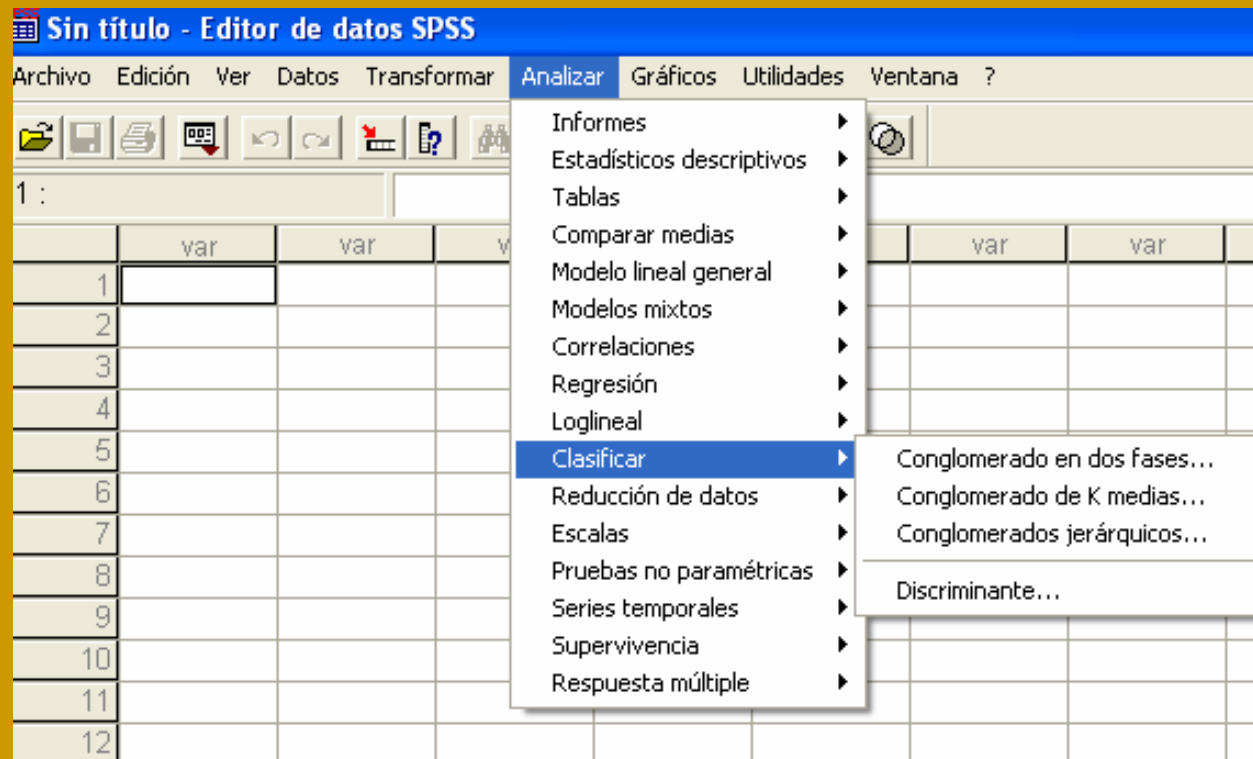

ANALISIS DE CLUSTER CON SPSS (PASW)

INMACULADA BARRERA

ANALISIS DE CLUSTER EN SPSS

Opción: **Analizar** → **Clasificar**



ANALISIS DE CLUSTER EN SPSS

Tres posibles OPCIONES

1.- Cluster en dos etapas

2.- K-means

3.- Jerárquicos

ANALISIS DE CLUSTER EN SPSS

1.- Cluster en dos etapas.- está pensado para minería de datos, es decir para estudios con un número de individuos grande que pueden tener problemas de clasificación con los otros procedimientos.

Otra peculiaridad es que permite trabajar conjuntamente con variables de tipo mixto (cuali y cuantitativas). Puede realizarse cuando el número de cluster es conocido a priori y también cuando no se conoce.

ANALISIS DE CLUSTER EN SPSS

2.- Cluster no jerárquicos .- sólo puede ser aplicado a variables cuantitativas y requiere conocer el número de cluster a priori.

Puede realizarse para un número de objetos relativamente grande pues no requiere el cálculo de todas las posibles distancias.

ANALISIS DE CLUSTER EN SPSS

3.- *Jerárquicos.-*

Para variables cuantitativas o bien para variables cualitativas

Si no se conoce el número de cluster a priori y cuando el número de objetos no es muy grande.

CLUSTER JERÁRQUICOS.-

El primer paso es la selección de variables:



Como se observa pueden etiquetarse los grupos con una de las variables del fichero.

También es posibles realizar conglomerados no para objetos sino para variables, (agrupar variables por el parecido que presentan en las respuestas de los individuos)

CLUSTER JERÁRQUICOS.-

OPCIÓN METODO:

-Podremos estandarizar las variables utilizadas en el análisis antes de utilizarlas el cálculo de las similaridades si fuese necesario. Los métodos disponibles son varios.

-Permite seleccionar la medida usada para ver el parecido entre individuos con distintas distancias dependiendo si la variable es binaria, frecuencias o de intervalo.

-Es posible también elegir el método para obtener los conglomerados Todos los vistos .

Los dos primeros vinculación Inter-grupos y dentro de grupos se corresponde a la opción denominada UPGMA (método del promedio) y una variante de este donde se consideran para el cálculo de la distancia media la correspondiente a todos los posibles pares del grupo resultante y no sólo a los formados con un elemento de cada grupo como en el anterior.

Análisis de conglomerados jerárquico: Método

Método de conglomeración: Vinculación inter-grupos

Medida:

- ☒ Intervalo: Distancia euclídea al cuadrado
Potencia: 2 Raíz: 2
- ☐ Frecuencias: Medida de Chi-cuadrado
- ☐ Binaria: Distancia euclídea al cuadrado
Presente: 1 Ausente: 0

Transformar valores:

Estandarizar: Ninguno

- ☒ Por variable
- ☐ Por caso

Transformar medidas:

- ☐ Valores absolutos
- ☐ Cambiar el signo
- ☐ Cambiar escala al rango 0-1

Continuar Cancelar Ayuda

ESTANDARIZAR

Análisis de conglomerados jerárquico: Método ✕

Método de conglomeración: Vinculación inter-grupos

Medida

☒ Intervalo: Distancia euclídea al cuadrado

Potencia: 2 Raíz: 2

☐ Frecuencias: Medida de Chi-cuadrado

☐ Binaria: Distancia euclídea al cuadrado

Presente: 1 Ausente: 0

Transformar valores

Estandarizar: Ninguno

Ninguno
Puntuaciones Z
Rango -1 a 1
Rango 0 a 1
Magnitud máxima de 1

Transformar medidas

☐ Valores absolutos

☐ Cambiar el signo

☐ Cambiar escala al rango 0-1

Continuar

Cancelar

Ayuda

MEDIDA

Análisis de conglomerados jerárquico: Método

Método de conglomeración: Vinculación inter-grupos

Medida:

☒ Intervalo: Distancia euclídea al cuadrado

☐ Frecuencias: Correlación de Pearson

☐ Binaria: Chebychev

Presente: 1 Ausente: 0

Transformar valores:

Estandarizar: Ninguno

☒ Por variable

☐ Por caso

Transformar medidas:

☐ Valores absolutos

☐ Cambiar el signo

☐ Cambiar escala al rango 0-1

Continuar

Cancelar

Ayuda

Análisis de conglomerados jerárquico: Método

Método de conglomeración: Vinculación inter-grupos

Medida:

☐ Intervalo: Distancia euclídea al cuadrado

Potencia: 2 Raíz: 2

☐ Frecuencias: Medida de Chi-cuadrado

☒ Binaria: Concordancia simple

Transformar valores:

Estandarizar: Ninguno

☐ Por variable

☐ Por caso

Transformar medidas:

☐ Valores absolutos

☐ Cambiar el signo

☐ Cambiar escala al rango 0-1


Continuar

Cancelar

Ayuda

13	ITALIA	2,90	13,70
14	HOLANDA	3,60	23,40

METODO

Análisis de conglomerados jerárquico: Método 

Método de conglomeración: Vinculación inter-grupos

Medida

☐ Intervalo: Vinculación inter-grupos
Vinculación intra-grupos
Vecino más próximo
Vecino más lejano
Agrupación de centroides
Agrupación de medianas

☐ Frecuencias: Medida de Chi-cuadrado

☒ Binaria: Concordancia simple

Presente: 1 Ausente: 0

Transformar valores

Estandarizar: Ninguno

☒ Por variable

☐ Por caso

Transformar medidas

☐ Valores absolutos

☐ Cambiar el signo

☐ Cambiar escala al rango 0-1

Continuar

Cancelar

Ayuda

CLUSTER JERÁRQUICOS.-

OPCIÓN ESTADÍSTICOS:

Historial muestra los casos o conglomerados combinados en cada etapa, las distancias entre los casos combinados y el último nivel del proceso de aglomeración en el que cada caso se unió al conglomerado correspondiente

Análisis de conglomerados jerárquico: Estadísticos

☒ Historial de conglomeración

☐ Matriz de distancias

Conglomerado de pertenencia

☒ Ninguno

☐ Solución única

Número de conglomerados:

☐ Rango de soluciones

Número mínimo de conglomerados:

Número máximo de conglomerados:

Continuar

Cancelar

Ayuda

Historial de conglomeración

Etapa	Conglomerado que se combina		Coeficientes	Etapa en la que el conglomerado aparece por primera vez		Próxima etapa
	Conglom erado 1	Conglom erado 2		Conglom erado 1	Conglom erado 2	
1	6	20	23,000	0	0	4
2	4	25	23,770	0	0	8
3	12	21	25,980	0	0	10
4	6	15	37,810	1	0	17
5	3	9	40,090	0	0	7
6	14	24	42,580	0	0	10
7	3	22	49,735	5	0	15
8	4	18	58,615	2	0	22
9	10	13	63,630	0	0	14
10	12	14	68,300	3	6	12
11	5	16	72,150	0	0	16
12	2	12	76,675	0	10	15
13	17	19	77,240	0	0	20
..

CLUSTER JERÁRQUICOS.-

「OPCIÓN ESTADÍSTICOS:

Matriz distancias

Conglomerado de pertenencia

nos da el conglomerado al que se asigna cada caso pudiendo elegir entre una única solución o un rango de soluciones En el ejemplo hemos seleccionado entre 2 y 3 cluster.

Análisis de conglomerados jerárquico: Estadísticos

☒ Historial de conglomeración

☐ Matriz de distancias

Conglomerado de pertenencia

☒ Ninguno

☐ Solución única

Número de conglomerados:

☐ Rango de soluciones

Número mínimo de conglomerados:

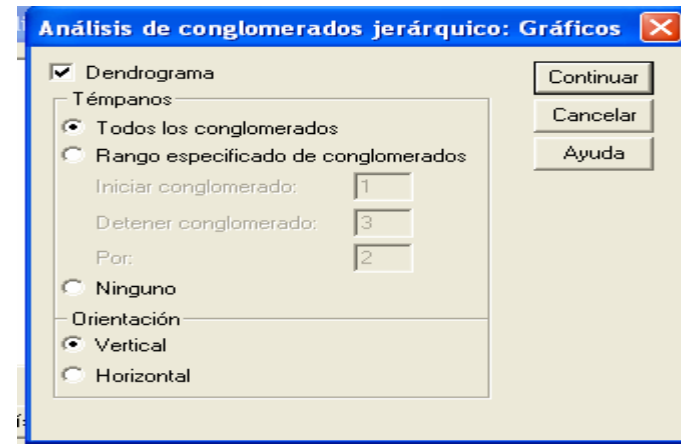
Número máximo de conglomerados:

Conglomerado de pertenencia		
Caso	3 conglome rados	2 conglome rados
1:ALBANIA	1	1
2:AUSTRIA	2	2
3:BELGICA	2	2
4:BULGARI	1	1
5:CHECOS	1	1
6:DINAMAR	2	2
7:ALE. Or	3	2
8:FINLAND	2	2
9:FRANCIA	2	2
10:GRECIA	1	1
11:HUNGRIA	1	1
12:IRLANDA	2	2
13:ITALIA	1	1
14:HO LANDA	2	2

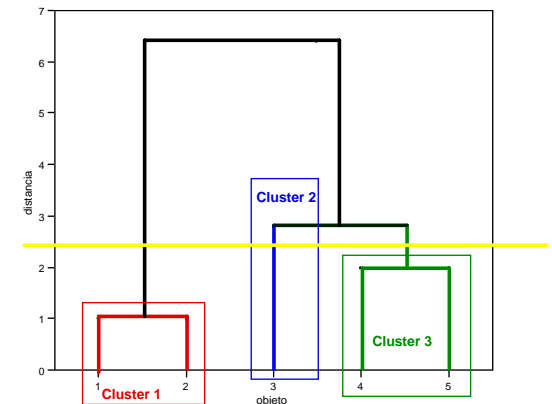
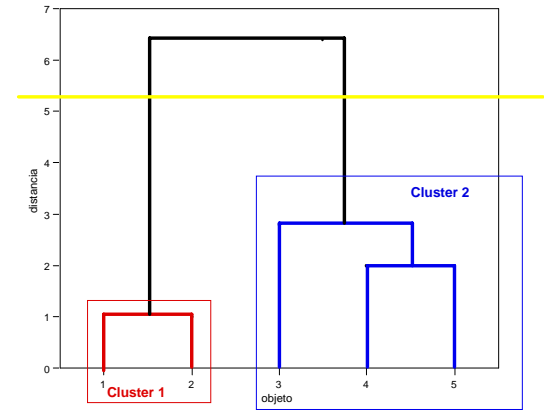
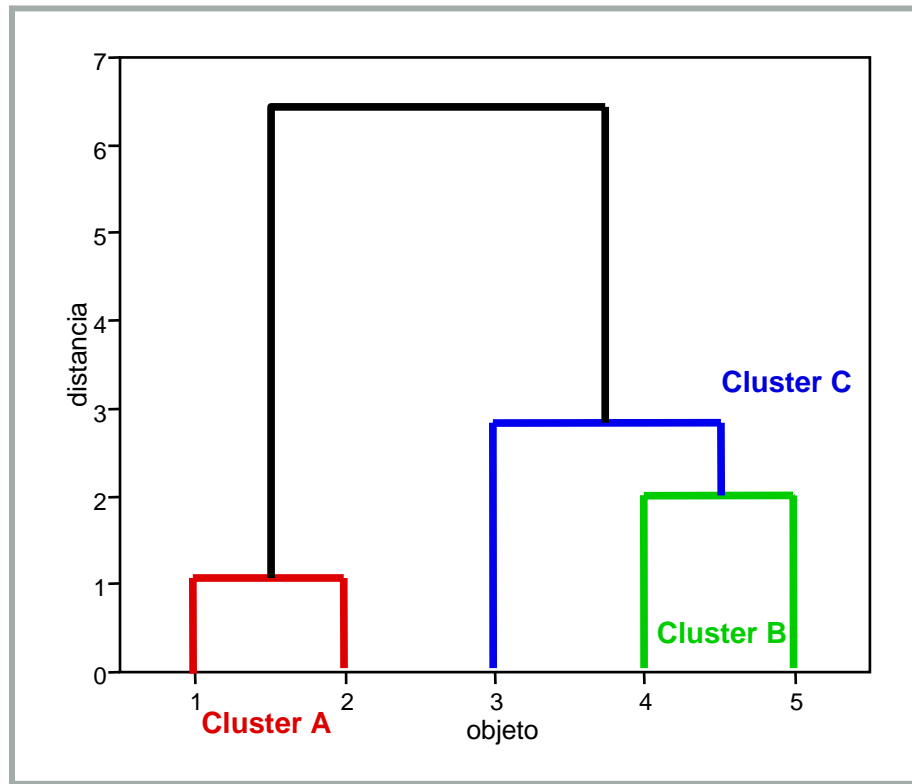
CLUSTER JERÁRQUICOS.-

」 OPCIÓN GRÁFICOS

Permite obtener el dendrograma y los vertical u horizontal icycle plots, o diagramas de témpanos.



CLUSTER JERÁRQUICOS.-



CLUSTER JERÁRQUICOS.-

OPCIÓN GUARDAR

Permite guardar los conglomerados de pertenencia para una solución única o para un rango de soluciones. Las variables guardadas pueden emplearse en análisis posteriores para explorar otras diferencias entre grupos.



PROCEDIMIENTO K-MEANS



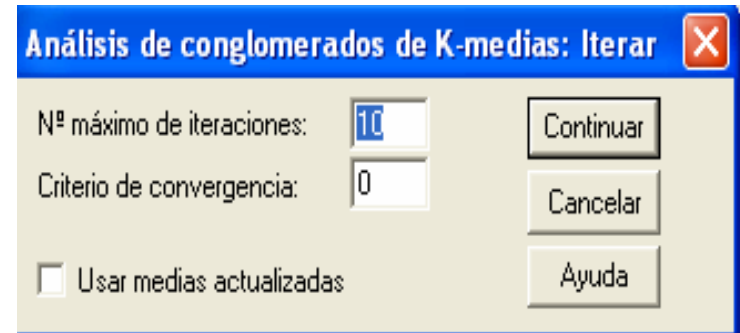
Una vez seleccionadas las variables y determinado el número de conglomerados que deseamos obtener podemos elegir entre iterar y clasificar o sólo clasificar. Para obtener máxima eficacia, podemos tomar una muestra de casos utilizar el método iterar y clasificar para determinar los centros de los conglomerados. Seleccionamos escribir finales en archivo. Después repetimos el análisis con sólo clasificar leyendo los iniciales del archivo anterior

PROCEDIMIENTO K-MEANS

」 OPCIÓN ITERAR

Para la opción iterar se puede determinar el número máximo de iteraciones, o bien fijar un criterio de convergencia mayor de cero y menor de uno.

La opción usar medias actualizadas recalcula centroides con cada individuo asignado al grupo, sino deselecciona esta opción no se recalculan hasta que todos los individuos están asignados.



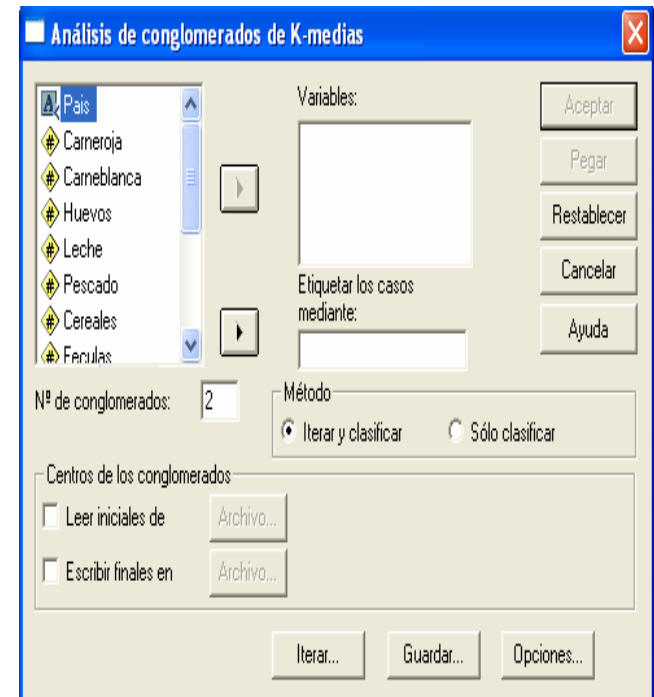
The screenshot shows a dialog box titled "Análisis de conglomerados de K-medias: Iterar" with a red 'X' button in the top right corner. The dialog contains the following elements:

- A label "Nº máximo de iteraciones:" followed by a text input field containing the value "10".
- A label "Criterio de convergencia:" followed by a text input field containing the value "0".
- A checkbox labeled "Usar medias actualizadas" which is currently unchecked.
- Three buttons on the right side: "Continuar", "Cancelar", and "Ayuda".

PROCEDIMIENTO K-MEANS

─ Opción guardar

permite crear una nueva variable que indica para cada caso el conglomerado al que pertenece y si se quiere otra variable con la distancia entre cada caso y su centro de clasificación.




PROCEDIMIENTO K-MEANS

BOTÓN OPCIONES

Centros iniciales de los conglomerados

	Conglomerado		
	1	2	3
Cereales	40,10	56,70	18,60
Feculas	4,00	1,10	5,20
frutossecos	5,40	3,70	1,50
frutasverduras	4,20	4,20	3,80


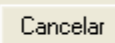
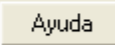
Análisis de conglomerados de K-medias: Opciones 

Estadísticos

- ☒ Centros de conglomerados iniciales
- ☐ Tabla de ANOVA
- ☐ Información del conglomerado para cada caso

Valores perdidos

- ☒ Excluir casos según lista
- ☐ Excluir casos según pareja

Distancias entre los centros de los conglomerados finales

Conglomerado	1	2	3
1		14,925	14,864
2	14,925		29,698
3	14,864	29,698	

PROCEDIMIENTO K-MEANS

」 **BOTÓN OPCIONES**

ANOVA

Análisis de conglomerados de K-medias: Opciones

Estadísticos

- ☒ Centros de conglomerados iniciales
- ☐ Tabla de ANOVA
- ☐ Información del conglomerado para cada caso

Valores perdidos

- ☒ Excluir casos según lista
- ☐ Excluir casos según pareja

Continuar

Cancelar

Ayuda

	Conglomerado		Error		F	Sig.
	Media cuadrática	gl	Media cuadrática	gl		
Cereales	1324,656	2	10,972	22	120,727	,000
Feculas	9,194	2	2,077	22	4,426	,024
frutossecos	15,636	2	2,880	22	5,429	,012
frutasverduras	1,832	2	3,383	22	,542	,589

Las pruebas F sólo se deben utilizar con una finalidad descriptiva puesto que los conglomerados han sido elegidos para maximizar las diferencias entre los casos en diferentes conglomerados. Los niveles críticos no son corregidos, por lo que no pueden interpretarse como pruebas de la hipótesis de que los centros de los conglomerados son iguales.

PROCEDIMIENTO K-MEANS

」 **BOTÓN OPCIONES**

Número de casos en cada conglomerado

Conglomerado	1	7,000
	2	3,000
	3	15,000
Válidos		25,000
Perdidos		,000

Análisis de conglomerados de K-medias: Opciones

Estadísticos

- ☒ Centros de conglomerados iniciales
- ☐ Tabla de ANOVA
- ☐ Información del conglomerado para cada caso

Valores perdidos

- ☒ Excluir casos según lista
- ☐ Excluir casos según pareja

Continuar

Cancelar

Ayuda

PROCEDIMIENTO DE CLUSTER EN DOS PASOS

- Está basado en un algoritmo que produce resultados óptimos si todas las variables son independientes, las continuas normalmente distribuidas y las categóricas multinomiales, pero funciona razonablemente bien en ausencia de estos supuestos.
- La solución final depende del orden de entrada de los datos. Para minimizar el efecto habríamos de ordenar el fichero de forma aleatoria.
- Pasos:
 - primer paso: formación de precluster* de los casos originales, Estos son clusters de los datos originales que se utilizarán en lugar de las filas del fichero original para realizar los *cluster jerárquicos en el segundo paso*. Todos los casos pertenecientes a un mismo precluster se tratan como un entidad sencilla.

PROCEDIMIENTO DE CLUSTER EN DOS PASOS

-Seleccionaremos las variables categóricas y continuas que formaran parte del análisis

-*Elegiremos las distancias:*

-Cuando se tengan **datos mixtos** la distancia que debemos de utilizar es el log-verosimilitud. La distancia entre dos clusters dependerá del decremento en el log-verosimilitud cuando ambas se combinan en un único cluster. Si se trata de **datos continuos** se puede usar la distancia euclídea entre los centros de los clusters.

The screenshot shows a software dialog box titled "Análisis de conglomerados en dos fases". It contains several sections for configuring a two-phase cluster analysis:

- Variables:** A list of variables on the left includes "Pais [Pais]", "Carneroja [Carneroja]", "Carneblanca [Carne...]", "Huevos [Huevos]", "Leche [Leche]", "Pescado [Pescado]", "Cereales [Cereales]", "Feculas [Feculas]", "frutossecos [frutose...", "frutasverduras [fruta...", and "Número de conglom...". Arrows point from this list to two empty boxes labeled "Variables categóricas:" and "Variables continuas:".
- Medida de distancia:** Two radio buttons are present: "Log-verosimilitud" (selected) and "Euclídea".
- Número de conglomerados:** Two options are shown: "Determinar automáticám." (selected) with a "Máximo:" value of 15, and "Especificar número fijo" with an empty "Número:" field.
- Recuento de variables continuas:** Two fields are shown: "Para tipificar:" with a value of 0, and "Asumidas como tipificadas:" with a value of 0.
- Criterio de conglomeración:** Two radio buttons are present: "Criterio bayesiano de Schwarz (BIC)" (selected) and "Criterio de información Akaike (AIC)".

On the right side of the dialog, there are buttons for "Aceptar", "Pegar", "Restabl.", "Cancelar", and "Ayuda". At the bottom, there are buttons for "Opciones...", "Gráficos...", and "Resultados...".

PROCEDIMIENTO DE CLUSTER EN DOS PASOS

La opción *número de clusters* permite especificar el número deseado de conglomerados o dejar que el algoritmo seleccione el número de clusters basado en dos criterios BIC (criterio Bayesiano) o AIC (criterio de información de Akaike).

El método requiere estandarización de todas las variables por lo que por defecto la efectúa y nos informa del número de variables a estandarizar.

The screenshot shows a software window titled "Análisis de conglomerados en dos fases". It contains several sections for configuring a two-phase cluster analysis. On the left, a list of variables includes "Pais [Pais]", "Carneroja [Carneroja]", "Carneblanca [Carne...", "Huevos [Huevos]", "Leche [Leche]", "Pescado [Pescado]", "Cereales [Cereales]", "Feculas [Feculas]", "frutossecos [frutosse...", "frutasverduras [fruta...", and "Número de conglom...". Two right-pointing arrows allow moving variables to the "Variables categóricas:" and "Variables continuas:" boxes on the right. Below the variable list, the "Medida de distancia" section has radio buttons for "Log-verosimilitud" (selected) and "Euclídea". The "Número de conglomerados" section has radio buttons for "Determinar automáticám." (selected) and "Especificar número fijo". The "Determinar automáticám." option includes a "Máximo:" field with the value "15". The "Especificar número fijo" option includes a "Número:" field. The "Criterio de conglomeración" section has radio buttons for "Criterio bayesiano de Schwarz (BIC)" (selected) and "Criterio de información Akaike (AIC)". The "Recuento de variables continuas" section shows "Para tipificar:" as 0 and "Asumidas como tipificadas:" as 0. On the far right, there are buttons for "Aceptar", "Pegar", "Restabl.", "Cancelar", and "Ayuda". At the bottom of the window are buttons for "Opciones...", "Gráficos...", and "Resultados...".

PROCEDIMIENTO DE CLUSTER EN DOS PASOS

Analisis de conglomerados en dos fases

Variables categóricas:

Variables continuas:

Medida de distancia:

Número de conglomerados:

Criterio de conglomeración:

Para tipificar:

Asumidas como tipificadas:

Aceptar Pegar Restablecer Cancelar Ayuda

Conglomerados en dos fases: Resultado

Resultado de Model Viewer

Gráficos y tablas (en Model Viewer)

Las variables especificadas como campos de evaluación pueden mostrarse opcionalmente en Visor de modelos como descriptores de conglomerado.

Variables:

Campos de evaluación:

Archivo de datos de trabajo:

Crear variable del conglomerado de pertenencia

Archivos XML:

Exportar modelo final

Nombre:

Examinar

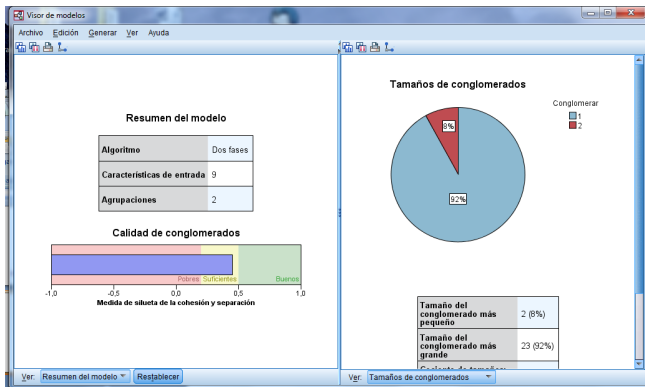
Exportar árbol CF

Nombre:

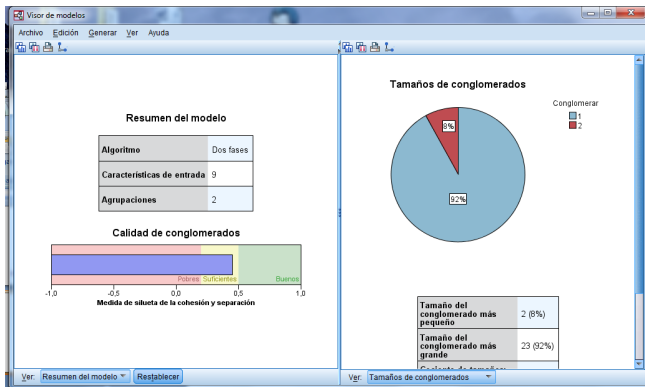
Examinar

Continuar Cancelar Ayuda

PROCEDIMIENTO DE CLUSTER EN DOS PASOS



PROCEDIMIENTO DE CLUSTER EN DOS PASOS



PROCEDIMIENTO DE CLUSTER EN DOS PASOS

OPCIÓN GRÁFICOS

Gráfico de importancia de variables : :

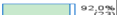

:Muestra varios gráficos que indican la importancia de cada variable en cada conglomerado.

Los resultados se pueden ordenar según el nivel de importancia de cada variable por conglomerado o por variable. En el primer caso para cada conglomerado se crearan gráficos por orden de importancia de variables. En el segundo caso para cada variable por conglomerados.

PROCEDIMIENTO DE CLUSTER EN DOS PASOS



Agrupaciones



Conglomerar Etiqueta	1	2
Descripción		
Tamaño	 92,0% (23)	 8,0% (2)
Funciones	<p>Frutos y vegetales 3,90</p> <p>Pescado 3,70</p> <p>Leche 18,00</p> <p>Carne Blanca 8,30</p> <p>Frutos secos 2,90</p> <p>Carne Roja 10,10</p> <p>Féculas 4,10</p> <p>Huevos 3,00</p> <p>Cereales 32,60</p>	<p>Frutos y vegetales 7,60</p> <p>Pescado 10,60</p> <p>Leche 6,80</p> <p>Carne Blanca 3,60</p> <p>Frutos secos 5,30</p> <p>Carne Roja 6,70</p> <p>Féculas 5,80</p> <p>Huevos 2,10</p> <p>Cereales 28,10</p>

Agrupaciones





Conglomerar Etiqueta	1	2
Descripción		
Tamaño	 92,0% (23)	 8,0% (2)
Funciones	<p>Pescado 3,70</p> <p>Frutos y vegetales 3,90</p> <p>Leche 18,00</p> <p>Carne Blanca 8,30</p> <p>Frutos secos 2,90</p> <p>Carne Roja 10,10</p> <p>Féculas 4,10</p> <p>Huevos 3,00</p> <p>Cereales 32,60</p>	<p>Carne Blanca 3,60</p> <p>Féculas 5,80</p> <p>Frutos y vegetales 7,60</p> <p>Carne Roja 6,70</p> <p>Leche 6,80</p> <p>Cereales 28,10</p> <p>Frutos secos 5,30</p> <p>Pescado 10,60</p> <p>Huevos 2,10</p>

PROCEDIMIENTO DE CLUSTER EN DOS PASOS

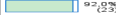

Agrupaciones



Conglomerar Etiqueta	1	2
Descripción		
Tamaño	 92,0% (23)	 8,0% (2)
Funciones	<p>Frutos y vegetales 3,90</p> <p>Pescado 3,70</p> <p>Leche 18,00</p> <p>Carne Blanca 8,30</p> <p>Frutos secos 2,90</p> <p>Carne Roja 10,10</p> <p>Féculas 4,10</p> <p>Huevos 3,00</p> <p>Cereales 32,60</p>	<p>Frutos y vegetales 7,60</p> <p>Pescado 10,60</p> <p>Leche 6,80</p> <p>Carne Blanca 3,60</p> <p>Frutos secos 5,30</p> <p>Carne Roja 6,70</p> <p>Féculas 5,80</p> <p>Huevos 2,10</p> <p>Cereales 28,10</p>

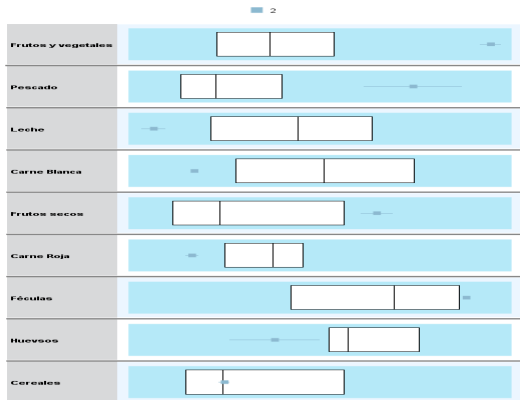
Agrupaciones



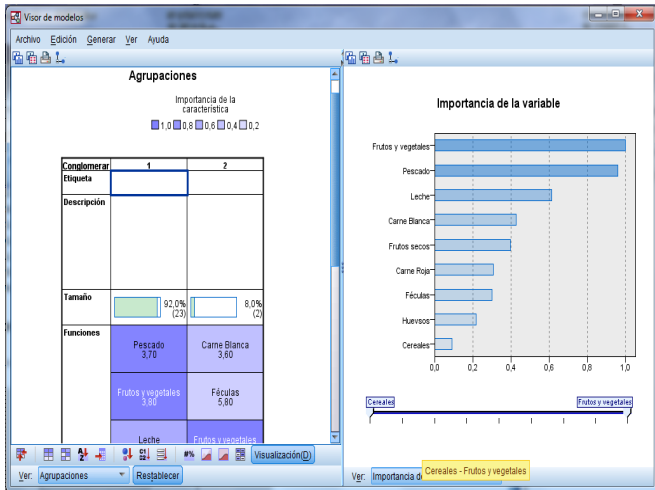
Conglomerar Etiqueta	1	2
Descripción		
Tamaño	 92,0% (23)	 8,0% (2)
Funciones	<p>Pescado 3,70</p> <p>Frutos y vegetales 3,90</p> <p>Leche 18,00</p> <p>Carne Blanca 8,30</p> <p>Frutos secos 2,90</p> <p>Carne Roja 10,10</p> <p>Féculas 4,10</p> <p>Huevos 3,00</p> <p>Cereales 32,60</p>	<p>Carne Blanca 3,60</p> <p>Féculas 5,80</p> <p>Frutos y vegetales 7,60</p> <p>Carne Roja 6,70</p> <p>Leche 6,80</p> <p>Cereales 28,10</p> <p>Frutos secos 5,30</p> <p>Pescado 10,60</p> <p>Huevos 2,10</p>

PROCEDIMIENTO DE CLUSTER EN DOS PASOS

Comparación de conglomerados



PROCEDIMIENTO DE CLUSTER EN DOS PASOS



PROCEDIMIENTO DE CLUSTER EN DOS PASOS

