciones en las variables seleccionadas. Es la opción por defecto.
- ○ **Variables**. Se agrupan las variables seleccionadas en la lista **Variables** a partir de las puntuaciones de los casos válidos del archivo de datos. Esta opción exige incluir en el análisis al menos tres variables. Al seleccionar esta opción se desactiva el botón **Guardar**.

Mostrar. Las opciones de este apartado permiten controlar el tipo de resultados que mostrará el *Visor* (ambas opciones están seleccionadas por defecto):

- ☐ **Estadísticos**. El *Visor* sólo muestra las tablas de resultados. Desactivando esta opción se anula el acceso al botón **Estadísticos**.
- ☐ **Gráficos**. El *Visor* sólo muestra los gráficos. Desactivando esta opción se anula el acceso al botón **Gráficos**.

Ejemplo (Análisis de conglomerados jerárquico)

Este ejemplo muestra cómo obtener un análisis de conglomerados jerárquico con las especificaciones que el procedimiento tiene establecidas por defecto. Seguiremos utilizando el archivo de datos *Coches.sav* que se encuentra en la misma carpeta en la que se ha instalado el SPSS.

Por motivos didácticos, vamos a trabajar únicamente con 15 vehículos seleccionados al azar del archivo *Coches.sav*. Para seleccionar los casos:

- En la ventana del *Editor de datos*, seleccionar la opción **Seleccionar casos...** del menú **Datos** para acceder al cuadro de diálogo *Seleccionar casos*.
- En el cuadro de diálogo *Seleccionar casos*, marcar la **Muestra aleatoria de casos** del apartado **Seleccionar** y pulsar en el botón **Muestra...** para acceder al subcuadro de diálogo *Seleccionar casos: Muestra aleatoria*.
- En el apartado **Tamaño de la muestra**, marcar la opción **Exactamente __ casos de los primeros __ casos** e introducir los valores 15 y 406 en los respectivos cuadros de texto. Pulsar en el botón **Continuar**.

Aceptando estas selecciones, el archivo de datos queda filtrado, dejando disponibles sólo 15 de los 406 casos existentes.

Para llevar a cabo el análisis de conglomerados jerárquico:

- En el cuadro de diálogo *Análisis de conglomerados jerárquico* (ver figura 22.1), seleccionar las variables *motor* (cilindrada en cc) y *cv* (potencia) y trasladarlas a la lista **Variables**.

Aceptando estas selecciones, el *Visor* ofrece los resultados que muestran las tablas 22.1 a la 22.3 y la figura 22.2.

La tabla 22.1 muestra un *resumen de los casos procesados*: el número y porcentaje de casos válidos analizados, el número y porcentaje de casos con valores perdidos en alguna de las variables incluidas en el análisis, y el tamaño total de la muestra, que no es otra cosa que la suma de los casos válidos y los perdidos. En dos notas a pie de tabla se indica el nombre de la medida utilizada para obtener la matriz de distancias (*Distancia euclídea al cuadrado*) y el método de conglomeración utilizado (*Vinculación promedio*). La solución obtenida puede depender en gran medida de la combinación de: el tipo de medida de las distancias y el método de conglomeración.

Tabla 22.1. Resumen de los casos procesados.

Casos[a,b]					
Válidos		Perdidos		Total	
N	Porcentaje	N	Porcentaje	N	Porcentaje
15	100.0	0	.0	15	100.0

a. Distancia euclídea al cuadrado.

b. Vinculación promedio (Inter-grupos).

La tabla 22.2 muestra el *historial del proceso de conglomeración*, etapa por etapa. En cada etapa se unen dos elementos. Como la muestra analizada tiene 15 casos, sólo se realizan 14 etapas de fusión.

La columna *Conglomerado que se combina* informa sobre los conglomerados (o casos) fundidos en cada etapa. En la primera etapa se han fundido los casos 72 y 146 del archivo de datos. Como el análisis se inicia con todos los casos separados en conglomerados individuales, la primera etapa siempre se refiere a casos individuales. A partir de ese momento, estos dos casos constituyen el conglomerado «72» y son indivisibles en las etapas posteriores.

La columna (*Coeficientes*) ofrece el valor de la distancia a la que se encuentran los casos antes de la fusión. En la primera etapa, la distancia de fusión entre los casos 72 y 146 vale 0, lo que significa que se trata de casos con idénticas puntuaciones en cilindrada y potencia.

La columna *Etapa en la que el conglomerado aparece por primera vez* recoge la etapa en la que se han formado los conglomerados que se están fundiendo en cada momento. El valor 0 indica que el conglomerado correspondiente es un caso individual. Un valor mayor que 0 indica el número de etapa en la que se formó el conglomerado. En nuestro ejemplo, en la etapa 2 se funden el elemento 72 y el 231. Inspeccionando las columnas correspondientes a la

primera aparición de estos elementos encontramos un 1 y un 0, lo que significa que el elemento 72 ya apareció en la etapa 1 y es un conglomerado (el 72-146), y que el elemento 231 es un caso individual.

La columna *Próxima etapa* indica la etapa en la que el conglomerado que se acaba de formar volverá a fundirse con otros elementos. Por ejemplo, el conglomerado 72-146-231 que se ha formado en la etapa 2, vuelve a fundirse con otros elementos en la etapa 7.

Tabla 22.2. 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	72	146	.000	0	0	2
2	72	231	25.000	1	0	7
3	117	178	36.000	0	0	5
4	126	181	2885.000	0	0	8
5	117	336	9874.000	3	0	6
6	117	275	19873.000	5	0	8
7	72	174	69023.336	2	0	10
8	117	126	141724.500	6	4	11
9	171	209	168325.000	0	0	12
10	72	145	360497.500	7	0	12
11	117	333	519727.344	8	0	14
12	72	171	1990572.750	10	9	13
13	20	72	7131117.500	0	12	14
14	20	117	10964185.000	13	11	0

El *diagrama de témpanos* de la figura 22.2 resume el proceso de fusión de manera gráfica. En las cabeceras de las columnas se encuentran los números de los casos individuales (cada columna etiquetada con un número representa un caso) y en las de las filas el número de conglomerados formados en cada etapa (cada fila representa una etapa del proceso de fusión). Las etapas comienzan en la parte inferior del diagrama y van progresando hacia arriba.

Inicialmente, se parte de 15 conglomerados individuales (es decir, tantos como casos analizados). En la primera etapa se funden dos casos individuales, quedando 14 conglomerados (13 individuales y 1 doble). Los casos fundidos en la primera etapa son el 72 y el 146, lo cual está representado con una marca que une las columnas correspondientes a esos dos casos. La información de la segunda etapa se encuentra una fila más arriba, momento en el que se funden el caso 231 y el conglomerado 72-146. En la tercera etapa (la fila de 12 conglomerados), se funden los casos individuales 117 y 178. En la cuarta etapa se funden el caso 336 y el conglomerado 117-178. Y así sucesivamente.

Figura 22.2. Diagrama de *témpanos* vertical.

Número de congl.	Caso																											
	333		181		126		275		336		178		117		209		171		145		174		231		146		72		20
1	X	X	X	X	X	X	X	X	X	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	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	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	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	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		X		X	X	X	X	X	X	X		X
7	X		X	X	X	X	X	X	X	X	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		X		X	X	X	X	X	X	X		X
9	X		X	X	X		X	X	X	X	X	X	X		X		X		X		X		X	X	X	X	X		X
10	X		X	X	X		X		X	X	X	X	X		X		X		X		X		X	X	X	X	X		X
11	X		X	X	X		X		X		X	X	X		X		X		X		X		X	X	X	X	X		X
12	X		X		X		X		X		X	X	X		X		X		X		X		X	X	X	X	X		X
13	X		X		X		X		X		X		X		X		X		X		X		X	X	X	X	X		X
14	X		X		X		X		X		X		X		X		X		X		X		X		X	X	X		X

En la penúltima etapa (segunda fila del diagrama), se funde el caso individual 20 con un gran conglomerado formado por los casos 209-171-145-174-231-146-72. En la última etapa, todos los casos se funden en un único conglomerado. Obviamente, el caso 20 debe ser un caso atípico de la muestra, muy distante de los demás casos, por lo que la solución de dos conglomerados muy probablemente no será satisfactoria (a menos que el propósito del análisis haya sido

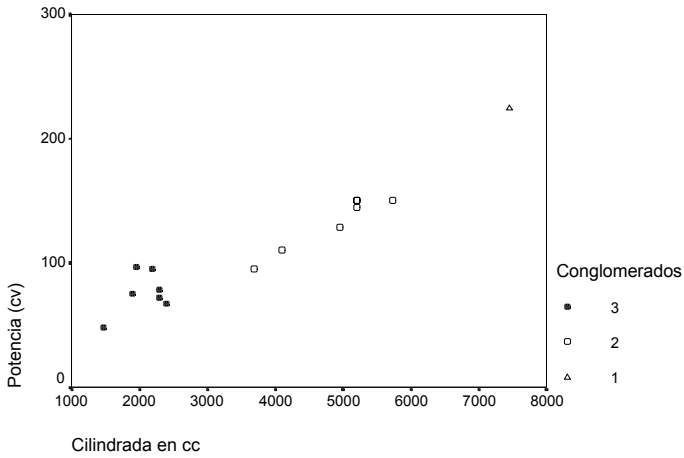
detectar el caso atípico). Y, dado que las sucesivas soluciones del análisis son jerárquicas, no podemos optar por la solución que separa los dos grandes conglomerados de la antepenúltima etapa sin mantener aislado el caso 22. Posiblemente la solución que mejor resuma la estructura de los datos sea la de tres conglomerados (ver figura 22.3). El diagrama de *témpanos* es de gran utilidad para identificar los elementos que constituyen cada una de las soluciones del análisis y cuáles han sido las formaciones previa y posterior a cada solución específica. Sin embargo, presenta el gran inconveniente de no informar en modo alguno de la distancia existente entre los conglomerados fundidos en cada etapa.

Cuando se intenta clasificar una muestra muy numerosa, el tamaño del diagrama es excesivamente ancho, lo que dificulta enormemente una inspección cómoda del mismo. En esos casos, existe la posibilidad de representar el diagrama en sentido horizontal.

El procedimiento de *Conglomerados jerárquicos* no ofrece ninguna tabla de resultados con los valores promedio de los conglomerados formados (los *centroides*) ya que su finalidad es permitir tomar una decisión sobre cuál es el número idóneo de conglomerados para representar la estructura interna de los datos. No obstante, es posible crear fácilmente la tabla de *centroides* a partir de las variables que el procedimiento permite crear en el archivo de datos (ver más abajo el apartado *Guardar*).

La figura 22.3 representa el diagrama de dispersión de los casos respecto a las dos variables de clasificación utilizadas: *motor* (cilindrada en cc) y *cv* (potencia). Los casos están marcados según el conglomerado al que han sido asignados al solicitar una solución de tres conglomerados (el diagrama se ha creado utilizando la información guardada por el procedimiento *Análisis de conglomerados jerárquico*). La solución de tres conglomerados parece satisfactoria: refleja la organización de los datos en las dos variables de clasificación. Y no parece que una solución con un mayor número de conglomerados pueda resumir mejor las distancias existentes entre los casos.

Figura 22.3. Diagrama de dispersión de las variables de agrupación.

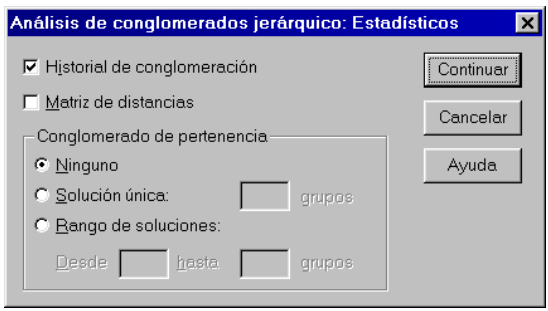


Estadísticos

Las opciones del subcuadro de diálogo estadísticos permiten solicitar estadísticos adicionales y anular la presentación del historial de conglomeración. Para acceder a estas opciones:

- Pulsar en el botón **Estadísticos...** del cuadro de diálogo *Análisis de conglomerados jerárquico* (ver figura 22.1) para acceder al subcuadro de diálogo *Análisis de conglomerados jerárquico: Estadísticos* que muestra la figura 22.4.

Figura 22.4. Cuadro de diálogo *Análisis de conglomerados jerárquico: Estadísticos*.



- **Historial de conglomeración**. Muestra una tabla que informa sobre los elementos (casos o variables) que son fundidos en cada etapa, sobre la distancia a la que se encuentran cuando son fundidos, y sobre las etapas previas y posteriores en las que aparecen los elementos implicados en cada etapa (ver tabla 22.2). Esta opción se encuentra activa por defecto; desactivándola se anula la presentación del historial de conglomeración.

☐ **Matriz de distancias**. Permite obtener la matriz de distancias entre los elementos analizados. Estas distancias pueden calcularse (ver más abajo el apartado *Medidas de distancia*) utilizando una medida de *similaridad* (grado de cercanía) o de *disimilaridad* (grado de lejanía). El tipo de matriz obtenida (de *similaridades* o de *disimilaridades*) depende de la medida seleccionada en el subcuadro de diálogo *Método* (ver más abajo el apartado *Métodos de conglomeración*).

Si el análisis contiene un gran número de elementos, la tabla puede llegar a ser muy voluminosa. La tabla 22.3 muestra la matriz de distancias para los 8 primeros casos de nuestro ejemplo. La tabla indica, en la cabecera de las columnas, que la medida utilizada es la distancia euclídea al cuadrado y, a pie de tabla, que se trata de una matriz de *disimilaridades*.

Tabla 22.3. Matriz de distancias entre los casos.

	distancia euclídea al cuadrado							
Caso	20	72	117	126	145	146	171	174
20		5045650	26669652	30880524	2967466	5045650	14222261	6294265
72	5045650		8514973	10961725	274576	0	2325601	69085
117	26669652	8514973		154458	11846565	8514973	1940978	7052274
126	30880524	10961725	154458		14705181	10961725	3190196	9293220
145	2967466	274576	11846565	14705181		274576	4197329	618237
146	5045650	0	8514973	10961725	274576		2325601	69085
171	14222261	2325601	1940978	3190196	4197329	2325601		1593800
174	6294265	69085	7052274	9293220	618237	69085	1593800	

Esta es una matriz de disimilaridades

Conglomerado de pertenencia. Las opciones de este apartado permiten controlar la presentación de la *tabla del conglomerado de pertenencia*. Esta tabla ofrece un listado de todos los casos analizados con indicación del conglomerado al que han sido asignados en cada etapa del análisis. Los casos aparecen listados en el mismo orden en el que se encuentran en el archivo de datos.

- **Ninguno**. Esta opción, que se encuentra activa por defecto, impide que el *Visor* muestre la *tabla del conglomerado de pertenencia*.
- **Solución única: __ grupos**. Permite obtener la *tabla del conglomerado de pertenencia* con la información correspondiente a una única solución: la establecida por el usuario. Para establecer el número de conglomerados de la solución que se desea obtener hay que introducir en el cuadro de texto un entero mayor que 1. La tabla 22.4 muestra la *tabla del conglomerado de pertenencia* con indicación del conglomerado al que pertenece cada caso al solicitar una solución de tres conglomerados.

Tabla 22.4. *Conglomerado de pertenencia* (solución de 3 conglomerados).

Caso	3 conglomerados
20	1
72	2
117	3
126	3
145	2
146	2
171	2
174	2
178	3
181	3
209	2
231	2
275	3
333	3
336	3

- **Rango de soluciones: Desde __ hasta __ grupos**. Permite obtener la *tabla del conglomerado de pertenencia* con la información correspondiente a varias soluciones (un *rango* de soluciones). Para establecer el número mínimo y máximo de conglomerados del rango de soluciones que se desea obtener hay que introducir, en el primer cuadro de texto, el número de conglomerados de la primera solución (la de menor número de conglomerados) y, en el segundo cuadro de texto, el número de conglomerados de la última solución (la de mayor número de conglomerados). Ambos valores deben ser enteros mayores que 1 y el primer valor debe ser menor que el segundo. La tabla 22.5 muestra la *tabla del conglomerado de pertenencia* con indicación del conglomerado al que pertenece cada caso en el rango de soluciones que van de 2 a 4 conglomerados.

Tabla 22.5. Tabla del conglomerado de pertenencia (soluciones de 2, 3 y 4 conglomerados).

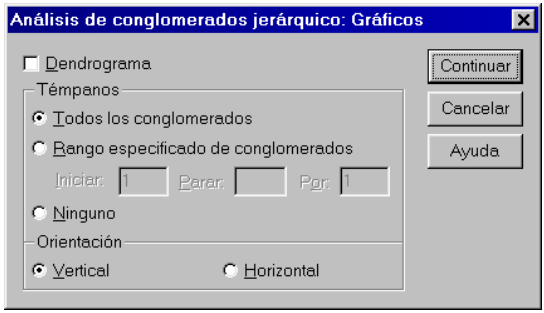
Caso	4 conglomerados	3 conglomerados	2 conglomerados
20	1	1	1
72	2	2	1
117	3	3	2
126	3	3	2
145	2	2	1
146	2	2	1
171	4	2	1
174	2	2	1
178	3	3	2
181	3	3	2
209	4	2	1
231	2	2	1
275	3	3	2
333	3	3	2
336	3	3	2

Gráficos

Las opciones del subcuadro de diálogo *Gráficos* permiten decidir qué tipos de gráficos se desea obtener. Para obtener esta información:

- Pulsar en el botón **Gráficos...** del cuadro de diálogo *Análisis de conglomerado jerárquico* (ver figura 22.1) para acceder al subcuadro de diálogo *Análisis de conglomerados jerárquico: Gráficos* que muestra la figura 22.5.

Figura 22.5. Cuadro de diálogo *Análisis de conglomerados jerárquicos: Gráficos*.



- **Dendrograma**. Muestra el dendrograma. Un dendrograma es un gráfico que combina la información del *diagrama de témpanos* y la del *historial de conglomeración*. En él, los conglomerados están representados mediante trazos horizontales y las etapas de la fusión mediante trazos verticales. La separación entre las etapas de la fusión es proporcional a la distancia a la que se están fundiendo los elementos en esa etapa (en una escala estandarizada de 25 puntos), por lo que fusiones de elementos muy próximos pueden no ser apreciables y confundirse bajo un único trazo vertical. Este gráfico es de gran utilidad para evaluar la homogeneidad de los conglomerados y facilita enormemente la decisión sobre el número óptimo de conglomerados.

Témpanos. Las opciones de este apartado permiten controlar algunos aspectos relacionados con el diagrama de témpanos:

- **Todos los conglomerados**. Esta opción, que se encuentra activa por defecto, ofrece una representación de los conglomerados de todas las etapas del análisis, es decir, una representación de todas las soluciones posibles (ver figura 22.2).
- **Rango especifico de conglomerados.** Permite seleccionar la representación de un subconjunto (rango) de soluciones. Para definir el rango de soluciones que se desea representar es necesario introducir tres valores. **Iniciar**: indica la solución con el menor número de conglomerados**. Parar**: indica la solución con el mayor número de conglomerados. **Por**: indica la cadencia (o incremento) con la se deben representar las soluciones del rango definido (manipular la cadencia de las soluciones representadas es especialmente interesante cuando el diagrama es muy extenso).
- **Ninguno**. Impide que el *Visor* muestre el diagrama de témpanos.

Orientación. Las opciones de este apartado permiten controlar la orientación del diagrama de témpanos:

- **Vertical**. Los casos se representan en las columnas y las etapas de la fusión en las filas. Esta opción se encuentra activa por defecto.
- **Horizontal**. Los casos se representan en las filas y las etapas de la fusión las columnas. Es el más apropiado para representar un gran número de elementos.

Ejemplo (Análisis de conglomerados jerárquico > Gráficos)

Este ejemplo muestra cómo obtener e interpretar los gráficos del análisis de conglomerados jerárquico. Para solicitar estos gráficos:

- En el cuadro de diálogo *Análisis de conglomerados jerárquico* (ver figura 22.1), seleccionar las variables *motor* y *cv* y trasladarlas a la lista **Variables**.
- Pulsar en el botón **Gráficos...** para acceder al subcuadro de diálogo *Análisis de conglomerados jerárquico: Gráficos* que muestra en la figura 22.5.
- Marcar la opción **Dendrograma**.
- En el apartado **Témpanos**, marcar la opción **Rango especificado de conglomerados** e introducir el valor 4 en el cuadro de texto **Parar**.

Aceptando estas selecciones, el *Visor* construye los gráficos que muestran las figuras 22.6 y 22.7.

La figura 22.6 ofrece el *diagrama de témpanos* vertical referido a las soluciones del rango establecido (1-4). La solución de un sólo conglomerado carece de interés y sólo se utiliza como punto de referencia (cualquier análisis de conglomerados jerárquico desemboca en esa solución, pero agrupar los casos en un único conglomerado es lo miso que no realizar el análisis).

Figura 22.6. Diagrama de témpanos del rango de soluciones 1 a 4.

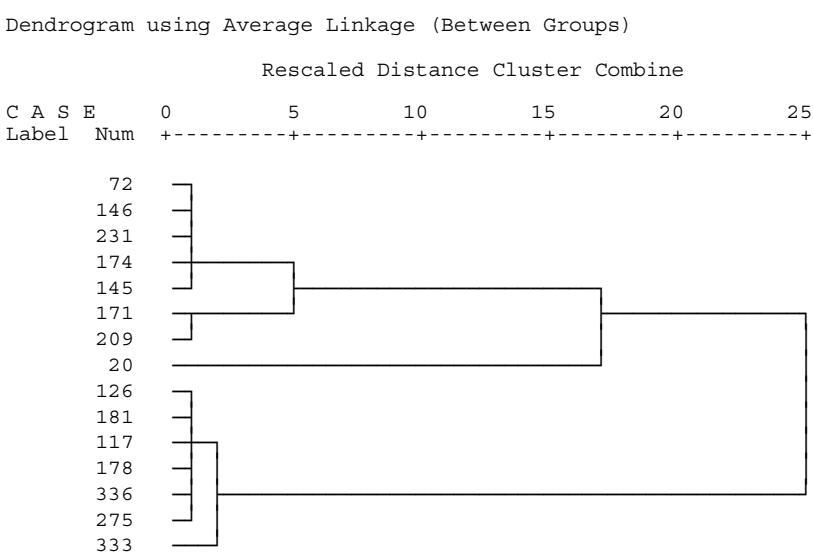
Número de congl.	Caso																															
	333		181		126		275		336		178		117		209		171		145		174		231		146		72		20			
1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
2	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
3	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	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	X	X	X	X	X	X	X	X	X			X	X

La solución de 2 conglomerados muestra un conglomerado formado por los casos 20-72-146-231-174-145-171-209 y otro conglomerado formado por todos los casos restantes. La solución de 3 conglomerados muestra un conglomerado formado por el caso 20, otro formado por los casos 72-146-231-174-145-171-209 (el conglomerado «72») y otro formado por los casos 117-178-336-275-126-181-333 (el conglomerado «117»). La solución de 4 conglomerados coincide con la de 3 excepto en lo referente al conglomerado «72», que ahora está dividido en dos: los casos 72-146-231-174-145 (conglomerado «72») por un lado, y los casos 171-209 (conglom-

erado «171») por otro. Exactamente esta misma información es la que ofrece la *tabla del conglomerado de pertenencia* (ver tabla 22.5).
La diferencia entre las soluciones con 3 y 4 conglomerados radica en la división del conglomerado «72» en dos conglomerados: el «72» y el «171». Ahora bien, puesto que desconocemos la distancia entre estos dos últimos conglomerados, resulta difícil valorar la pertinencia de efectuar o no esa división. No obstante, en el dendrograma de la figura 22.7 podemos encontrar información adicional.

La figura 22.7 muestra el dendrograma de los 15 vehículos de nuestro ejemplo. En un dendrograma, además de estar representadas las etapas del proceso de fusión, también lo están las distancias existentes entre los elementos fundidos. Pero las distancias no están representadas en su escala original sino en una escala estandarizada de 25 puntos. Las líneas verticales identifican elementos fundidos (conglomerados); y la posición de las líneas verticales indica la distancia existente entre los elementos fundidos.

Figura 22.7. Dendrograma.



Por la tabla del *historial de conglomeración* (tabla 22.2) sabemos ya que la mayor distancia entre conglomerados vale 10.964.185,00 (entre el conglomerado «20 y el «117») y la menor 0,00 (entre el vehículo 72 y el 146); Pues bien, puesto que las distancias representadas en el dendrograma están reescaladas, a la distancia mayor (10.964.185,00) le corresponde un valor de 25 y a la menor (0,00) un valor de 1.

No obstante, el dendrograma suele asignar también una distancia de 1 a las fusiones de las primeras etapas (pues en ellas las distancias suelen ser muy pequeñas en comparación con las distancias de las etapas finales), lo cual impide averiguar el orden en el que se han producido las primeras fusiones (para ello hay que recurrir a la tabla del *historial de conglomeración*). El dendrograma de la figura 22.7, por ejemplo, asigna una distancia de 1 a los elementos fundidos en las 10 primeras etapas (a pesar de que el proceso de fusión consta de 14 etapas, sólo hay 4 líneas verticales con valor distinto de 1, lo cual puede dar una idea de la precisión con la que el dendrograma representa las distancias). Así pues, puesto que en la parte baja (izquierda) de la escala no es posible distinguir el orden de fusión de los conglomerados, conviene señalar que, en esa zona, el dendrograma está produciendo una falsa impresión. Dado que el proceso de aglomeración se realiza siempre de manera binaria (fundiendo dos elementos en cada etapa), si pudiéramos utilizar una lupa para aumentar la resolución de la zona baja de la escala, los tamos verticales aparecerían escalonados en distintas posiciones de la escala de 25 puntos.

La primera etapa de la fusión se representa en el extremo izquierdo del dendrograma; la última etapa, en el extremo derecho. En el dendrograma de la figura 22.7 puede observarse que el caso 20 se encuentra muy distante del resto de los casos, pues la primera fusión en la que interviene se produce a una distancia de 17 puntos (la distancia que va desde el origen de la escala hasta el punto 17). Por el contrario, los conglomerados «72» y «171» se funden a una distancia de 5 puntos (la distancia que va desde el origen de la escala hasta el punto 5). Los valores exactos de las distancias entre estos elementos se encuentran en la tabla del *historial de conglomeración* (ver tabla 22.2): la distancia del caso 20 al conglomerado «72» vale 7.131.117,50, y la distancia entre los conglomerados «72» y «171» vale 1.990.572,75. Podemos comprobar que la escala estandarizada del dendrograma refleja con suficiente precisión la proporcionalidad existente entre las distancias originales.

Las fusiones que se producen cerca del origen de la escala (izquierda) indican que el conglomerado formado es bastante homogéneo. Por el contrario, La fusiones que se producen en la zona final de la escala (derecha) indican que el conglomerado formado es bastante heterogéneo. Para tomar una decisión sobre cuál ha de ser el número de conglomerados idóneo puede recorrerse el dendrograma de derecha a izquierda y detener la atención allí donde las

líneas verticales están unidas al origen de la escala con trazos horizontales cortos (o no demasiado largos). Tras esto, bastará con seguir cada línea horizontal hacia la izquierda para identificar los casos que componen cada conglomerado.

En el dendrograma de nuestro ejemplo (figura 22.7) parece razonable adoptar una solución de 3 conglomerados:

Conglomerado 1: 20
Conglomerado 2: 72-146-231-174-145-171-209
Conglomerado 3: 126-181-117-178-336-275-333

Los números asignados a estos tres conglomerados se corresponden con los asignados por el propio procedimiento (ver, por ejemplo, las tablas 22.4 y 22.5).

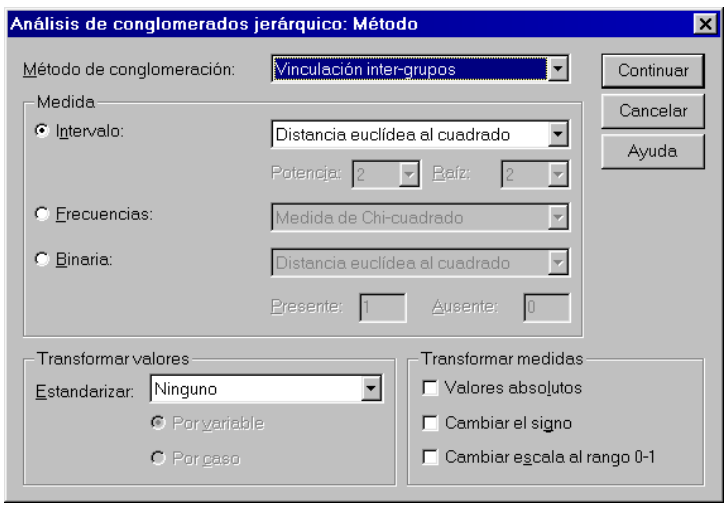
Por supuesto, si se desea obtener un número preestablecido de conglomerados, bastará con partir el dendrograma verticalmente por donde se encuentre ese número de líneas verticales y seguir cada línea horizontal hacia la izquierda para identificar los casos que componen cada conglomerado.

Método

Las opciones del cuadro de diálogo *Método* permiten seleccionar un *método de conglomeración* y el tipo de medida que se desea utilizar para evaluar las distancias entre los elementos. También permiten transformar las puntuaciones originales y las medidas de distancia resultantes. Las selecciones de este cuadro de diálogo determinan la solución obtenida. Distintas combinaciones de opciones pueden dar como resultado soluciones muy distintas. Para personalizar estas opciones:

- Pulsar en el botón **Método...** del cuadro de diálogo *Análisis de conglomerados jerárquico* (ver figura 22.1) para acceder al subcuadro de diálogo *Análisis de conglomerados jerárquico: Método* que muestra la figura 22.8.

Figura 22.8. Cuadro de diálogo *Análisis de conglomerados jerárquico: Método*.



Método de conglomeración

La primera opción del cuadro de diálogo permite seleccionar un *método de conglomeración*. Según hemos señalado ya, el análisis de conglomerados jerárquico siempre evoluciona paso a paso, uniendo en cada paso los dos elementos de la matriz de distancias que se encuentran más próximos entre sí. En cada paso se funden dos elementos o grupos de elementos.

Una vez calculada la matriz de distancias, los dos elementos más próximos (los más similares o menos distantes) son fundidos en un mismo conglomerado. Estos dos casos que constituyen el primer conglomerado (en este momento son sólo dos casos por tratarse del primer paso del procedimiento) constituyen una unidad que, como tal, posee su propia distancia respecto al resto de los elementos de la matriz de distancias. La matriz inicial de los *n*x*n* sujetos (o *p*x*p* variables) cambia (pues dos de sus filas –y dos de sus columnas– han sido fundidas en una) transformándose en una matriz (*n*–1)(*n*–1). Tras recalcular las distancias, en la siguiente etapa del análisis se vuelven a seleccionar los dos elementos de la matriz más próximos entre sí y son fundidos en un nuevo conglomerado. Por supuesto, los dos elementos fundidos en esta segunda etapa pueden ser dos casos individuales o un caso individual y el conglomerado ya formado en la primera etapa. En este momento, la matriz de distancias de dimensiones (*n*–1)x(*n*-1) se transforma en una matriz de distancias de dimensiones (*n*–2)x(*n*–2), lo que exige volver a calcular las distancias del nuevo conglomerado respecto al resto de elementos de la matriz. El proceso continua paso a paso hasta que, finalmente, se consigue fundir en un único conglomerado a todos los elementos de la matriz de distancias (de dimensiones finales 2x2). En ese punto termina el análisis. Pues bien, los *métodos de conglomeración* son los procedimientos mediante los cuales es posible volver a calcular las distancias entre los nuevos elementos en cada etapa del proceso de fusión.

Lógicamente, en todo este proceso de fusión no existe una solución única, sino tantas como pasos da el proceso. La decisión sobre qué solución se considera más satisfactoria puede tomarse en cualquier etapa del proceso, pero lo más lógico y habitual es postergar esta decisión hasta el momento en que el análisis ha concluido.

Conviene señalar que el método de conglomeración utilizado para recalcular las distancias en cada etapa del proceso de fusión puede determinar de manera sustantiva la calidad de la solución alcanzada. La idoneidad y eficacia del método de conglomeración seleccionado dependerá en gran medida de la propia estructura de los datos y de la forma multivariante de la nube de puntos.

Método de vinculación por el vecino más próximo

El método de *vinculación simple*, *enlace simple*, o *por el vecino más próximo*, comienza seleccionando y fundiendo los dos elementos de la matriz de distancias que se encuentran más próximos. La distancia de este nuevo conglomerado respecto a los restantes elementos de la matriz se calcula como la menor de las distancias entre cada elemento del conglomerado y el resto de elementos de la matriz. En los pasos sucesivos, la distancia entre dos conglomerados se calcula como la distancia entre sus dos elementos más próximos. Así, la distancia d_{AB} entre los conglomerados A y B se calcula mediante:

$$d_{AB} = \min (d_{ij})$$

donde d_{ij} es la distancia entre los elementos i y j, el primero perteneciente al conglomerado A y el segundo al conglomerado B.

Método de vinculación por el vecino más lejano

El método de *vinculación completa*, *enlace completo*, o *por el vecino más lejano*, se comporta de manera opuesta al anterior. La distancia entre dos conglomerados se calcula como la distancia entre sus dos elementos más alejados. Es decir, la distancia entre dos conglomerados A y B se calcula como:

$$d_{AB} = \max (d_{ij})$$

Método de vinculación inter-grupos

El método de *vinculación promedio*, o de *vinculación inter-grupo*, presenta la ventaja sobre los dos métodos anteriores de aprovechar la información de todos los miembros de los dos conglomerados que se comparan. La distancia entre dos conglomerados se calcula como la distancia promedio existente entre todos los pares de elementos de ambos conglomerados:

$$d_{AB} = \frac{1}{n_A n_B} \sum_{i \in A} \sum_{j \in B} d_{ij}$$

Tanto este método como el de Ward son sensibles a posibles transformaciones monótonas de los datos.

Método de Ward

Este método fue propuesto por Ward (1963), quien argumentó que los conglomerados debían constituirse de tal manera que, al fundirse dos elementos, la pérdida de información resultante de la fusión fuera mínima. En este contexto, la *cantidad de información* se cuantifica como la suma de las distancias al cuadrado de cada elemento respecto al centroide del conglomerado al que pertenece (*SCE = Suma de Cuadrados Error*). Para ello, se comienza calculando, en cada conglomerado, el vector de medias de todas la variables, es decir, el *centroide multivariante*. A continuación, se calculan las distancias euclídeas al cuadrado entre cada elemento y los centroides (vector de medias) de todos los conglomerados. Por último, se suman las distancias correspondientes a todos los elementos.

En cada paso se unen aquellos conglomerados (o elementos) que dan lugar a un menor incremento de la *SCE*, es decir, de la suma de cuadrados de las distancias intra-conglomerado. La *SCE* se define como:

$$SCE = \sum_{j=1}^{k} \left(\sum_{i=1}^{n_j} X_{ij}^2 - \frac{1}{n_j} \left(\sum_{i=1}^{n_j} X_{ij} \right)^2 \right)$$

Método de agrupación de centroides

El método de *agrupación de centroides* calcula la distancia entre dos conglomerados como la distancia entre sus vectores de medias. Con este método, la matriz de distancias original sólo se utiliza en la primera etapa. En las etapas sucesivas se utiliza la matriz de distancias actualizada en la etapa previa. En cada etapa, el algoritmo utiliza la información de los dos conglomerados (o elementos) fundidos en la etapa previa y el conglomerado (o elemento) que se intentará fundir en esa etapa. La distancia entre el conglomerado *AB* y el elemento *C* se calcula como:

$$d_{(AB)C} = \frac{n_A}{n_A + n_B} d_{AC} + \frac{n_B}{n_A + n_B} d_{BC} - \frac{n_A n_B}{(n_A + n_B)^2} d_{AB}$$

Una desventaja de este método es que la distancia entre dos conglomerados puede disminuir a medida que progresa el análisis, ya que los conglomerados fundidos en los últimos pasos son más diferentes entre sí que los que se funden en las primeras etapas.

En este método, el centroide de un conglomerado es la combinación ponderada de los dos centroides de sus dos últimos elementos (o conglomerados), siendo las ponderaciones proporcionales a los tamaños de los conglomerados.

Método de agrupación de medianas

En el método de *agrupación de medianas*, los dos conglomerados (o elementos) que se combinan reciben idéntica ponderación en el cálculo del nuevo centroide combinado, independientemente del tamaño de cada uno de los conglomerados (o elementos). Esto permite que, a la hora de caracterizar a los conglomerados resultantes, los conglomerados pequeños tengan la misma importancia que los conglomerados grandes. Dado un conglomerado *AB* y un elemento *C*, la nueva distancia del conglomerado al elemento se calcula como:

$$d_{(AB)C} = \frac{d_{AC} + d_{BC}}{2} - \frac{d_{AB}}{4}$$

Al igual que en el procedimiento anterior, la matriz de distancias utilizada en cada etapa para los cálculos es la matriz del paso previo. Para una discusión más detallada de los métodos de aglomeración puede consultarse Anderberg (1973).

Medidas de distancia

Uno de los aspectos clave del análisis de conglomerados es la elección de la medida que se desea utilizar para cuantificar la *distancia* entre los elementos. El procedimiento *Análisis de conglomerados jerárquico* permite elegir entre un gran número de medidas de distancia que se diferencian por el tipo de datos para el que han sido diseñadas: cuantitativos, categóricos, dicotómicos.

Estas medidas también se diferencian por el tipo de distancia evaluada: similaridad o disimilaridad. Las medidas de ***similaridad*** evalúan el grado de parecido o proximidad existente entre dos elementos. Los valores más altos indican mayor parecido o proximidad entre los elementos comparados: cuando dos elementos se encuentran juntos, el valor de las medidas es máximo. El coeficiente de correlación de Pearson es, quizá, la medida de similaridad más ampliamente utilizada.

Las medidas de ***disimilaridad*** ponen el énfasis sobre el grado de diferencia o lejanía existente entre dos elementos. Los valores más altos indican mayor diferencia o lejanía entre los elementos comparados: cuando dos elementos se encuentran juntos, la distancia es nula. Las medidas de disimilaridad son las que han pasado al vocabulario común con la acepción de *medidas de distancia*. La distancia euclídea (la longitud del segmento lineal que une dos elementos) es, quizá, la medida de disimilaridad más conocida.

Las opciones del apartado **Medida** (ver figura 22.8) permiten seleccionar la medida que se desea utilizar para evaluar la *distancia* entre los elementos. Las medidas se encuentran agrupadas en función del tipo de datos para el que son pertinentes (todas las variables seleccionadas para el análisis deben compartir el mismo tipo de nivel de medida). Conviene no olvidar que las elecciones que se hagan en este apartado afectarán al cálculo de la matriz de distancias y, consecuentemente, pueden condicionar de forma importante las soluciones alcanzadas.

En el listado que se ofrece a continuación, las fórmulas reciben el mismo nombre que les asigna la sintaxis del SPSS.

Intervalo

Esta opción incluye medidas de similaridad y disimilaridad para datos cuantitativos obtenidos con una escala de medida de intervalo o razón.

- ***Distancia euclídea***. Medida de disimilaridad utilizada por defecto para datos de intervalo. Raíz cuadrada de la suma de los cuadrados de las diferencias entre los valores de las variables:

$$EUCLID\,(X,Y) = \sqrt{\sum_i (X_i - Y_i)^2}$$

- ***Distancia euclídea al cuadrado***. Medida de disimilaridad. Suma de los cuadrados de las diferencias entre los valores de las variables:

$$SEUCLID\,(X,Y) = \sum_i (X_i - Y_i)^2$$

- ***Coseno***. Medida de similaridad. Medida estrechamente relacionada con el coeficiente de correlación de Pearson. Es el coseno del ángulo formado por dos vectores de puntuaciones. Tiene un máximo de 1 y un mínimo de -1:

$$COSINE\,(X,Y) = \frac{\sum_i^n X_i Y_i}{\sqrt{\left(\sum_i^n X_i^2\right)\left(\sum_i^n Y_i^2\right)}}$$

- ***Correlación de Pearson***. Medida de similaridad angular con las variables en escala tipificada. Se trata de una medida típica de relación lineal entre variables. Toma valores entre –1 y 1:

$$CORRELATION\,(X,Y) = \frac{\sum_i^n z_{x_i} z_{y_i}}{n-1}$$

donde n es el tamaño de la muestra y z_x y z_y son las puntuaciones tipificadas del sujeto i en las variables X e Y, que son las variables entre las que se calcula la distancia.

- ***Chebychev***. Medida de disimilaridad. Diferencia más grande en valor absoluto entre los valores de dos variables:

$$CHEBYCHEV(X,Y) = máx_i |X_i - Y_i|$$

- ***Bloques***. Medida de disimilaridad. También llamada distancia *absoluta*, distancia de *ciudad*, de *Manhatan*, y del *taxista*. Es la suma de los valores absolutos de las diferencias entre los valores de dos variables:

$$BLOCK(X,Y) = \sum_i |X_i - Y_i|$$

- ***Minkowsky***. Medida de disimilaridad basada en la distancia euclídea. Raíz de orden p de la suma de las potencias de orden p de los valores absolutos de las diferencias entre los valores de dos variables:

$$MINKOWSKI(X,Y) = \left(\sum_i |X_i - Y_i|^p \right)^{\frac{1}{p}}$$

donde p es cualquier número entero positivo.

- ***Personalizada***. Medida de disimilaridad basada en la distancia euclídea. Raíz de orden r de la suma de las potencias de orden p de los valores absolutos de las diferencias entre los valores de dos variables:

$$POWER(X,Y) = \left(\sum_i |X_i - Y_i|^p \right)^{\frac{1}{r}}$$

donde p y r son dos números enteros positivos cualesquiera.

Frecuencias

Esta opción incluye dos medidas de disimilaridad para datos categóricos. Ambas se basan en el estadístico *chi-cuadrado* de independencia para tablas de contingencia bidimensionales.

- ***Chi-cuadrado***. Medida de disimilaridad utilizada por defecto para datos categóricos. Se basa en las divergencias existentes entre las frecuencias observadas y el modelo de independencia. La magnitud de esta medida depende del tamaño muestral. Los valores esperados se obtienen asumiendo independencia entre las variables:

$$CHISQ(X,Y) = \sqrt{\sum_i [X_i - E(X_i)]^2/E(X_i) + \sum_i [Y_i - E(Y_i)]^2/E(Y_i)}$$

- ***Phi-cuadrado***. Medida de disimilaridad. La medida *chi-cuadrado* normalizada por la raíz cuadrada del número de casos. Su valor no depende del tamaño muestral:

$$PHI2(X,Y) = CHISQ(X,Y) \Big/ \sqrt{n}$$

Binaria

Las medidas para datos *binarios* se utilizan con variables dicotómicas, es decir, con variables cuyos valores reflejan la presencia o ausencia de la característica medida. Generalmente, la *presencia* de la característica se codifica con el valor 1 y la *ausencia* con el valor 0. La tabla 22.6 muestra una tabla de contingencia 2x2 con la notación utilizada al resumir los datos referidos a dos variables dicotómicas.

Tabla 22.6. Tabla de contingencia con dos variables dicotómicas.

		Variable Y_i: 1	Variable Y_i: 0	
Variable X_i	1	a	b	a + b
	0	c	d	c + d
		a + c	b + d	n

En la tabla, *n* se refiere al número total de casos, *a* se refiere al número de casos que comparten la presencia de ambas características, *d* se refiere al número de casos que comparten la ausencia de ambas características (*a* y *d* son las *concordancias*), y *b* y *c* se refieren el número de casos que presentan una característica y no la otra (las *discordancias*).

Existe un gran número de medidas para calcular la distancia entre los elementos de una tabla de contingencia de estas características. Estas medidas difieren, básicamente, en la importancia que conceden a cada casilla de la tabla. Se considera que dos elementos son tanto más similares entre sí cuanto mayor número de presencias o ausencias comparten. Pero las presencias y las ausencias no tienen por qué tener la misma importancia al valorar la similaridad. Si dos sujetos responden sí a la pregunta "¿Ha padecido alguna enfermedad grave en los últimos tres meses?", esa concordancia posee mucho mayor valor informativo que si ambos sujetos responden no. Sin embargo, si dos sujetos responden sí a la pregunta "¿Ha ido alguna vez a la playa en verano?", esa concordancia posee mucho menor valor informativo que si ambos sujetos responden no.

Por esta razón, algunas medidas no tienen en cuenta las ausencias conjuntas (*d*); otras conceden más importancia a las concordancias que a las discordancias, o al revés; otras sólo tienen

en cuenta las presencias conjuntas; otras, las ausencias; etc. Puesto que cada una de ellas pone el énfasis en un aspecto concreto de la tabla, la decisión sobre qué medida conviene utilizar no es una cuestión trivial. Sobre todo si tenemos en cuenta que muchas de ellas no arrojan resultados equivalentes (no son monótonas entre sí, pudiendo darse inversiones de valores en los elementos comparados) y que el cambio de codificación de las presencias-ausencias (el cambio de ceros por unos y de unos por ceros) también puede hacer variar el resultado.

Las fórmulas que se ofrecen a continuación están diseñadas para evaluar la distancia entre *dos variables a partir de un cierto número de casos*. No obstante, intercambiando en la tabla 22.6 las variables *X* e *Y* por los casos *i* y *j*, las fórmulas que se ofrecen pueden utilizarse para calcular la distancia entre *dos casos a partir de un cierto número de variables*.

- ***Distancia euclídea*** . Medida de disimilaridad. Versión binaria de la distancia euclídea. Su valor mínimo es 0, pero no tiene máximo:

$$BEUCLID\,(X,Y) = \sqrt{b + c}$$

- ***Distancia euclídea al cuadrado***. Medida de disimilaridad. Su valor mínimo es 0, pero no tiene máximo:

$$BSEUCLID\,(X,Y) = b + c$$

- ***Diferencia de tamaño***. Medida de disimilaridad. Su valor mínimo es 0, pero no tiene máximo:

$$SIZE\,(X,Y) = \frac{(b - c)^2}{(a + b + c + d)^2}$$

- ***Diferencia de configuración***. Medida de disimilaridad. Toma valores entre 0 y 1:

$$PATTERN\,(X,Y) = \frac{bc}{(a + b + c + d)^2}$$

- ***Varianza***. Medida de disimilaridad. Su valor mínimo es 0, pero no tiene máximo:

$$VARIANCE\,(X,Y) = \frac{b + c}{4(a + b + c + d)^2}$$

- ***Dispersión***. Medida de similaridad. Toma valores entre 0 y 1:

$$DISPER\,(X,Y) = \frac{ad - bc}{(a + b + c + d)^2}$$

- ***Forma***. Medida de disimilaridad. No tiene límite inferior ni superior:

$$BSHAPE\,(X,Y) = \frac{(a + b + c + d)(b + c) - (b - c)^2}{(a + b + c + d)^2}$$

- ***Concordancia simple*** o *emparejamiento simple*. Medida de similaridad. Es el cociente entre el número de concordancias y el número total de características:

$$SM\,(X,Y) = \frac{a + b}{n}$$

- ***Coeficiente Phi (de cuatro puntos)***. Medida de similaridad. Versión binaria de coeficiente de correlación de Pearson. Es la medida de asociación más utilizada para datos binarios. Toma valores entre 0 y 1:

$$PHI\,(X,Y) = \frac{ad - bc}{\sqrt{(a + b)(a + c)(b + d)(c + d)}}$$

- ***Lambda de Goodman y Kruskal***. Medida de similaridad. Evalúa el grado en que el estado de una característica en una variable (presente o ausente) puede predecirse a partir del estado de esa característica en la otra variable. En concreto, *lambda* mide la reducción proporcional del error de predicción que se consigue al utilizar una variable como predictora de la otra cuando las direcciones de la predicción son de igual importancia. *Lambda* toma valores entre 0 y 1:

$$LAMBDA\,(X,Y) = \frac{t_1 - t_2}{2(a + b + c + d)}$$

donde:

$$t_1 = máx(a, b) + máx(c, d) + máx(a, c) + máx(b, d)$$
$$t_2 = máx(a+c, b+d) + máx(a+b, c+d)$$

- ***D de Andemberg***. Medida de similaridad. Al igual que *lambda*, evalúa la capacidad predictiva de una variable sobre otra. Y, al igual que *lambda*, mide la reducción en la probabilidad del error de predicción cuando una de las variables es utilizada para predecir la otra. Toma valores entre 0 y 1:

$$D\,(X,Y) = \frac{t_1 + t_2}{2(a + b + c + d)}$$

donde t_1 y t_2 se definen de la misma manera que en la medida *lambda* de Goodman y Kruskal.

- ***Dice***. Medida de similaridad. También conocida como medida de Czekanowski o de Sorenson. No tiene en cuenta las ausencias conjuntas, pero concede valor doble a las presencias conjuntas:

$$DICE\,(X,Y) = \frac{2a}{2a + b + c}$$

- ***Hamann***. Medida de similaridad. Probabilidad de que la característica medida se encuentre en el mismo estado en las dos variables (presente o ausente en ambas), menos la probabilidad de que la característica se encuentre en distinto estado en ambas variables (presente en una y ausente en otra). Toma valores entre –1 y 1:

$$HAMANN\,(X,Y) = \frac{(a + d) - (b + c)}{a + b + c + d}$$

- ***Jaccard***. Medida de similaridad. Medida conocida también como *tasa de similaridad*. No tiene en cuenta las ausencias conjuntas (*d*) y pondera por igual las concordancias y las discordancias:

$$JACCARD\,(X,Y) = \frac{a}{a + b + c}$$

- ***Kulczynski 1***. Medida de similaridad. Excluye las ausencias conjuntas del numerador y las concordancias del denominador. Esta medida tiene un límite inferior de 0, pero no tiene límite superior. Y no es posible calcularla si no existen discordancias (es decir, si $b = c = 0$). En ese caso, el procedimiento asigna un valor arbitrario de 9999,999 como límite superior tanto si no hay discordancias como si el valor de la medida excede de ese valor:

$$K1\,(X,Y) = \frac{a}{b + c}$$

- ***Kulczynski 2***. Medida de similaridad. Probabilidad condicional de que la característica medida esté presente en una variable dado que lo está en la otra. La medida final es el promedio de las dos medidas posibles: $P(X|Y)$ y $P(Y|X)$. Toma valores entre 0 y 1:

$$KK2\,(X,Y) = \frac{a/(a + b) + a/(a + c)}{2}$$

- ***Lance y Williams***. Medida de disimilaridad. También se conoce como el *coeficiente no métrico de Bray-Curtis*. Toma valores entre 0 y 1:

$$BLWMN\,(X,Y) = \frac{b + c}{2a + b + c}$$

- ***Ochiai***. Medida de similaridad. Versión binaria del coseno. Toma valores entre 0 y 1:

$$OCHIAI\,(X,Y) = \sqrt{\left(\frac{a}{a+b}\right)\left(\frac{a}{a+c}\right)}$$

- ***Rogers y Tanimoto***. Medida de similaridad. Incluye las ausencias conjuntas tanto en el numerador como en el denominador y concede doble valor a las disimilaridades:

$$RT1\,(X,Y) = \frac{a + d}{a + d + 2(b + c)}$$

- ***Russel y Rao***. Medida de similaridad. Es el producto escalar binario:

$$RR\,(X,Y) = \frac{a}{n}$$

- ***Sokal y Sneath 1***. Medida de similaridad. Incluye las ausencias conjuntas tanto en el numerador como en el denominador y concede doble valor a las similaridades:

$$SS1\,(X,Y) = \frac{2(a + d)}{2(a + d) + b + c}$$

- ***Sokal y Sneath 2***. Medida de similaridad. Excluye las ausencias conjuntas y concede doble valor a las disimilaridades:

$$SS2\,(X,Y) = \frac{a}{a + 2(b + c)}$$

- ***Sokal y Sneath 3***. Medida de similaridad. Excluye las concordancias del denominador. Esta medida tiene un límite inferior de 0, pero no tiene límite superior. Y no es posible calcularla si no existen discordancias (es decir, si $b = c = 0$). En ese caso, el programa asigna un valor arbitrario de 9999,999 como límite superior tanto si no hay discordancias como si el valor de la medida excede de ese valor:

$$SS3\,(X,Y) = \frac{a + d}{b + c}$$

- ***Sokal y Sneath 4***. Medida de similaridad. Probabilidad condicional de que la característica medida se encuentre en el mismo estado (presente o ausente) en las dos variables. La medida final es el promedio de las dos medidas posibles: $P(X|Y)$ y $P(Y|X)$. Toma valores entre 0 y 1:

$$SS4\,(X,Y) = \frac{a/(a + b) + a/(a + c) + d/(b + d) + d/(c + d)}{4}$$

- ***Sokal y Sneath 5***. Medida de similaridad. Toma valores entre 0 y 1:

$$SS5\,(X,Y) = \frac{ad}{\sqrt{(a + b)(a + c)(b + d)(c + d)}}$$

- ***Y de Yule***. Medida de similaridad. El coeficiente de coligación Y de Yule es una función de los productos cruzados en una tabla 2x2. Toma valores entre –1 y 1:

$$Y\,(X,Y) = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}$$

- ***Q de Yule***. Medida de similaridad. Versión para tablas 2x2 de la medida ordinal *gamma* de Goodman y Kruskal. También es una función de los productos cruzados. Toma valores entre –1 y 1:

$$Q(X,Y) = \frac{ad - bc}{ad + bc}$$

Para una descripción más detallada de todas estas as medidas de distancia puede consultarse Anderberg (1973) o Romesburg (1984).

Transformar valores

Muchas de las medidas de distancia (por ejemplo, la distancia euclídea y el resto de medidas derivadas de ella) no son invariantes respecto a la métrica de los datos, ya que las diferencias existentes entre las variables con puntuaciones muy altas pueden anular las diferencias existentes entre las variables con puntuaciones bajas. En el archivo *Coches.sav* que venimos utilizando, la variable *motor* (cilindrada en cc) adopta valores comprendidos entre 66 y 7456; por el contrario, la variable *cilindr* (número de cilindros) adopta valores comprendidos entre 3 y 8. Lógicamente, al calcular la distancia entre dos vehículos, la diferencia en las cilindradas tendrá mucha más presencia que la diferencia en el número de cilindros.

Para resolver este problema suele recomendarse no utilizar las puntuaciones directas de las variables (los datos en bruto) sino las puntuaciones transformadas a escalas del mismo rango (escala 0-1, escala típica, etc.).

Las opciones del apartado **Transformar valores** (ver figura 22.8) permiten elegir entre distintos tipos de transformación, así como si la transformación se desea hacer tomando como referencia los casos o las variables. La transformación elegida se aplica a todos los elementos del análisis. Estas opciones no están disponibles cuando se selecciona una medida de distancia binaria. En todos los casos es posible seleccionar los *elementos* (casos o variables) que se desea transformar. Las opciones de transformación son:

- **Ninguno**. No se aplica ningún método de transformación.
- **Puntuaciones *Z***. A cada valor se le resta la media del elemento y esa diferencia se divide por la desviación típica del elemento. Se obtienen valores estandarizados con media 0 y desviación típica 1. Si la desviación típica vale 0, se asigna un 0 a todos los valores.
- **Rango -1 a 1**. Cada valor se divide por el rango o amplitud del elemento. Se obtienen valores estandarizados con amplitud 2 en una escala cuya unidad de medida es el rango o amplitud del elemento. Si el rango o amplitud vale cero, no se efectúa la transformación.
- **Rango 0 a 1**. A cada valor se le resta el valor más pequeño del elemento y esa diferencia se divide entre el rango o amplitud del elemento. Se obtienen valores estandarizados comprendidos entre 0 y 1. Si el rango vale 0, se asigna un 0,5 a todos los valores.

- **Magnitud máxima de 1**. Cada valor se divide por el valor más grande del elemento. Se obtienen valores estandarizados con un máximo de 1 y un mínimo variable pero nunca menor de 0. Si el valor más grande vale 0, se divide por el valor absoluto del valor más pequeño y se suma 1.
- **Media 1**. Divide cada valor por la media del elemento. Se obtienen valores estandarizados con media igual a 1, y en una escala cuya unidad de medida es la media del elemento. Si la media vale 0, se suma un 1 a todos los valores.
- **Desviación típica 1**. Divide cada valor por la desviación típica del elemento. Se obtienen valores estandarizados con desviación típica igual a 1 y en una escala cuya unidad de medida es la desviación típica media del elemento. Si la desviación típica vale 0, no se efectúa la transformación.

Transformar medidas

Las opciones del apartado **Transformar medidas** (ver figura 22.8) permiten transformar los valores de la matriz de distancias. Si se selecciona más de una transformación, el procedimiento las realiza en el siguiente orden:

- **Valores absolutos**. Valor absoluto de las distancias calculadas.
- **Cambiar el signo**. Cambia el signo de las distancias calculadas, transformando las medidas de similaridad en medidas de disimilaridad y viceversa.
- **Cambiar escala al rango 0-1**. Se resta a todos los valores de la matriz de distancias la distancia más pequeña y cada nueva distancia se divide por el rango o amplitud de todas las distancias. Se obtienen así valores que oscilan entre 0 y 1.

Ejemplo (Análisis de conglomerados > Método)

Este ejemplo muestra cómo llevar a cabo un análisis de conglomerados jerárquico de variables y se discute la influencia del método de conglomeración y de la medida de distancia utilizados sobre el resultado del análisis.

- En el cuadro de diálogo *Análisis de conglomerados jerárquicos* (ver figura 22.1), trasladar las variables *consumo*, *motor* (cilindrada), *cv* (potencia), *peso*, *acel* (aceleración) y *cilindr* (nº de cilindros) a la lista **Variables**.
- Seleccionar la opción **Variables** del apartado **Conglomerar**.
- Pulsar en el botón **Estadísticos...** para acceder al subcuadro de diálogo *Análisis de conglomerados jerárquico: Estadísticos* (ver figura 22.4).
- Marcar la opción **Matriz de distancias**. Pulsar el botón **Continuar**.
- Pulsar en el botón **Gráficos...** para acceder al subcuadro de diálogo *Análisis de conglomerados jerárquico: Gráficos* (ver figura 22.5).
- Marcar la opción **Dendrograma**. Pulsar el botón **Continuar**.

Aceptando estas selecciones, el *Visor* ofrece los resultados que muestran las tablas 22.7 a 22.9 y las figuras 22.9 y 22.10.

La tabla 22.7 muestra un *resumen de los casos procesados*. Los 15 casos con valor perdido en alguna de las variables no intervienen en el cálculo de las distancias. En una nota a pie de tabla se indica el nombre de la medida utilizada para obtener la matriz de distancias: la *distancia euclídea al cuadrado* (esta es la medida de distancia que el programa utiliza por defecto).

Tabla 22.7. Resumen de los casos procesados.

Casos[a]					
Valid		Perdidos		Total	
N	Porcentaje	N	Porcentaje	N	Porcentaje
391	96.3%	15	3.7%	406	100.0%

a. Distancia euclídea al cuadrado

La tabla 22.8 muestra la matriz de las distancias entre los elementos procesados. Puesto que hemos decidido conglomerar *variables*, la matriz de distancias muestra las distancias entre las variables. (Aunque la diagonal está blanqueada, en realidad todas las casillas contienen ceros, es decir, el valor de la distancia euclídea mínima posible).

Tabla 22.8. Matriz de distancias.

	Archivo matricial de entrada						
Caso	Consumo (l/100Km)	Cilindrada en cc	Potencia (CV)	Peso total (kg)	Aceleración 0 a 100	Año del modelo	Número de cilindros
Consumo (l/100Km)		5071082496	3860925	405358240	20133	1657506	15669
Cilindrada en cc	5071082496		4803001344	2701554176	5066977280	4918516224	5087751680
Potencia (CV)	3860925	4803001344		331556096	3709066	933811	4344718
Peso total (kg)	405358240	2701554176	331556096		403080160	358392640	410201696
Aceleración 0 a 100 km/h	20133	5066977280	3709066	403080160		1435214	45607
Año del modelo	1657506	4918516224	933811	358392640	1435214		1953059
Número de cilindros	15669	5087751680	4344718	410201696	45607	1953059	

La tabla 22.9 muestra el historial de conglomeración. Puede observarse que las distancias de fusión (columna *Coeficientes*) aumentan rápidamente conforme avanzan las etapas. El titular que precede al historial de conglomeración en el *Visor* indica que se ha utilizado el método de conglomeración *vinculación promedio inter-grupos*.

Tabla 22.9. 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	7	15669.000	0	0	2
2	1	5	32869.949	1	0	4
3	3	6	933811.000	0	0	4
4	1	3	2826748.000	2	3	5
5	1	4	381717760.000	4	0	6
6	1	2	4608146944.000	5	0	0

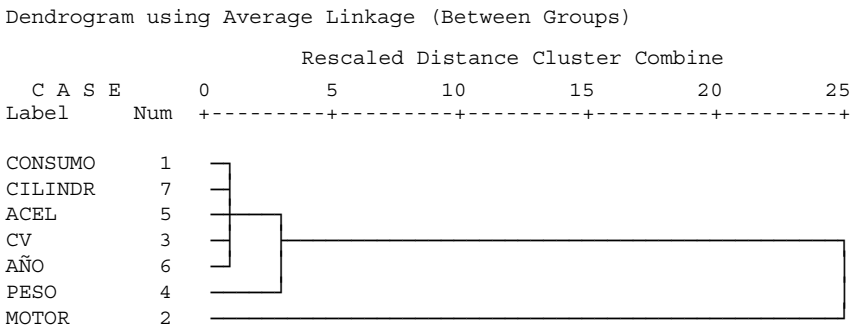
La figura 22.9 ofrece el *diagrama de témpanos*. El diagrama hace evidente que el proceso de fusión de elementos se realiza paso a paso, fundiendo únicamente dos elementos en cada paso.

Figura 22.9. Diagrama de témpanos.

Número de conglomerados	Cilindrada		Peso		Año del modelo		Potencia		Aceleración		Nº de cilindros		Consumo
1	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
3	X		X		X	X	X	X	X	X	X	X	X
4	X		X		X	X	X		X	X	X	X	X
5	X		X		X		X		X	X	X	X	X
6	X		X		X		X		X		X	X	X

La figura 22.10 muestra el *dendrograma*. Según veremos más adelante, este dendrograma resulta bastante engañoso. Por ejemplo, la variable *motor* parece ser la única que se diferencia del resto, pues se une a ellas sólo en la etapa final y a distancia muy superior al resto de las distancias de fusión. Además, la gran distancia a la que esta variable se incorpora al conglomerado final está enmascarando las distancias de fusión del resto de variables: sabemos, por el diagrama de témpanos y por el historial de conglomeración, que las variables *consumo*, *aceleración* y *número de cilindros* se funden muy pronto, y que las variables potencia y año del modelo podrían formar un conglomerado separado. Sin embargo, la solución no parece demasiado clara.

Figura 22.10. Dendrograma.



El motivo por el que la solución ha resultado tan poco ilustrativa de las relaciones existentes entre las variables hay que buscarlo en la medida de distancia utilizada. La distancia euclídea al cuadrado es una medida muy afectada por las diferencias de métrica existentes entre las variables.

La tabla 22.10 muestra algunos estadísticos descriptivos de las variables utilizadas en el análisis (esta tabla se ha obtenido con el procedimiento **Estadísticos descriptivos** > **Descriptivos** del menú **Analizar**). Puede apreciarse con claridad que la métrica de las variables *motor* (cilindrada en cc) y *peso* es muy diferente de la métrica del resto de variables. El rango de la cilindrada es, por ejemplo, 358 veces mayor que el del consumo, y 1478 veces mayor que el del número de cilindros.

Tabla 22.10. Estadísticos descriptivos.

	N	Rango	Mínimo	Máximo	Media	Desv. típ.
Consumo (l/100Km)	398	21	5	26	11.23	3.95
Cilindrada en cc	406	7390	66	7456	3179.73	1724.01
Potencia (CV)	400	184	46	230	104.83	38.52
Peso total (kg)	406	1469	244	1713	989.51	283.28
Aceleración 0 a 100 km/h	406	17	8	25	15.50	2.82
Año del modelo	406	82	0	82	75.75	5.31
Número de cilindros	405	5	3	8	5.47	1.71
N válido (según lista)	391					

Si la métrica original de las variables fuera relevante para el análisis (como, por ejemplo, cuando las puntuaciones representan un coste económico: saldos medios de distintos tipos de cuentas bancarias, etc.), sería conveniente y apropiado realizar el análisis utilizando la métrica original de las variables.

Sin embargo, si el interés del estudio se centra en las relaciones existentes entre las variables, es necesario homogeneizar la métrica de las variables antes de iniciar el análisis para que las diferencias de métrica no constituyan un problema.

Vamos a repetir el mismo análisis, pero transformando los valores originales en puntuaciones típicas. Para ello:

- Recuperar el cuadro de diálogo *Análisis de conglomerados jerárquicos* (ver figura 22.1), trasladar las variables *consumo*, *motor* (cilindrada), *cv* (potencia), *peso*, *acel* (aceleración) y *cilindr* (nº de cilindros) a la lista **Variables**, y seleccionar la opción **Variables** del apartado **Conglomerar**.

- Pulsar en el botón **Estadísticos...** para acceder al subcuadro de diálogo *Análisis de conglomerados jerárquico: Estadísticos* (ver figura 22.4) y marcar la opción **Matriz de distancias**. Pulsar el botón **Continuar**.

- Pulsar en el botón **Gráficos...** para acceder al subcuadro de diálogo *Análisis de conglomerados jerárquico: Gráficos* (ver figura 22.5) y marcar la opción **Dendrograma**. Pulsar el botón **Continuar**.

- Pulsar en el botón **Método...** para acceder al subcuadro de diálogo *Análisis de conglomerados jerárquicos: Método* (ver figura 22.8) y en el apartado **Transformar valores**, seleccionar la opción **Puntuaciones *Z*** de la lista **Estandarizar**. Pulsar el botón **Continuar**.

Aceptando estas selecciones, el *Visor* ofrece, entre otros, los resultados que muestran la tabla 22.11 y las figuras 22.11 y 22.12.

Comparando la nueva matriz de distancias (tabla 22.11) con la matriz de distancias anterior (tabla 22.8) puede comprobarse que los valores de la matriz son ahora mucho más homogéneos. Ya no es la variable *motor* (potencia) la que posee las distancias más grandes con el resto de elementos, sino la variable *acel* (aceleración). Es evidente que algo ha cambiado al tipificar los valores originales.

Tabla 22.11. Matriz de distancias.

	Archivo matricial de entrada						
Caso	Consumo (l/100Km)	Cilindrada en cc	Potencia (CV)	Peso total (kg)	Aceleración 0 a 100 km/h	Año del modelo	Número de cilindros
Consumo (l/100Km)		99.38	113.66	89.62	1152.97	1210.38	123.16
Cilindrada en cc	99.38		79.21	51.57	1207.64	1066.48	38.33
Potencia (CV)	113.66	79.21		106.94	1327.03	1100.55	122.98
Peso total (kg)	89.62	51.57	106.94		1111.85	1016.44	80.24
Aceleración 0 a 100 km/h	1152.97	1207.64	1327.03	1111.85		548.95	1178.54
Año del modelo	1210.38	1066.48	1100.55	1016.44	548.95		1046.44
Número de cilindros	123.16	38.33	122.98	80.24	1178.54	1046.44	

La figura 22.11 muestra el *diagrama de témpanos* del nuevo análisis. Las dos variables que primero se funden son el número de cilindros y la cilindrada, que a su vez se funden con el peso en la segunda etapa. Esta solución es claramente distinta de la obtenida anteriormente, en la que la variable cilindrada (debido, sin duda, a su métrica) se mantenía muy distante de las demás variables, a las que sólo se unía en el conglomerado final.

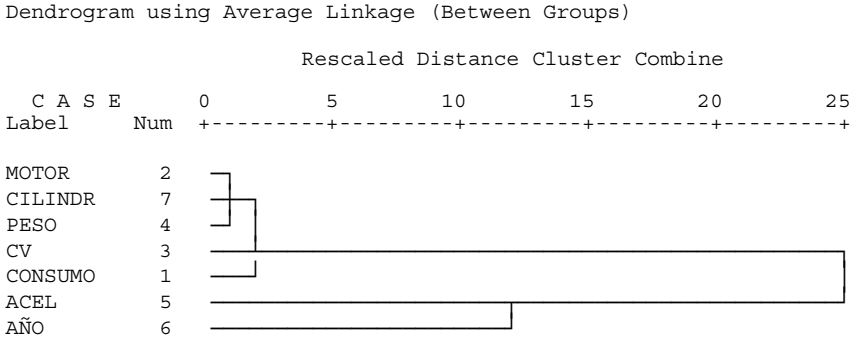
Figura 22.11. Diagrama de témpanos.

Número de conglomerados	Año del modelo		Aceleración		Potencia		Peso		Nº de cilindros		Cilindrada		Consumo)
1	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
3	X		X		X	X	X	X	X	X	X	X	X
4	X		X		X	X	X	X	X	X	X		X
5	X		X		X		X	X	X	X	X		X
6	X		X		X		X		X	X	X		X

El dendrograma del nuevo conjunto de soluciones (figura 22.12) muestra un conglomerado de tres variables muy próximas, la cilindrada (motor), el número de cilindros y el peso. A este conglomerado se une la potencia (*cv*) y posteriormente el consumo (para apreciar el orden en que se funden estas últimas es necesario consultar el diagrama de témpanos). Por otro lado, la aceleración se une al año del modelo con una distancia de fusión mucho mayor que las distancias de fusión de las primeras etapas. Por fin, este último conglomerado se funde con el resto de variables en la última etapa.

El dendrograma sugiere que la mejor solución es la de tres conglomerados. Uno formado por las variables relativas al tamaño del vehículo: cilindrada (*motor*), número de cilindros (*cilindr*), peso, potencia (*cv*) y consumo; otro formado por la variable aceleración (*acel*); y otro más formado por la variable año del modelo.

Figura 22.12. Dendrograma.



Si dudamos de cuál es el número óptimo de conglomerados, podemos utilizar el procedimiento *Análisis factorial* para validar la solución obtenida. Esto es posible hacerlo siempre y cuando pueda utilizarse el coeficiente de correlación de Pearson para cuantificar las distancias entre las variables. Si las variables analizadas son categóricas u ordinales no es recomendable utilizar el análisis factorial como método de validación.

La tabla 22.12 muestra el resultado del análisis factorial con la solución rotada mediante el método *Varimax*. Puede comprobarse que la solución obtenida es idéntica a la alcanzada con el análisis de conglomerados jerárquico. En el primer factor saturan las variables relativas al tamaño del vehículo (peso, cilindrada, número de cilindros, consumo y potencia); en el segundo factor satura, fundamentalmente, la variable aceleración; y en el tercer factor, la variable año del modelo.

Esto es exactamente lo que hemos encontrado con el análisis de conglomerados jerárquico. Sin embargo, la solución del análisis factorial es más flexible en un sentido: permite que las variables saturen en varios factores de manera simultánea. Así, podemos apreciar, por ejemplo, que la aceleración se encuentra positivamente relacionada con la potencia, o que los coches más antiguos consumen más. Esto no es posible apreciarlo en el análisis de conglomerados porque, una vez que se ha formado un conglomerado en una etapa, ya es indivisible en las etapas posteriores.

Tabla 22.12. Matriz factorial de los componentes rotados.

	Componente		
	1	2	3
Peso total (kg)	.962	.146	-.103
Cilindrada en cc	.925	.287	-.150
Número de cilindros	.916	.247	-.127
Consumo (l/100Km)	.850	.192	-.401
Potencia (CV)	.798	.501	-.206
Aceleración 0 a 100 km/h	-.279	-.947	.128
Año del modelo	-.190	-.126	.970

Método de extracción: Análisis de componentes principales.
Método de rotación: Normalización Varimax con Kaiser.

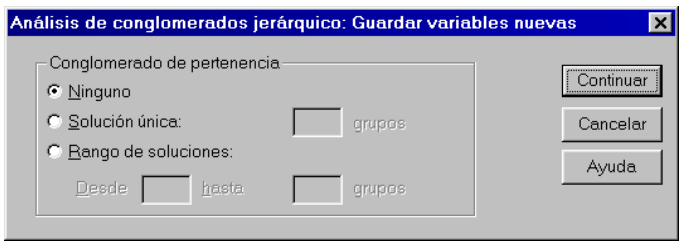
Cuando la estructura de los elementos analizados es clara, cualquiera de los métodos de conglomeración dará lugar a datos similares. El dendrograma resultante utilizando, por ejemplo, el método de la mediana o el del vecino más próximo es muy similar al dendrograma de la figura 22.12.

Guardar

Las opciones del cuadro de diálogo *Guardar* permiten crear en el *Editor de datos* variables nuevas basadas en los resultados del análisis. Para crear estas variables:

- Pulsar en el botón **Guardar...** del cuadro de diálogo *Análisis de conglomerado jerárquico* (ver figura 22.1) para acceder al subcuadro de diálogo *Análisis de conglomerados: Guardar variables nuevas* que muestra la figura 22.13.

Figura 22.13. Cuadro de diálogo *Análisis de conglomerados jerárquicos: Guardar variables nuevas*.



Conglomerado de pertenencia. Las opciones de este apartado permiten crear y guardar una o más variables con valores indicando el conglomerado al que ha sido asignado cada caso. Estas variables pueden emplearse posteriormente en otros análisis para, por ejemplo, explorar diferencias entre los grupos.

Estas opciones sólo están disponibles si se ha seleccionado la opción **Casos** del apartado **Conglomerar** en el cuadro de diálogo principal (ver figura 22.1):

- **Ninguno**. No crea ninguna variable. Es la opción por defecto.
- **Solución única**. Guarda una única variable cuyos valores indican el conglomerado al que ha sido asignado cada caso en la solución de *k* conglomerados. El cuadro de texto ***k* grupos** permite introducir el número de conglomerados de la solución que se desea obtener.

- **Rango de soluciones**. Guarda un conjunto de variables cuyos valores indican el conglomerado al que ha sido asignado cada caso en las distintas soluciones del rango seleccionado.

  Los cuadros de texto **Desde *k* hasta *p* grupos** permiten definir el rango de soluciones que se desea obtener. Para ello, hay que introducir para *k* un número entero que indique el número de conglomerados de la solución con menos conglomerados; y para *p*, un número entero que indique el número de conglomerados de la solución con más conglomerados.

Ejemplo (Análisis de conglomerados jerárquico > Guardar)

Este ejemplo muestra cómo guardar algunas variables basadas en los resultados del análisis de conglomerados jerárquico y cómo utilizar esas variables con otras técnicas de análisis estadístico. Para facilitar la representación de los resultados utilizaremos una muestra aleatoria de 40 casos. Para seleccionar los casos:

- En la ventana del *Editor de datos*, seleccionar la opción **Seleccionar casos** del menú **Datos** para acceder al cuadro de diálogo *Seleccionar casos*.
- Marcar la opción **Muestra aleatoria de casos** del apartado **Seleccionar** y pulsar el botón **Muestra...** para acceder al subcuadro de diálogo *Seleccionar casos: Muestra aleatoria*.
- En el apartado **Tamaño de la muestra**, seleccionar la expresión **Exactamente *n* casos de los primeros *N* casos** e introducir los valores $n = 40$ y $N = 406$ en los correspondientes cuadros de texto. Pulsar en el botón Continuar para volver al cuadro de diálogo principal.

Aceptando estas selecciones, el archivo de datos queda filtrado, dejando disponibles sólo 40 de los 406 casos existentes.

Para llevar a cabo el análisis de conglomerados jerárquico y guardar los resultados del análisis como nuevas variables del archivo de datos:

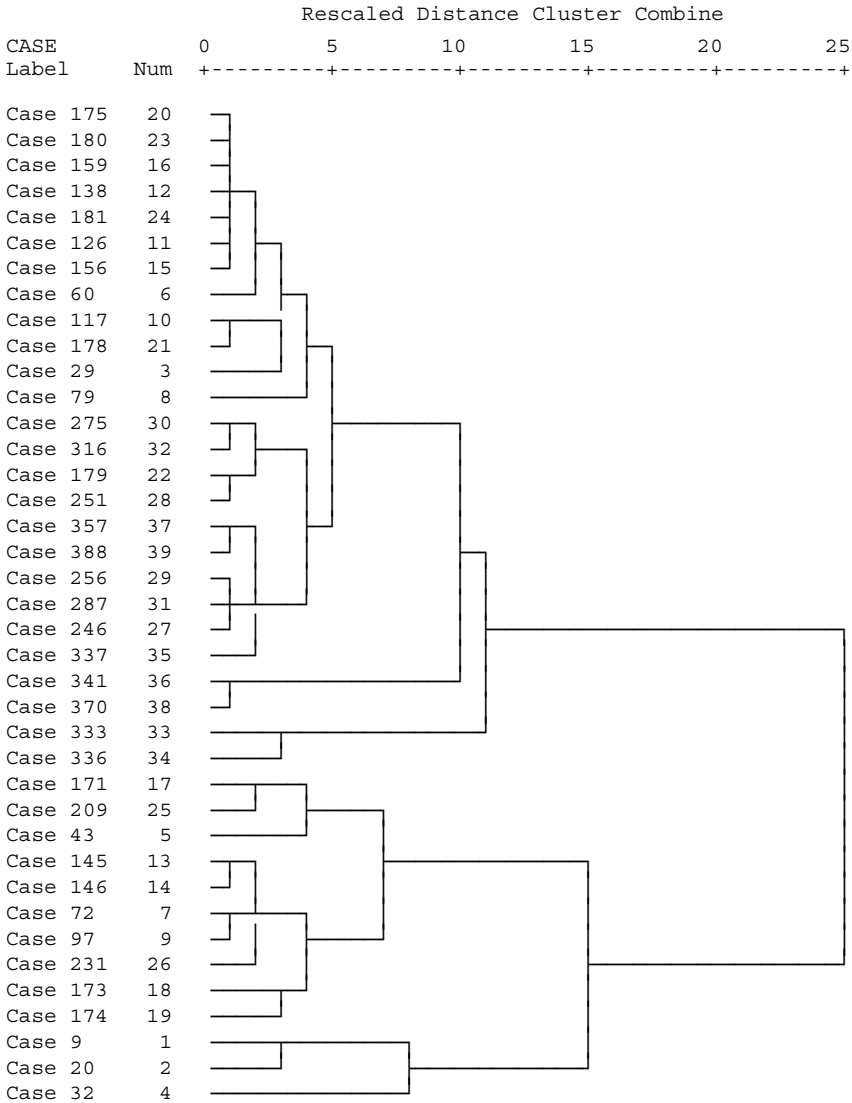
- En el cuadro de diálogo *Análisis de conglomerados jerárquicos* (ver figura 22.1) seleccionar las variables *consumo*, *motor*, *cv*, *peso*, *acel*, *año* y *cilindr* y trasladarlas a la lista **Variables**.
- Marcar la opción **Casos** del apartado **Conglomerar**.
- Desactivar la opción **Estadísticos** del apartado **Mostrar** para que el *Visor* sólo muestre los gráficos.
- Pulsar en el botón **Gráficos...** para acceder al subcuadro de diálogo *Análisis de conglomerados jerárquico: Gráficos* (ver figura 22.4).
- Marcar la opción **Dendrograma** y la opción **Ninguno** del apartado **Témpanos**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- Pulsar el botón **Método...** para acceder al subcuadro de diálogo *Análisis de conglomerados jerárquico: Método* (ver figura 22.8).
- En el apartado **Transformar valores**, seleccionar la opción **Puntuaciones *Z*** de la lista desplegable **Estandarizar**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- Pulsar en el botón **Guardar...** para acceder al subcuadro de diálogo *Análisis de conglomerados jerárquico: Guardar variables nuevas* (ver figura 22.13).
- En el apartado **Conglomerado de pertenencia**, seleccionar la opción **Solución única: __ grupos** e introducir el valor 3 en el cuadro de texto.

Aceptando estas selecciones, el *Visor* ofrece el dendrograma que muestra la figura 22.14 y crea en el *Editor de datos* una nueva variable a la que asigna el nombre *clu3_1*.

El dendrograma de la figura 22.14 ayuda a tomar la decisión sobre el número óptimo de conglomerados. La figura permite apreciar la existencia de dos conglomerados claros. Pero también podría ser apropiada una solución de tres conglomerados, manteniendo los casos 9, 30 y 32 en un conglomerado separado.

En el archivo de datos se ha almacenado una variable que contiene el conglomerado al que pertenece cada caso, para la solución de 3 conglomerados. Esta nueva variable, denominada *clu3_1*, puede ser utilizada con otros procedimientos de análisis. El nombre asignado a la nueva variable, *clu3_1*, tiene este significado: *clu* = *cluster* (conglomerado en inglés); *3* = clasificación correspondiente a la solución de 3 conglomerados; *_1* = primera variable creada con el nombre *clu3* en la sesión actual).

Figura 22.14. Dendrograma para el método de Vinculación inter-grupos.



Las tablas 22.13 y 22.14 muestran las medias de cada una de las variables utilizadas en cada uno de los conglomerados del análisis. La tabla 22.13 se basa en las puntuaciones originales. La tabla 22.14, en las puntuaciones tipificadas. Estas tablas se han obtenido con el procedimiento **Comparar medias > Medias** del menú **Analizar**, tomando como *dependientes* las siete variables incluidas en el análisis y como *independiente* la variable *clu3_1*. (La tabla basada en las puntuaciones originales es similar a la ofrecida como opción en el procedimiento *Análisis de conglomerados de K medias*.

En las columnas se encuentran los vectores de medias de los conglomerados (vectores a los que ya nos hemos referido como ***centroides***). Los *centroides* son de utilidad para conocer cuáles son las variables en las que más se diferencian los conglomerados; por tanto, ayudan a conocer la constitución interna de los conglomerados y a interpretar la solución. Esta información puede resultar especialmente interesante debido a que el análisis de conglomerados no utiliza selectivamente las variables con mejores propiedades estadísticas, sino que utiliza todas las variables designadas por el usuario.

Tabla 22.13. Centroides de los conglomerados en puntuaciones directas.

		Conglomerados			
		1	2	3	Total
Consumo (l/100Km)	Media	19.33	8.46	15.20	11.03
	N	3	26	10	39
Cilindrada en cc	Media	6937.00	1839.77	4770.20	2983.26
	N	3	26	10	39
Potencia (CV)	Media	221.67	81.85	128.90	104.67
	N	3	26	10	39
Peso total (kg)	Media	1347.00	784.35	1278.20	954.26
	N	3	26	10	39
Aceleración 0 a 100 km/h	Media	11.33	15.88	14.39	15.15
	N	3	26	10	39
Año del modelo	Media	70.00	76.42	74.20	75.36
	N	3	26	10	39
Número de cilindros	Media	8.00	4.08	7.40	5.23
	N	3	26	10	39

Los *centroides* indican (tanto los originales como los tipificados), por ejemplo, que el conglomerado 1 está más cerca del 3 que del 2 en todas las variables exceptuando la aceleración y el año del modelo (de hecho, los conglomerados 1 y 3 son los que se funden en la solución de 2 conglomerados). El conglomerado 1 está constituido por los 3 vehículos de mayor *tamaño del motor* (es decir, los tres vehículos con mayores consumo, cilindrada, potencia, peso y número de cilindros), aceleración rápida y mucha antigüedad . El conglomerado 2 está formado por los

26 vehículos con menor *tamaño del motor*, aceleración lenta y poca antigüedad. El conglomerado 3 consta de 10 vehículos con gran *tamaño del motor* (aunque no tanto como los del conglomerado 1), pero con aceleración lenta y poca antigüedad.

Figura 22.14. Centroides de los conglomerados en puntuaciones típicas.

		Conglomerados			
Puntuaciones típicas		1	2	3	Total
Consumo (l/100Km)	Media	1.934	-.597	.972	.000
	N	3	26	10	39
Cilindrada en cc	Media	2.119	-.661	.937	-.038
	N	3	26	10	39
Potencia (CV)	Media	2.589	-.545	.509	-.034
	N	3	26	10	39
Peso total (kg)	Media	1.332	-.629	1.092	-.037
	N	3	26	10	39
Aceleración 0 a 100 km/h	Media	-1.368	.302	-.245	.034
	N	3	26	10	39
Año del modelo	Media	-1.510	.346	-.296	.039
	N	3	26	10	39
Número de cilindros	Media	1.477	-.669	1.148	-.038
	N	3	26	10	39

Otra cosa interesante que puede hacerse con la variable del conglomerado de pertenencia es comprobar si existen diferencias significativas entre los conglomerados obtenidos. Para ello, podemos utilizar el ***análisis de varianza*** de un factor tomando como variable *independiente* o *factor* la variable que contiene información sobre el conglomerado al que pertenece cada sujeto (*clus3_1*) y como variables *dependientes* cada una de las variables incluidas en el análisis. (Podemos obtener un ANOVA de un factor con el mismo procedimiento utilizado para obtener la tabla de medias (**Comparar medias** > **Medias**), o con el procedimiento **Comparar medias** > **ANOVA de un factor**).

La tabla 22.15 muestra un resumen con los resultados del análisis de varianza. Esta tabla es equivalente a la tabla resumen del ANOVA que ofrece el procedimiento *Análisis de conglomerados de K medias*. Los niveles críticos (*Sig.*) asociados a cada estadístico de contraste son orientativos, pues están calculados sin efectuar ningún tipo de control sobre la tasa de error (es decir, sobre la probabilidad de cometer errores tipo I). Por otro lado, los sujetos no han sido asignados aleatoriamente a los distintos conglomerados, como asume el análisis de varianza. No obstante, los resultados de la tabla ayudan a valorar si los conglomerados son diferentes entre sí y qué variables contribuyen a hacerlos diferentes.

Tabla 22.15. Tabal resumen del ANOVA.

		Suma de cuadrados	gl	Media cuadrática	F	Sig.
Consumo (l/100Km)	Inter-grupos	552.246	2	276.123	66.836	.000
	Intra-grupos	148.728	36	4.131		
	Total	700.974	38			
Cilindrada en cc	Inter-grupos	112824575.221	2	56412287.610	187.318	.000
	Intra-grupos	10841696.215	36	301158.228		
	Total	123666271.436	38			
Potencia (CV)	Inter-grupos	60479.715	2	30239.858	80.230	.000
	Intra-grupos	13568.951	36	376.915		
	Total	74048.667	38			
Peso total (kg)	Inter-grupos	2262743.951	2	1131371.976	52.811	.000
	Intra-grupos	771223.485	36	21422.875		
	Total	3033967.436	38			
Aceleración 0 a 100 km/h	Inter-grupos	63.361	2	31.681	5.362	.009
	Intra-grupos	212.696	36	5.908		
	Total	276.057	38			
Año del modelo	Inter-grupos	129.028	2	64.514	7.493	.002
	Intra-grupos	309.946	36	8.610		
	Total	438.974	38			
Número de cilindros	Inter-grupos	104.677	2	52.338	103.265	.000
	Intra-grupos	18.246	36	.507		
	Total	122.923	38			

En nuestro ejemplo, todos los estadísticos *F* de la tabla tienen asociados niveles críticos muy pequeños, por lo que podemos pensar que todas las variables incluidas en el análisis son útiles desde el punto de vista de la clasificación de los casos. Si alguna de las variables tuviera asociado un nivel crítico alto (por encima de 0,05) deberíamos sospechar que esa variable carece de relevancia para efectuar la conglomeración.

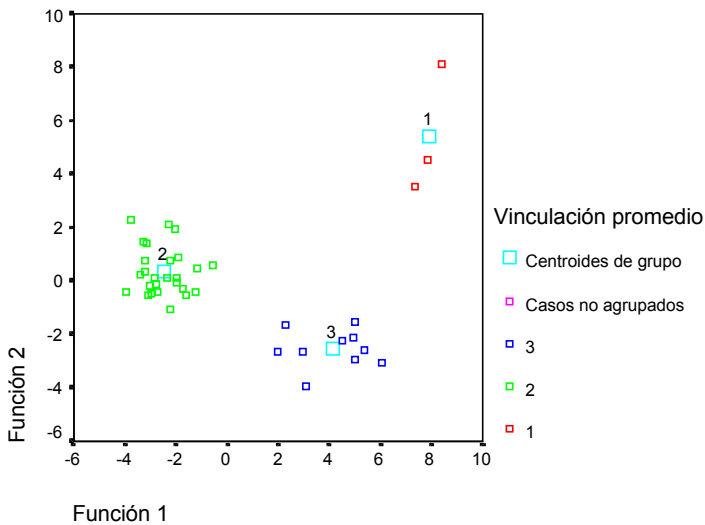
Cuando la solución del análisis de conglomerados contiene más de dos conglomerados, pueden realizarse comparaciones múltiples entre las medias de los conglomerados (comparaciones *post-hoc*) para averiguar qué conglomerados difieren entre sí en cada una de las variables.

Otro análisis interesante que podemos hacer con la variable del conglomerado de pertenencia es un ***análisis discriminante***. Utilizando como variable de clasificación el conglomerado de pertenencia podemos facilitar la interpretación de las diferencias existentes entre los conglomerados, ya que las variables independientes se encuentran muy relacionadas entre sí.

La figura 22.15 muestra la distribución de los casos en el espacio definido por las dos funciones discriminantes. Los casos del conglomerado 1 obtienen puntuaciones altas en las dos funciones. Los casos del segundo conglomerado obtienen puntuaciones bajas en la primera función y puntuaciones medias en la segunda. Los casos del tercer conglomerado obtienen puntuaciones medias-altas en la primera función y puntuaciones medias-bajas en la segunda. El diagrama de dispersión muestra con claridad que los casos de un conglomerado se encuentran bien diferenciados de los casos de los restantes conglomerados.

El diagrama también muestra la ubicación concreta de los tres centroides en las funciones discriminantes.

Figura 22.15. Diagrama de dispersión de las funciones discriminantes.



La tabla 22.16 muestra la matriz de los *coeficientes estandarizados* de las dos funciones discriminantes obtenidas. La primera función atribuye mayor importancia a la cilindrada y al consumo; la segunda atribuye mayor importancia a la potencia, a la aceleración y al peso.

Podemos pensar que los casos del primer conglomerado son coches de gran consumo y cilindrada, poco peso y gran potencia (puntuaciones altas en ambas funciones), digamos que son coches potentes. El segundo conglomerado está constituido por coches de reducida cilindrada y poco consumo (función 1) y potencia y aceleración medias (función 2). El tercer conglomerado esta compuesto por coches de cilindrada y consumo medio-alto (función 1), poca potencia y mucho peso (función 2).

Tabla 22.16. Coeficientes estandarizados del análisis discriminante.

	Función	
	1	2
Consumo (l/100Km)	.499	-.170
Cilindrada en cc	.908	.192
Potencia (CV)	-.192	2.173
Peso total (kg)	-.181	-1.191
Aceleración 0 a 100 km/h	.209	1.091
Año del modelo	-.142	.175
Número de cilindros	.256	-.814

La tabla 22.17 recoge las correlaciones entre las variables de agrupación y las funciones discriminantes. La primera función contiene información sobre el tamaño del motor y la segunda función información sobre la potencia del motor.

Tabla 22.17. Matriz de estructura del análisis discriminante.

	Función	
	1	2
Cilindrada en cc	.850	.095
Número de cilindros	.624	-.189
Potencia (CV)	.522	.355
Consumo (l/100Km)	.509	.015
Peso total (kg)	.445	-.151
Año del modelo	-.160	-.104
Aceleración 0 a 100 km/h	-.134	-.095

< Fin del capítulo 22 >