Capítulo 21
Análisis de conglomerados (I): El procedimiento *Conglomerados de K medias*

Introducción

El análisis de conglomerados (en inglés, *cluster analisys*) es una técnica multivariante que permite agrupar los *casos* o *variables* de un archivo de datos en función del parecido o similaridad existente entre ellos.

Como técnica de *agrupación de variables*, el análisis de conglomerados es similar al análisis factorial; pero, mientras que la *factorización* es más bien poco flexible en algunos de sus supuestos (linealidad, normalidad, variables cuantitativas, etc.) y siempre estima de la misma manera la matriz de distancias, la *aglomeración* es menos restrictiva en sus supuestos (no exige linealidad, ni simetría, permite variables categóricas, etc.) y admite varios métodos de estimación de la matriz de distancias.

Como técnica de *agrupación de casos*, el análisis de conglomerados es similar al análisis discriminante. Sin embargo, mientras que el análisis discriminante efectúa la clasificación tomando como referencia un criterio o variable dependiente (los grupos de clasificación), el análisis de conglomerados permite detectar el número óptimo de grupos y su composición únicamente a partir de la similaridad existente entre los casos; además, el análisis de conglomerados no asume ninguna distribución específica para las variables. Por simplicidad, en este capítulo se comienza exponiendo la agrupación de casos.

El programa SPSS dispone de dos tipos de análisis de conglomerados: el análisis de conglomerados *jerárquico* y el análisis de conglomerados de *K medias*. El método *jerárquico* es idóneo para determinar el número óptimo de conglomerados existente en los datos y el contenido de los mismos. El método de *K medias* permite procesar un número ilimitado de casos, pero sólo permite utilizar un método de aglomeración y requiere que se proponga previamente el número de conglomerados que se desea obtener.

Ambos métodos de análisis son de tipo *aglomerativo*, en el sentido de que, partiendo del análisis de los casos individuales, intentan ir agrupando casos hasta llegar a la formación de grupos o conglomerados homogéneos. Pero existen también métodos *divisivos* que parten de la muestra global como un sólo grupo y la van dividiendo en subgrupos hasta llegar a la formación de grupos o conglomerados homogéneos con un número relativamente reducido de sujetos. Las técnicas divisivas son especialmente adecuadas para el análisis de variables categóricas. Uno de los métodos divisivos más difundido es el CHAID (disponible como un procedimiento autónomo dentro del módulo *AnswerTree* de SPSS).

Análisis de conglomerados de *K medias*

El análisis de conglomerados de *K medias* es un método de *agrupación de casos* que se basa en las distancias existentes entre ellos en un conjunto de variables (este método de aglomeración no permite agrupar variables). Versiones anteriores del procedimiento comenzaban el análisis con la asignación de los *K* primeros casos a los *centros* de los *K* conglomerados (los *centros* multivariantes de los conglomerados se denominan *centroides*). En la versión actual se comienza seleccionado los *K* casos más distantes entre sí (el usuario debe determinar inicialmente el número *K* de conglomerados que desea obtener). Y a continuación se inicia la lectura secuencial del archivo de datos asignando cada caso al *centro* más próximo y actualizando el valor de los *centros* a medida que se van incorporando nuevos casos. Una vez que todos los casos han sido asignados a uno de los *K* conglomerados, se inicia un proceso iterativo para calcular los *centroides* finales de esos *K* conglomerados.

El análisis de conglomerados de *K* medias es especialmente útil cuando se dispone de un gran número de casos. Existe la posibilidad de utilizar la técnica de manera exploratoria, clasificando los casos e iterando para encontrar la ubicación de los *centroides*, o sólo como técnica de clasificación, clasificando los casos a partir de *centroides* conocidos suministrados por el usuario. Cuando se utiliza como técnica exploratoria, es habitual que el usuario desconozca el número idóneo de conglomerados, por lo que es conveniente repetir el análisis con distinto número de conglomerados y comparar las soluciones obtenidas; en estos casos también puede utilizarse el método análisis de conglomerados *jerárquico* con una submuestra de casos.

Para llevar a cabo un Análisis de conglomerados de *K* medias:

- Seleccionar la opción **Clasificar** > **Conglomerados de *K* medias** del menú **Analizar** para acceder al cuadro de diálogo *Análisis de conglomerados de K-medias* que muestra la figura 21.1.

Figura 21.1. Cuadro de diálogo Análisis de conglomerados de K-medias.

La lista de variables del archivo de datos ofrece un listado con todas las variables del archivo (numéricas y de cadena), pero las variables de cadena sólo pueden utilizarse para etiquetar casos. Para obtener un análisis de conglomerados de *K medias*:

- Seleccionar las variables numéricas que se desea a utilizar para diferenciar a los sujetos 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 a la lista **Etiquetar casos mediante**.

Nº de conglomerados. En este cuadro de texto se encuentra seleccionada por defecto la solución de dos conglomerados. Para solicitar un número mayor de conglomerados, introducir el número deseado en el cuadro.

Método. Las opciones de este apartado permiten indicar si los *centros* de los conglomerados deben o no ser estimados iterativamente:

- **Iterar y clasificar**. El procedimiento se encarga de estimar los *centros* iterativamente y de clasificar a las sujetos con arreglo a los *centros* estimados.
- **Sólo clasificar**. Se clasifica a los sujetos según los *centros* iniciales (sin actualizar sus valores iterativamente). Al marcar esta opción se desactiva el botón **Iterar...**, impidiendo esto el acceso a las especificaciones del proceso de iteración. Esta opción suele utilizarse junto con el botón **Centros>>** (ver más adelante).

Ejemplo (Análisis de conglomerados de K medias)

Este ejemplo muestra cómo obtener un análisis de conglomerados con las especificaciones que el procedimiento tiene establecidas por defecto. En todos los ejemplos de este capítulo utilizaremos el archivo de datos *Coches.sav*, que se encuentra en la misma carpeta en la que se ha instalado el SPSS.

Para hacer más comprensible la representación gráfica de los resultados, vamos a comenzar utilizando únicamente el 20 % de los casos de la muestra. Para seleccionar el 20 % de 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*.
- Seleccionar 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**, introducir el valor 20 en recuadro de texto de la opción **Aproximadamente *p* % de todos los casos**. Pulsar el botón **Continuar**.

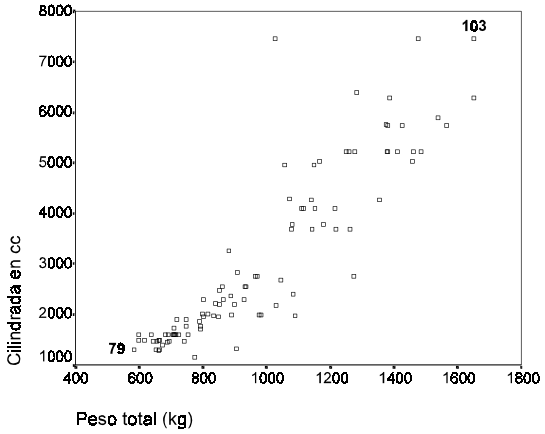
Aceptando estas selecciones, el archivo de datos queda filtrado, dejando disponibles sólo 98 de los 406 casos existentes.

Vamos a comenzar representando la distancia existente entre los casos en dos variables de interés. Para ello:

- Seleccionar la opción **Interactivos** > **Diagrama de dispersión** del menú **Gráficos**.
- Desplazar la variable *peso* (peso total del vehículo en kg) al eje abscisas y la variable *motor* (cilindrada en cc) al eje de ordenadas.

Aceptando estas elecciones, el *Visor de resultados* genera el diagrama de dispersión que muestra la figura 21.2.

Figura 21.2. Diagrama de dispersión del peso por la cilindrada.



En el diagrama de dispersión están representados los valores en *peso* y *motor* de los 98 casos seleccionados. Se puede apreciar que existe un grupo de vehículos relativamente numeroso con peso y cilindrada reducidos y otro grupo más disperso de vehículos de mayor peso y mayor cilindrada. Se han identificado, mediante el número de caso, los dos vehículos aparentemente más alejados entre sí (el caso 79 y el 103). La nube de puntos, por tanto, incita a pensar que existen al menos dos grupos naturales de casos.

Para clasificar los casos en dos grupos:

- Seleccionar la opción **Clasificar** > **Conglomerado de *K* medias** del menú **Analizar** para acceder al cuadro de diálogo *Análisis de conglomerados de K-medias* que muestra la figura 21.1.
- Trasladar las variables *motor* y *peso* a la lista **Variables**.

Aceptando estas selecciones, el *Visor* ofrece los resultados que muestran las tablas 21.1 a la 21.4. La tabla 21.1 contiene los *centros iniciales*, es decir, los valores que corresponden, en las dos variables de clasificación utilizadas, a los dos casos que han sido elegidos como centros respectivos de los dos conglomerados solicitados. Inspeccionando el archivo de datos es fácil comprobar que esos dos casos son el 103 (*Conglomerado 1*) y el 79 (*Conglomerado 2*), los mismos que han sido identificados en el diagrama de dispersión.

Tabla 21.1. Centros iniciales de los conglomerados.

	Conglomerado	
	1	2
Cilindrada en cc	7456	1147
Peso total (kg)	1650	776

Una vez seleccionados los centros de los conglomerados, cada caso es asignado al conglomerado de cuyo centro se encuentra más próximo y comienza un proceso de ubicación iterativa de los *centros*. En la primera iteración se reasignan los casos por su distancia al nuevo *centro* y, tras la reasignación, se vuelve a actualizar el valor del *centro*. En la siguiente iteración se vuelven a reasignar los casos y a actualizar el valor del *centro*. Etc. La tabla 21.2 resume el *historial de iteraciones* (5 en nuestro ejemplo) con indicación del cambio (desplazamiento) experimentado por cada *centro* en cada iteración. Puede observarse que, conforme avanzan las iteraciones, el desplazamiento de los *centros* se va haciendo más y más pequeño, hasta llegar a la quinta iteración, en la que ya no existe desplazamiento alguno.

El proceso de iteración se detiene, por defecto, cuando se alcanzan 10 iteraciones o cuando de una iteración a otra no se produce ningún cambio en la ubicación de los *centroides* (cambio = 0). En nuestro ejemplo, el proceso ha finalizado antes de alcanzar 10 iteraciones porque en la quinta ya no se produce ningún cambio.

Tabla 21.2. Tabla del historial de iteraciones.

	Cambio en los centros de los conglomerados	
Iteración	1	2
1	1747.624	1115.589
2	366.905	199.335
3	99.619	52.974
4	176.349	110.532
5	.000	.000

La tabla 21.3 ofrece los *centros de los conglomerados finales*, es decir, los *centros* de los conglomerados tras el proceso de actualización iterativa. Comparando los *centros finales* (tras la iteración) de esta tabla con los *centros iniciales* (antes de la iteración) de la tabla 21.1 se puede apreciar con claridad un desplazamiento del *centro* del conglomerado 1 hacia la parte inferior del plano definido por las dos variables de clasificación y un desplazamiento del *centro* del conglomerado 2 hacia la parte superior.

Esta tabla es de gran utilidad para interpretar la constitución de los conglomerados pues resume los valores centrales de cada conglomerado en las variables de interés. La interpretación de los resultados de nuestro ejemplo es simple: el primer conglomerado está constituido por vehículos de gran cilindrada y mucho peso, mientras que segundo conglomerado está constituido por los vehículos de cilindrada reducida y poco peso.

Tabla 21.3. Centros de los conglomerados finales.

	Conglomerado	
	1	2
Cilindrada en cc	5093	1902
Peso total (kg)	1295	801

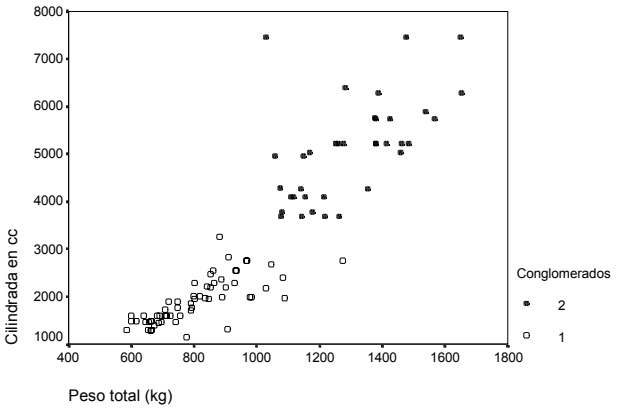
Por último, la tabla 21.4 informa sobre el *número de casos* asignado a cada conglomerado. En nuestro ejemplo, los tamaños de los conglomerados no son demasiado diferentes, pero dadas las características de la técnica es muy frecuente encontrarse con que alguno de los conglomerados finales contenga tan sólo un número muy reducido de casos atípicos.

Tabla 21.4. Número de casos en cada conglomerado final.

Conglomerado	1	36
	2	62
Válidos		98
Perdidos		0

Obteniendo ahora un diagrama de dispersión con marcas distintas para los casos de uno y otro conglomerado podemos formarnos una idea bastante precisa de las características de cada conglomerado. La figura 21.3 muestra este diagrama de dispersión (ver más abajo cómo guardar en el archivo de datos una variable cuyos valores indiquen a qué conglomerado pertenece cada caso).

Figura 21.3. Diagrama de dispersión con los dos conglomerados.



Los conglomerados obtenidos en nuestro ejemplo parecen captar de manera bastante adecuada la fisonomía de los datos: se distingue con claridad un conglomerado *inferior* (poco *peso* y poca *cilindrada*) y un conglomerado *superior* (mucho *peso* y mucha *cilindrada*). Sin embargo, no necesariamente siempre que se utiliza el análisis de conglomerados debemos esperar resultados tan claros (grupos tan claramente definidos). En la frontera entre ambos conglomerados pueden encontrarse casos que aparenten estar más cerca del conglomerado al que no han sido asignados. En nuestro ejemplo, un caso del conglomerado 1 parece alejarse de su centro hacia la derecha (el alejamiento se da particularmente en relación con la variable *peso*: es un vehículo que pesa más que el resto de vehículos de su grupo) y acercarse a los casos fronterizos del conglomerado 2.

Por otra parte, si decidiéramos obtener una solución de tres conglomerados, es bastante plausible pensar que el vehículo de la parte superior izquierda del conglomerado 2 podría constituir su propio conglomerado individual. Para entender lo que esta sucediendo es muy importante saber cómo se están midiendo las distancias entre los casos.

Medida de la distancia

El procedimiento *Análisis de conglomerados de K medias* siempre utiliza, para medir la distancia entre los casos, la *distancia euclídea*: la longitud de la recta que une ambos casos. La distancia euclídea se calcula de la siguiente manera:

$$d_{ii'} = \sqrt{\sum_j (X_{ij} - X_{i'j})^2}$$

donde X se refiere a las puntuaciones obtenidas por el caso i y el caso i' ($i \neq i'$) en cada una de las $j = 1, 2, ..., p$ variables incluidas en el análisis (el sumatorio de la expresión incluirá p términos, es decir, tantos como variables). Por ejemplo, la distancia euclídea entre el caso 103 y el caso 79 (ver tabla 21.1) es:

$$d_{103-79} = \sqrt{(7456-1147)^2 + (1650-776)^2} = 6369{,}25$$

Esta distancia es conceptualmente fácil de entender y sirve tanto para variables cuantitativas continuas como para variables ordinales. Pero, como contrapartida, es muy sensible a la métrica de las variables. En nuestra muestra, mientras que el *peso* toma valores comprendidos entre 585 y 1651 (rango = 1066), la cilindrada lo hace entre 1147 y 7456 (rango = 6309). Es decir, la cilindrada presenta una variabilidad más de diez veces mayor que el *peso*. Cabe esperar, por tanto, que, en la mayoría de los casos del archivo, las diferencias en *cilindrada* sean mayores que las diferencias en *peso*. Y este hecho quedará reflejado en la distancia euclídea.

Para eliminar del cálculo de las distancias el efecto debido a las diferencias en la métrica de las variables, se acostumbra a transformar las variables antes del análisis de manera que todas ellas tengan variabilidades similares. Entre las transformaciones disponibles, una bastante utilizada que permite igualar tanto la métrica como la variabilidad de las variables es la *tipificación*, es decir, la transformación en puntuaciones z con media 0 y varianza 1. El *Análisis de conglomerados de K medias* no incluye, entre sus opciones, la tipificación de las variables; si se desea incluir en el análisis las variables tipificadas, es necesario efectuar la transformación antes de iniciar el análisis. Para tipificar variables:

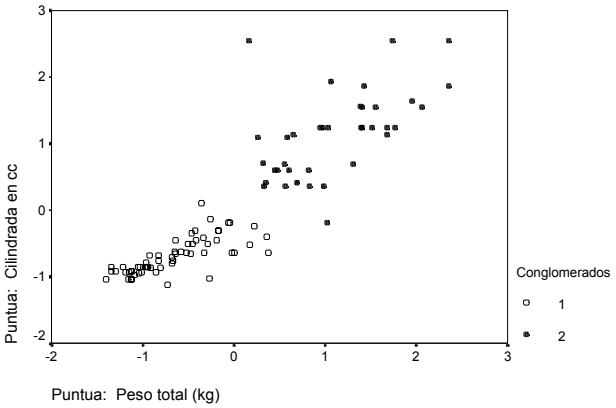
- Seleccionar la opción **Estadísticos descriptivos** > **Descriptivos** del menú **Analizar** para acceder al cuadro de diálogo *Descriptivos*.
- Seleccionar la(s) variable(s) que se desea transformar y trasladarlas a la lista **Variables**.
- Marcar la opción **Guardar valores tipificados como variables**.

Pulsando el botón **Aceptar**, el SPSS crea en el *Editor de datos* las variables tipificadas correspondientes a cada una de las variables seleccionadas en el cuadro de diálogo (las nuevas variables mantienen el nombre original con una *z* delante).

La figura 21.4 muestra el diagrama de dispersión obtenido al aplicar una análisis de conglomerados con los mismos casos y las mismas variables del ejemplo anterior, pero utilizando las variables tipificadas. Ahora, las agrupaciones resultantes parecen corresponderse mejor con

la estructura real de los datos. El caso que en la solución anterior pertenecía al primer conglomerado (*inferior*) pero cuyo peso era sensiblemente mayor que el del resto de vehículos de ese conglomerado ha pasado ahora pertenecer al segundo conglomerado (*superior*).

Figura 21.4. Solución de dos conglomerados utilizando las variables tipificadas.



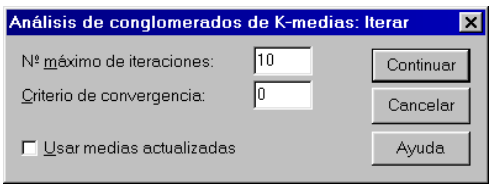
Sin embargo, aunque la solución puede clarificarse al utilizar variables tipificadas, esta estrategia posee un inconveniente que no podemos pasar por alto: se dificulta la interpretación de los resultados. Al tipificar las variables, los resultados de las tablas que informan de la ubicación de los *centroides* se encuentran también en escala tipificada, por lo que la ubicación relativa de los mismos no es interpretable en términos de las unidades de medida originales.

Iterar

El subcuadro de diálogo *Iterar* permite controlar algunos detalles relacionados con el proceso de iteración utilizado para el cálculo de los *centroides* finales. El acceso a este subcuadro de diálogo sólo es posible si se selecciona la opción **Iterar y clasificar** en el cuadro de diálogo principal. Para controlar las opciones relacionadas con el proceso de iteración:

- Pulsar en el botón **Iterar...** del cuadro de diálogo *Análisis de conglomerado de K-medias* (ver figura 21.1) para acceder al subcuadro de diálogo *Análisis de conglomerados de K-medias: Iterar* que muestra la figura 21.5.

Figura 21.5. Subcuadro de diálogo *Iterar*.



Nº máximo de iteraciones. Este cuadro de texto sirve para limitar el número de iteraciones que el algoritmo del procedimiento *K-medias* puede llevar a cabo. El proceso de iteración se detiene después del número de iteraciones especificado, incluso aunque no se haya alcanzado el criterio de convergencia. Puede introducirse un valor entre 1 y 999 (introduciendo el valor 0 se reproduce el algoritmo utilizado por el comando *Quick Cluster* de las versiones del SPSS anteriores a la 5.0).

Criterio de convergencia. Permite modificar el criterio de convergencia utilizado por el SPSS para detener el proceso de iteración. El valor de este criterio es, por defecto, cero, pero puede cambiarse introduciendo un valor diferente en el cuadro de texto. El valor introducido representa la proporción de la distancia mínima existente entre los *centros iniciales* de los conglomerados. Por tratarse de una proporción, este valor debe ser mayor o igual que cero y menor o igual que uno. Si se introduce un valor de, por ejemplo, 0,02, el proceso de iteración se detendrá cuando entre una iteración y la siguiente no se consiga desplazar ninguno de los *centros* una distancia superior al dos por ciento de la menor de las distancias existentes entre cualquiera de los *centros iniciales*. La tabla del historial de las iteraciones muestra, en una nota a pie de tabla, el desplazamiento obtenido en la última iteración (se haya alcanzado o no el criterio de convergencia).

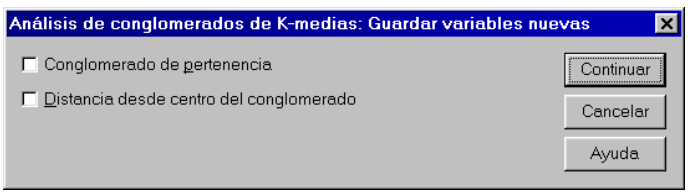
☐ **Usar medias actualizadas**. Permite solicitar la actualización de los centros de los conglomerados. Cuando se asigna un caso a uno de los conglomerados se calcula de nuevo el valor del centro del conglomerado. Cuando se selecciona la actualización de los centros de los conglomerados, el orden de los casos en el archivo de datos puede afectar a la solución obtenida. Si no se selecciona esta opción, los centros de los conglomerados finales se calculan después de la clasificación de todos los casos.

Guardar

Las opciones del subcuadro de diálogo guardar permiten guardar en el archivo de datos la información de clasificación para cada caso. Con ello se puede utilizar esta información en otros procedimientos. La información que se almacena es la misma que la presentada en el *Visor* en la tabla de información para cada caso (ver tabla 21.6). Para almacenar la información en el archivo de datos:

▶ Pulsar en el botón **Guardar** del cuadro de diálogo *Análisis de conglomerado de K-medias* (ver figura 21.1) para acceder al subcuadro de diálogo *Análisis de conglomerados de K-medias: Guardar* que muestra la figura 21.6.

Figura 21.6. Subcuadro de diálogo *Guardar variables nuevas*.



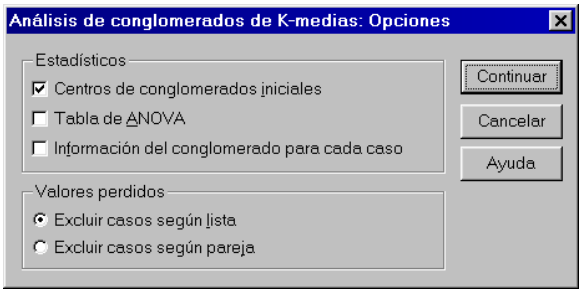
- **Conglomerado de pertenencia**. Crea una variable en el *Editor de datos* (con nombre *qcl_#*) cuyos valores indican el conglomerado final al que pertenece cada caso. Los valores de la nueva variable van desde 1 hasta el número de conglomerados. Esta información es útil, por ejemplo, para construir un diagrama de dispersión con marcas distintas para los casos pertenecientes a distintos conglomerados, o para llevar a cabo un análisis discriminante con intención de identificar la importancia relativa de cada variable en la diferenciación entre conglomerados.
- **Distancia desde el centro del conglomerado**. Crea una variable en el *Editor de datos* cuyos valores indican la distancia euclídea existente entre cada caso y el *centro* del conglomerado al que ha sido asignado.

Opciones

El cuadro de diálogo *Opciones* permite obtener algunos estadísticos y controlar el tratamiento que se desea dar a los valores perdidos. Para acceder a las opciones:

- Pulsar en el botón **Opciones...** del cuadro de diálogo *Análisis de conglomerados de K-medias* (ver figura 21.1) para acceder al subcuadro de diálogo *Análisis de conglomerados de K-medias: Opciones* que muestra la figura 21.7.

Figura 21.7. Subcuadro de diálogo *Opciones*.



Estadísticos. Las opciones de este apartado permiten seleccionar algunos estadísticos adicionales:

- **Centros de conglomerados iniciales**. Muestra una tabla con los casos que el procedimiento selecciona como *centros iniciales* de los conglomerados (ver tabla 21.1). Esta opción se encuentra seleccionada por defecto.
- **Tabla de ANOVA**. Muestra la tabla resumen del análisis de varianza con un estadístico *F* univariante para cada una de las variables incluidas en el análisis (tabla 21.5). El análisis de varianza se obtiene tomando los grupos definidos por los *conglomerados* como *factor* y cada una de las variables incluidas en el análisis como *variable dependiente*. Una nota al pie de tabla informa de que los estadísticos *F* sólo deben utilizarse

con una finalidad descriptiva pues los casos no se han asignado aleatoriamente a los conglomerados sino que se han asignado intentando optimizar las diferencias entre los conglomerados. Además, los niveles críticos asociados a los estadísticos *F* no deben ser interpretados de la manera habitual pues el procedimiento *K-medias* no aplica ningún tipo de corrección sobre la tasa de error (es decir, sobre la probabilidad de cometer errores tipo I cuando se llevan a cabo muchos contrastes). Lógicamente, la tabla de ANOVA no se muestra cuando todos los casos son asignados a un único conglomerado.

Tabla 21.5. Tabla resumen del ANOVA.

	Conglomerado		Error			
	Media cuadrática	gl	Media cuadrática	gl	F	Sig.
Cilindrada en cc	231888537.860	1	571249.969	96	405.932	.000
Peso total (kg)	5545469.651	1	23612.092	96	234.857	.000

☐ **Información del conglomerado para cada caso**. Muestra un listado de todos los casos utilizados en el análisis con indicación del conglomerado al que ha sido asignado cada caso y la distancia euclídea existente entre cada caso y el *centro* de su conglomerado (tabla 21.6). También muestra la distancia euclídea existente entre los *centros* de los conglomerados finales (tabla 21.7). Los casos se muestran en el mismo orden en el que se encuentran en el archivo de datos.

Tabla 21.6. Conglomerado de pertenencia y distancias a sus respectivos *centros*.

Número de caso	Conglomerado	Distancia
1	1	140.880
5	1	204.594
9	1	2370.203
10	1	1298.386
11	2	359.395
12	1	647.977
14	1	1187.008
20	1	2378.330
...	...	...
380	2	312.717
387	2	203.440
388	2	310.165
391	2	142.163
394	2	432.819
397	2	656.903
399	2	466.322
402	2	412.778

Tabla 21.7. Distancias entre los centros de los conglomerados finales.

Conglomerado	1	2
1		3228.772
2	3228.772	

Valores perdidos. Las opciones des este cuadro permiten controlar el tratamiento que se desea dar a los valores perdidos.

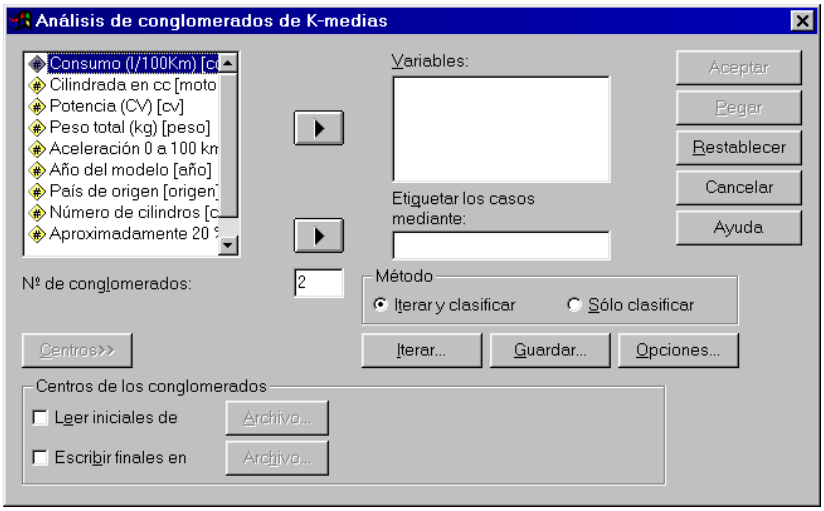
- **Excluir casos según lista**. Se excluyen los casos con valor perdido en cualquiera de las variables incluidas en el análisis. Es la opción por defecto.
- **Excluir casos según pareja**. Asigna los casos a los conglomerados a partir de las distancias calculadas en todas las variables en las que no tengan valores perdidos.

Centros

Ampliando el cuadro de diálogo, al pulsar en el botón centros, es posible escribir y leer los centros de los conglomerados. Para acceder a utilizar centros de conglomerados externos:

- Pulsar en el botón **Centros>>** del cuadro de diálogo *Análisis de conglomerado de K-medias* (ver figura 21.1) para expandir el cuadro de diálogo tal como muestra la figura 21.8.

Figura 21.8. Cuadro de diálogo Análisis de conglomerados de K-medias ampliado.



Centros de los conglomerados. El nuevo apartado obtenido al expandir el cuadro de diálogo original posee dos opciones:

- **Leer iniciales de**. Permite al usuario decidir qué valor deben tomar los *centros* de los conglomerados. El botón **Archivo...** sirve para indicar el nombre y ruta del archivo que contiene los valores de los *centros*. El nombre del archivo seleccionado se muestra junto al botón.

Lo habitual es designar un archivo resultante de una ejecución previa (guardado con la opción **Escribir finales en**) y en conjunción con la opción **Sólo clasificar** del apartado **Método**.

- **Escribir finales en**. Guarda los *centros* de los conglomerados finales en un archivo de datos externo. Este archivo puede utilizarse posteriormente para la clasificación de nuevos casos. El botón **Archivo...** permite asignar nombre y ruta al archivo de destino. El nombre del archivo seleccionado se muestra junto al botón.

Los archivos de datos utilizados por estas dos opciones contienen variables con nombres especiales reconocidas automáticamente por el sistema. No es recomendable generar libremente la estructura de estos archivos; es preferible dejar que sea el propio procedimiento el que los genere.