KIWITEC.
HIGH QUALITY TECH COURSES

Curso:

SPSS STATISTICS

Módulo I:

ANÁLISIS DE CONGLOMERADOS o CLUSTER ANALYSIS

www.kiwitec.es - secretaria@kiwitec.es

I. Introducción

El análisis de conglomerados o *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.), la *aglomeración* es menos restrictiva en sus supuestos (no exige linealidad, ni simetría, permite variables categóricas, etc.).

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.

En SPSS se disponen 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.

II. Análisis de conglomerados de K medias

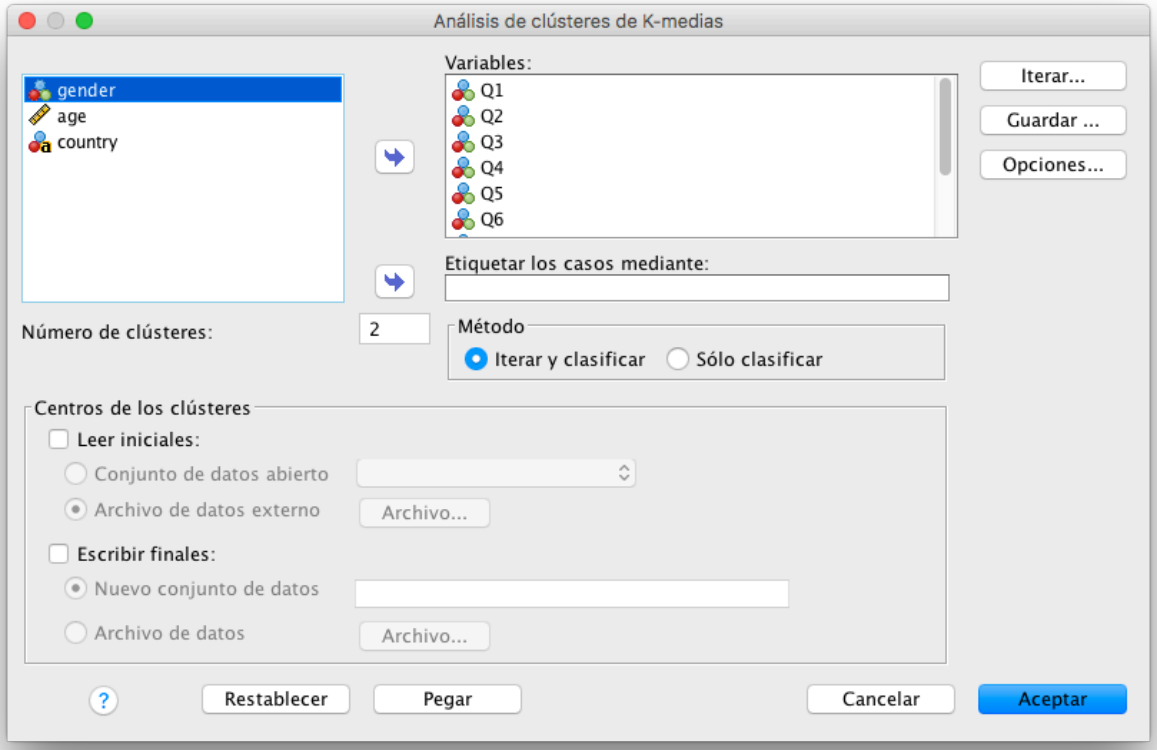
El análisis de conglomerados de K medias es un método de *agrupación de casos* que se basa en las distancias existentes entre ellos en un conjunto de variables (este método de aglomeración no permite agrupar variables). Se comienza seleccionado los K casos más distantes entre sí (el usuario debe determinar 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 hay que seleccionar la opción **Clasificar > Conglomerados de K medias** del menú **Analizar** para acceder a su cuadro de dialogo:

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* deben seleccionarse las variables numéricas que se desea a utilizar para diferenciar a los sujetos y formar los conglomerados, y trasladarlas a la lista **Variables.** Opcionalmente, se puede seleccionar una variable para identificar los casos (por

ejemplo un DNI, o un número de cliente o contrato) en las tablas de resultados y en los gráficos y trasladarla a la lista **Etiquetar casos mediante.**



Adicionalmente debe suministrarse el **Nº de conglomerados** que se desean extraer**.** En este cuadro de texto se encuentra seleccionada por defecto la solución de dos conglomerados. Para solicitar un número mayor de conglomerados, basta con introducir el número deseado en el cuadro.

1. Método de estimación

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

• **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 el acceso a las especificaciones del proceso de iteración. Esta opción suele utilizarse junto con la opción **Centros**, que se explica más adelante.

• **Iterar y clasificar.** El procedimiento se encarga de estimar los *centros* iterativamente y de clasificar a las sujetos con arreglo a los *centros* estimados.

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. Conforme avanzan las iteraciones, el desplazamiento de los centros se va haciendo más y más pequeño. 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).

Centros de clústeres iniciales

	Clúster 1	Clúster 2
Q1	4	1
Q2	4	1
Q3	3	1
Q4	4	1
Q5	0	4
Q6	3	1
Q7	4	1
Q8	0	4
Q9	1	4
Q10	1	4

Centros de clústeres finales

	Clúster 1	Clúster 2
Q1	4	3
Q2	4	3
Q3	2	3
Q4	3	3
Q5	2	3
Q6	3	2
Q7	3	2
Q8	2	3
Q9	2	3
Q10	2	3

En este procedimiento, SPSS muestra los *centros de los conglomerados finales* frente a los *centros iniciales* (antes de la iteración). Esto es de gran utilidad para interpretar la constitución de los conglomerados pues resume los valores centrales de cada conglomerado en las variables de interés.

Historial de iteraciones[a]

Iteración	Cambiar en centros de clústeres 1	2
1	3,666	3,794
2	,126	,111
3	,018	,017
4	,002	,002
5	,000	,000

a. Convergencia conseguida debido a que no hay ningún cambio en los centros de clústeres o un cambio pequeño. El cambio de la coordenada máxima absoluta para cualquier centro es ,000. La iteración actual es 5. La distancia mínimo entre los centros iniciales es 9,695.

En cualquiera de los métodos, SPSS informa sobre el número de casos asignado a cada conglomerado.

Número de casos en cada clúster

Clúster	1	2386,000
	2	2370,000
Válidos		4756,000
Perdidos		,000

Si se representase mediante un diagrama de dispersión con marcas distintas para los casos de uno y otro conglomerado podría formarse una idea bastante precisa de las características de cada conglomerado.

2. 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 (su rango o recorrido). Si dos variables tienen un rango muy diferente, 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.

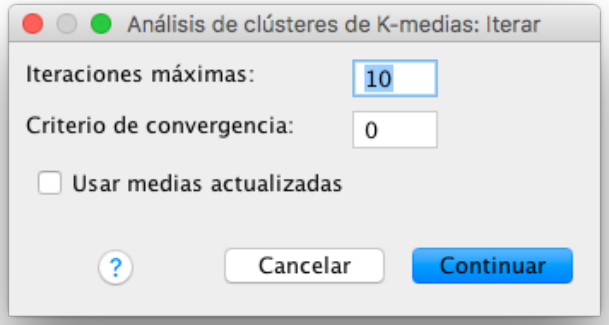
Para tipificar variables basta con seleccionar la opción **Estadísticos descriptivos** > **Descriptivos** del menú **Analizar** para acceder al cuadro de diálogo Descriptivos. Desde él, pueden seleccionarse la(s) variable(s) que se desea transformar y trasladarlas a la lista **Variables**, y por último marcar la opción **Guardar valores tipificados como variables**.

Pulsando el botón **Aceptar**, 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).

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.

3. 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 hay que pulsar en el botón **Iterar...** del cuadro de diálogo *Análisis de conglomerado de K-medias*. En él se pueden configurar las siguientes opciones:



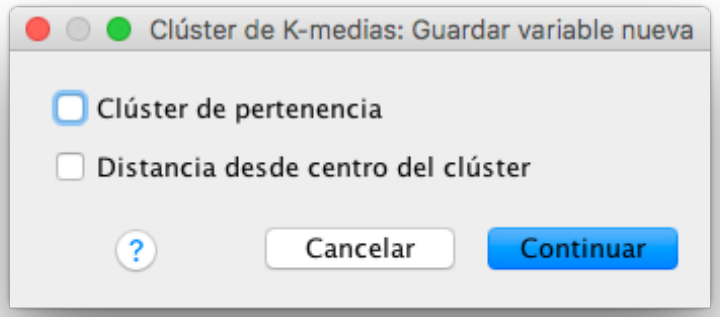
Nº máximo de iteraciones. Este cuadro de texto sirve para limitar el número de iteraciones que el algoritmo del procedimiento *K-medias* puede llevar a cabo. El proceso de iteración se detiene después del número de iteraciones especificado, incluso aunque no se haya alcanzado el criterio de convergencia. Puede introducirse un valor entre 1 y 999.

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 clasificar todos los casos.

4. 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. Para almacenar la información en el archivo de datos basta con pulsar en el botón **Guardar** del cuadro de diálogo *Análisis de conglomerado de K-medias* para acceder al subcuadro de diálogo *Análisis de conglomerados de K- medias: Guardar*:

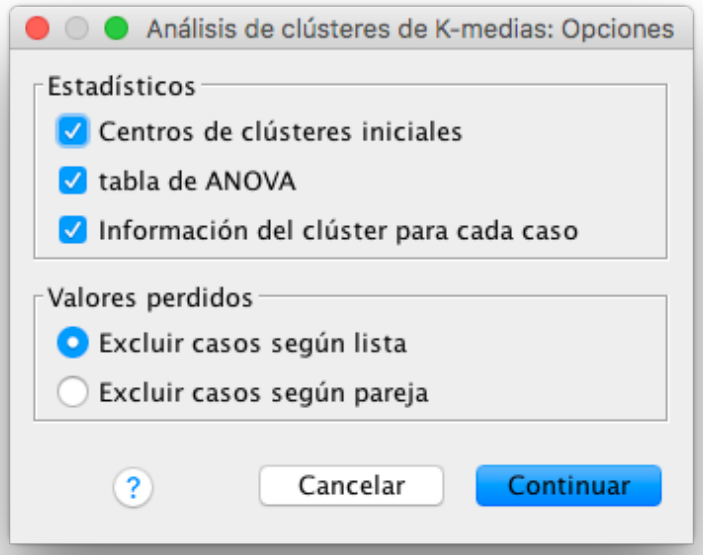


• **Conglomerado de pertenencia.** Crea una variable en el *Editor de datos* (con nombre *qcl_#*) cuyos valores indican el conglomerado final al que pertenece cada caso. Los valores de la nueva variable van desde 1 hasta el número de conglomerados. Esta información es útil, por ejemplo, para construir 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.

5. 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 basta con pulsar sobre el botón **Opciones**... del cuadro de diálogo *Análisis de conglomerados de K-medias*:



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. 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.

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.

ANOVA

	Clúster		Error			
	Media cuadrática	gl	Media cuadrática	gl	F	Sig.
Q1	1208,860	1	,504	4754	2398,439	,000
Q2	834,784	1	,518	4754	1612,421	,000
Q3	1947,857	1	,524	4754	3714,967	,000
Q4	683,011	1	,525	4754	1300,107	,000
Q5	1879,750	1	,591	4754	3180,691	,000
Q6	1892,199	1	,467	4754	4054,003	,000
Q7	1793,713	1	,521	4754	3440,183	,000
Q8	1231,245	1	,692	4754	1779,410	,000
Q9	1818,200	1	,616	4754	2950,222	,000
Q10	2788,953	1	,590	4754	4725,540	,000

Las pruebas F sólo se deben utilizar con fines descriptivos porque los clústeres se han elegido para maximizar las diferencias entre los casos de distintos clústeres. Los niveles de significación observados no están corregidos para esto y, por lo tanto, no se pueden interpretar como pruebas de la hipótesis de que los medias de clúster son iguales.

- **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. También muestra la distancia euclídea existente entre los centros de los conglomerados finales. Los casos se muestran en el mismo orden en el que se encuentran en el archivo de datos.

Valores perdidos. Las opciones de 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.

6. Centros

En el cuadro de diálogo, bajo la opción centros, es posible escribir y leer los centros de los conglomerados:

- **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.

III. Análisis de conglomerados jerárquico

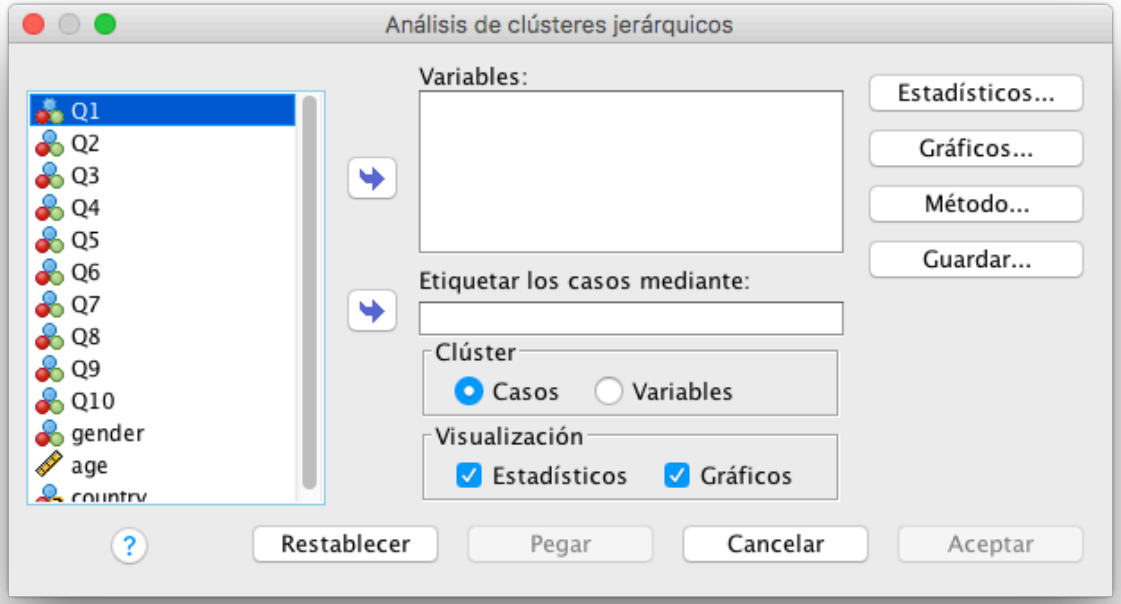
A diferencia de lo que ocurre con el procedimiento *Análisis de conglomerados de K medias*, el procedimiento *Análisis de conglomerados jerárquico* permite aglomerar tanto casos como variables y elegir entre una gran variedad de métodos de aglomeración y medidas de distancia. Pero la diferencia fundamental 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 basta con seleccionar la opción **Clasificar** > **Conglomerados jerárquicos** del menú *Analizar*:



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**.

Cluster. 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.**

Visualización. 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 ac- ceso al botón **Gráficos.**

Lo primero que muestra SPSS es 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 (por defecto *Distancia euclídea al cuadrado*) y el método de conglomeración utilizado (por defecto *Vinculación promedio*). La solución obtenida puede depender en gran medida de la combinación del tipo de medida de las distancias y el método de conglomeración.

Resumen de procesamiento de casos[a]

Casos					
Válido		Perdidos		Total	
N	Porcentaje	N	Porcentaje	N	Porcentaje
4756	100,0%	0	0,0%	4756	100,0%

a. Distancia euclídea al cuadrado utilizada

Posteriormente se muestra una tabla con el *historial del proceso de conglomeración*, etapa por etapa. En cada etapa se unen dos elementos.

Historial de conglomeración

Etapa	Clúster combinado		Coeficientes	Primera aparición del clúster de etapa		Etapa siguiente
	Clúster 1	Clúster 2		Clúster 1	Clúster 2	
1	6	7	2359,000	0	0	6
2	1	2	2504,000	0	0	4
3	9	10	3225,000	0	0	7
4	1	4	3299,000	2	0	6
5	3	5	3497,000	0	0	7
6	1	6	4823,167	4	1	9
7	3	9	4876,000	5	3	8
8	3	8	5414,500	7	0	9
9	1	3	13023,560	6	8	0

La columna Conglomerado que se combina informa sobre los conglomerados (o casos) fundidos en cada etapa. Como el análisis se inicia con todos los casos separados en conglomerados individuales,

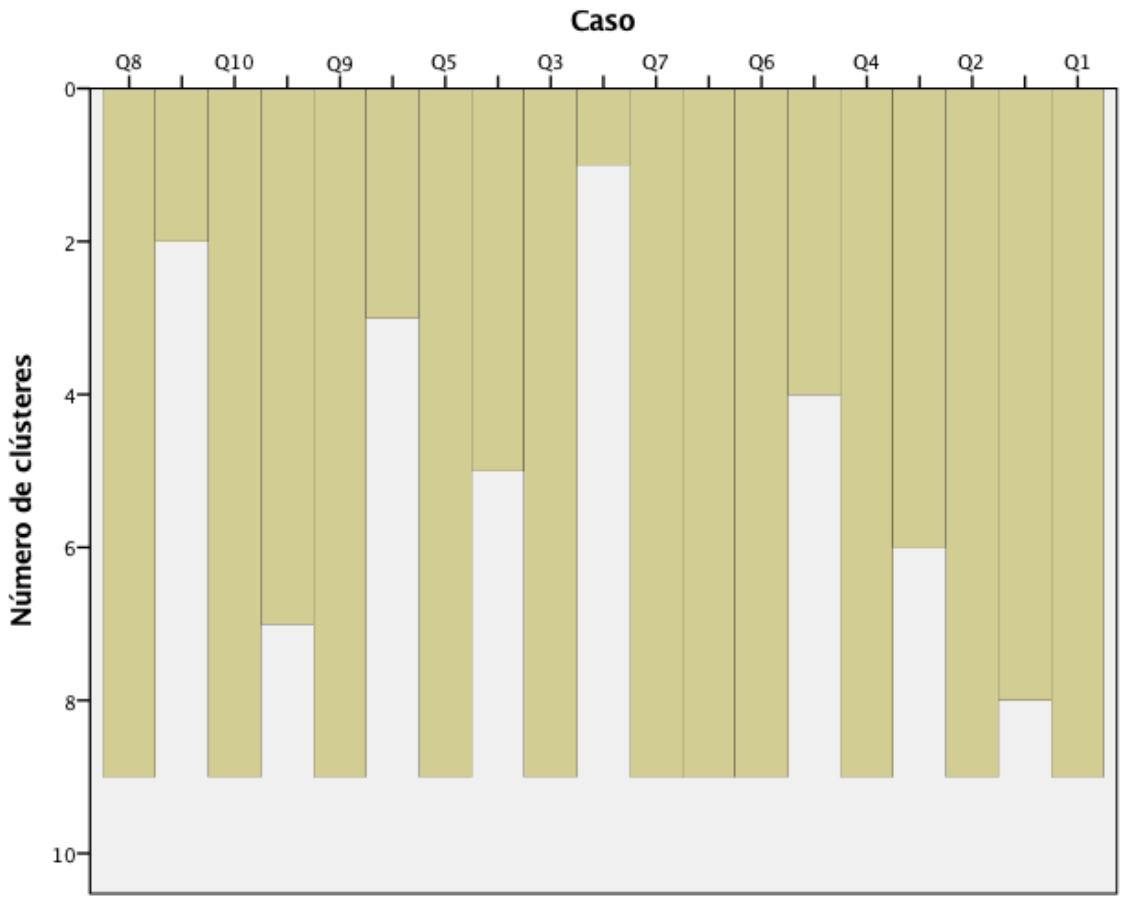
la primera etapa siempre se refiere a casos individuales. A partir de ese momento, los dos casos unidos constituyen un único conglomerado y son indivisibles en las etapas posteriores.

La columna *(Coeficientes)* ofrece el valor de la distancia a la que se encuentran los casos antes de la fusión. Si en algún momento la distancia de fusión entre los casos vale 0, esto significa que se trata de casos con idénticas puntuaciones.

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. Un valor de 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.

La columna *Próxima etapa* indica la etapa en la que el conglomerado que se acaba de formar volverá a fundirse con otros elementos.

El *diagrama de témpanos* 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.



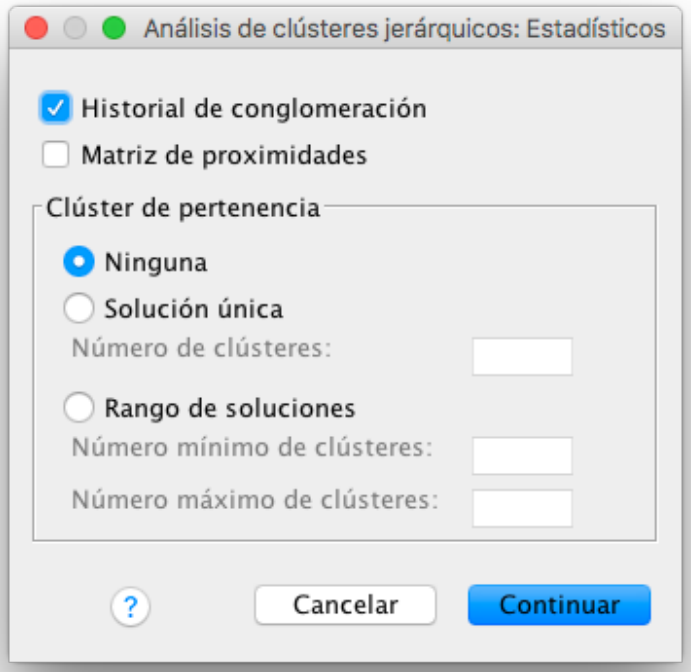
El diagrama de *témpanos* es de gran utilidad para identificar los elementos que constituyen cada una de las soluciones del análisis y cuáles han sido las formaciones previa y posterior a cada solución específica. Sin embargo, presenta el gran inconveniente de no informar en modo alguno de la distancia 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*).

1. 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 es necesario pulsar en el botón Estadísticos... del cuadro de diálogo Análisis de conglomerados jerárquico:



• **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. Esta opción se encuentra activa por defecto; desactivándola se anula la presentación del historial de conglomeración.

• **Matriz de proximidades.** Permite obtener la matriz de distancias entre los elementos analiza- dos. 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 le- janía). El tipo de matriz obtenida (de *similaridades* o de *disimilaridades*) depende de la medida seleccionada en el subcuadro de diálogo Método.

Si el análisis contiene un gran número de elementos, la tabla puede llegar a ser muy voluminosa.

Cúster 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.

Clúster de pertenencia

Caso	2 clústeres
Q1	1
Q2	1
Q3	2
Q4	1
Q5	2
Q6	1
Q7	1
Q8	2
Q9	2
Q10	2

- **Ninguno.** Esta opción, que se encuentra activa por defecto impide que el *Visor* muestre la tabla del *conglomerado de pertenencia*.

- **Solución única.** 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.

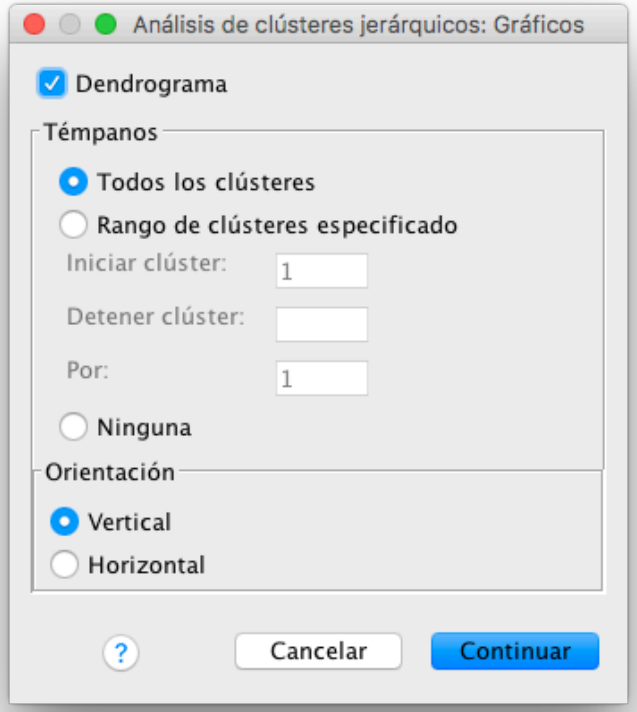
- **Rango de soluciones**. Permite obtener la *tabla del conglomerado de pertenencia* con la información correspondiente a varias soluciones (un rango de soluciones). Para establecer el número mínimo y máximo de conglomerados del rango de soluciones que se desea obtener hay que introducir, en el primer 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 en- teros mayores que 1 y el primer valor debe ser menor que el segundo.

Clúster de pertenencia

Caso	4 clústeres	3 clústeres	2 clústeres
Q1	1	1	1
Q2	1	1	1
Q3	2	2	2
Q4	1	1	1
Q5	2	2	2
Q6	1	1	1
Q7	1	1	1
Q8	3	3	2
Q9	4	2	2
Q10	4	2	2

2. Gráficos

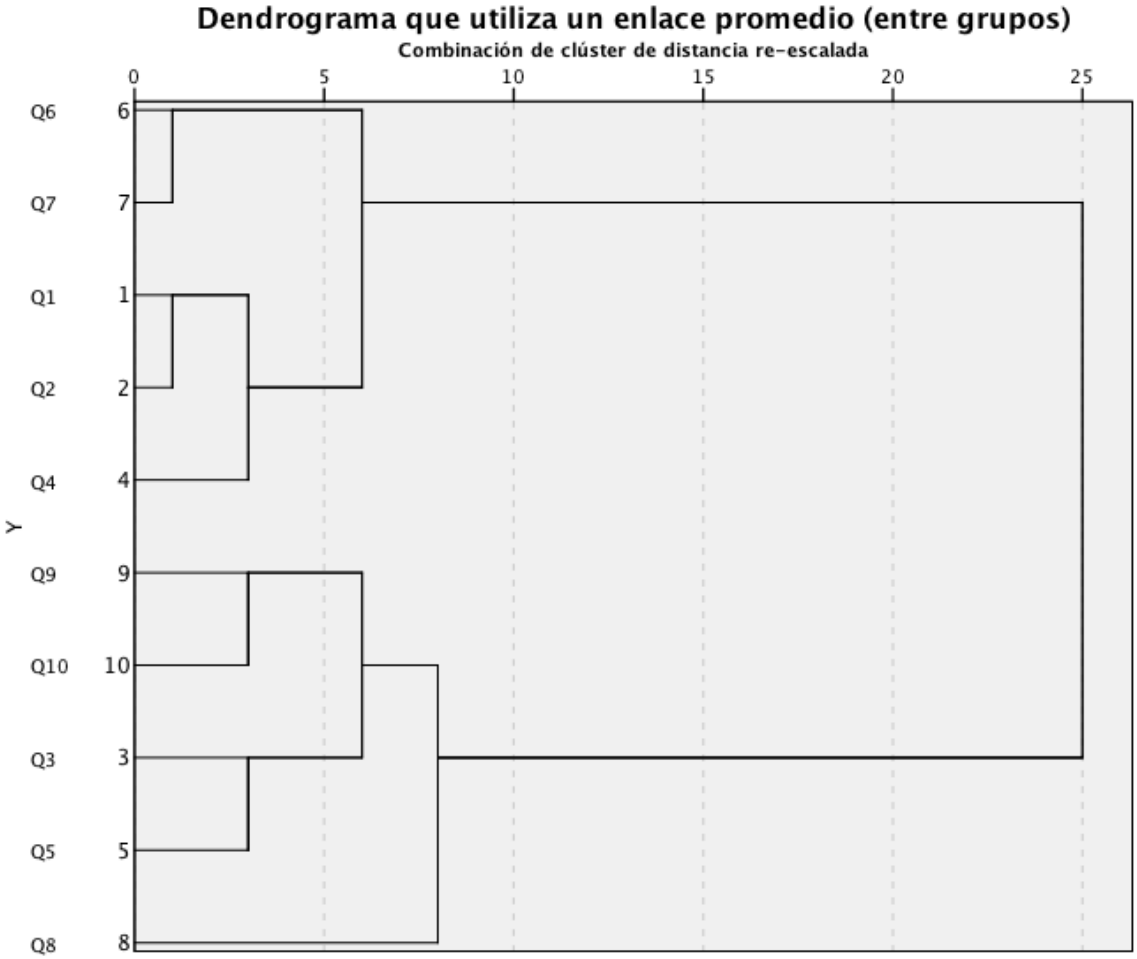
Las opciones del subcuadro de diálogo Gráficos permiten decidir qué tipos de gráficos se desea obtener. Para obtener esta información hay que pulsar en el botón **Gráficos**... del cuadro de diálogo *Análisis de conglomerado jerárquico*.



• **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.

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. 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.



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.

- **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.

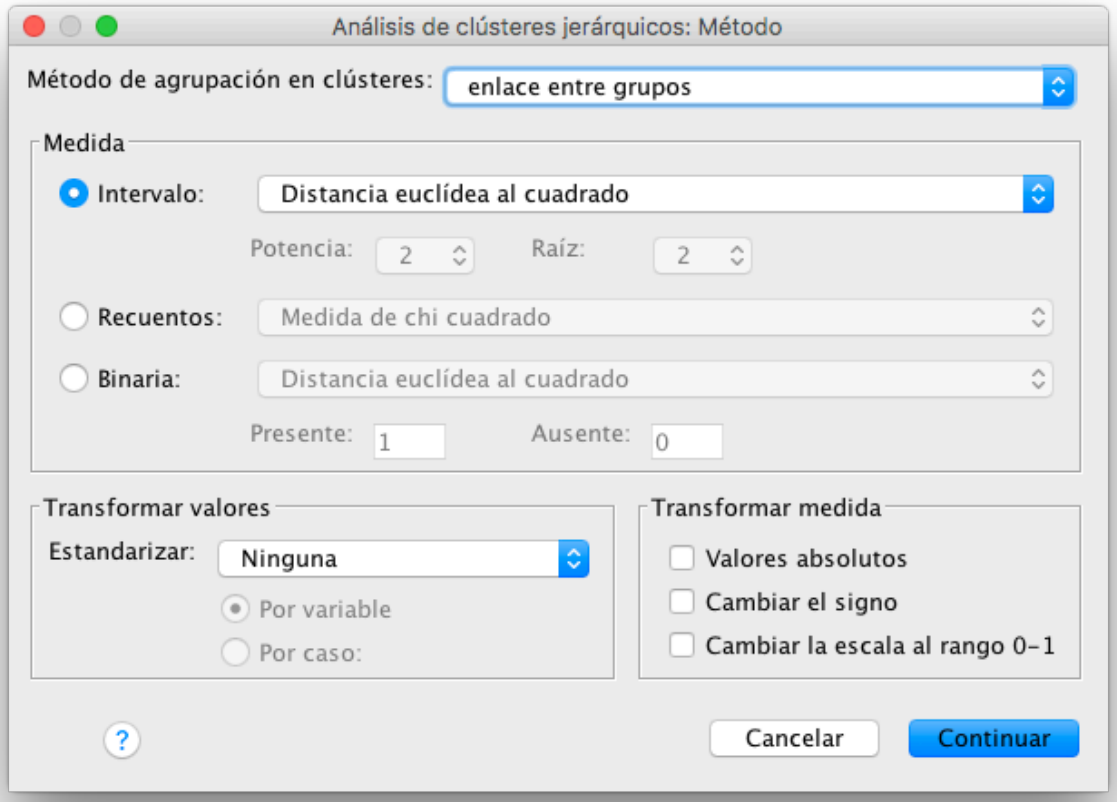
Orientación. Las opciones de este apartado permiten controlar la orientación del diagrama de témpanos:

• **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.

3. 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 basta con pulsar en el botón Método... del cuadro de diálogo Análisis de conglomerados jerárquico:



i. Método de agrupación en clústeres

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 nxn sujetos (o pxp 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 dAB entre

los conglomerados A y B se calcula mediante:

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

donde dij 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:

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

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:

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

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.

ii. Medida

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. Análisis de conglomerados jerárquico

Las opciones del apartado Medida 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:

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

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

$$CORRELATION(X,Y) = \frac{\sum_i Z_{X_i} Z_{Y_i}}{n-1}$$

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

$$CHEBICHEV(X,Y) = max(|X_i - Y_i|)$$

- **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: donde p es cualquier número entero positivo.

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

- **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: donde p y r son dos números enteros positivos cualesquiera.

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

Recuento

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:

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

- **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:

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

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.

		Variable Y_i		
		1	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 anterior 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:

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

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

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

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

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

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

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

- **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:

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

- **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:

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

donde:

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

- **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:

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

  donde t1 y t2 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:

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

- **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:

$$K2(X,Y) = \frac{\frac{a}{(a+b)} + \frac{a}{(a+c)}}{2}$$

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

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

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

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

- **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:

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

- **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:

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

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

$$SS2(X,Y) = \frac{1}{1 + 2 \cdot (b+c)}$$

- **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:

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

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

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

- **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:

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

- **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:

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

iii. 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.

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 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.

iv. Transformar medidas

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

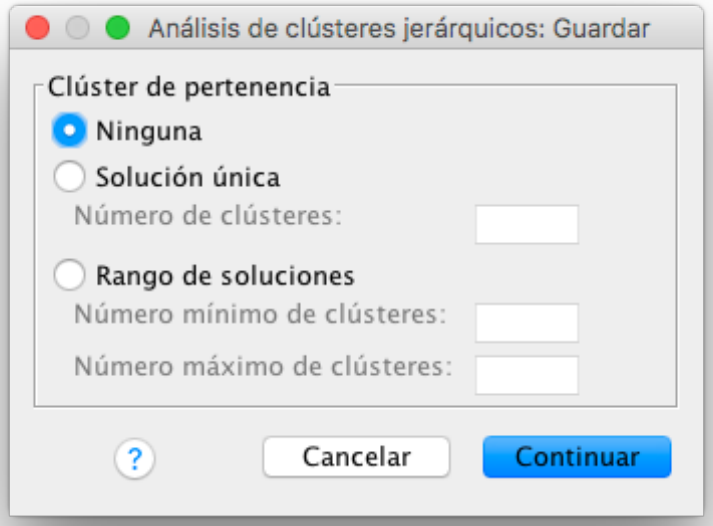
- **Valores absolutos.** Valor absoluto de las distancias calculadas.

- **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.

4. 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: hay que pulsar en el botón Guardar... del cuadro de diálogo Análisis de conglomerado para acceder al subcuadro de diálogo:



Cluúter 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.

- **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 Número de clústeres 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 Número mínimo de clústeres y Número máximo de clústeres permiten definir el rango de soluciones que se desea obtener. Para ello, hay que introducir un número

entero que indique el número de conglomerados de la solución con menos conglomerados; y un número entero que indique el número de conglomerados de la solución con más conglomerados.