Análisis Cluster en SPSS

M. Dolores Martínez Miranda

Profesora del Dpto. Estadística e I.O.

Universidad de Granada

Referencias bibliográficas

1. **Hair, J.F., Anderson, R.E., Tatham, R.L. y Black, W.C. (1999) Análisis Multivariante (5ª edición). Ed. Prentice Hall.**
2. **Pérez, C. (2001) Técnicas estadísticas con SPSS. Ed. Prentice Hall.**

Motivación: Necesidad de diseñar una estrategia que permita definir grupos de objetos homogéneos. Tarea de clasificación.

Aplicabilidad: Psicología, biología, sociología, economía, ingeniería, investigación de mercados, etc.

Análisis Cluster: Técnica multivariante cuyo principal propósito es agrupar objetos formando conglomerados (*clusters*) de objetos con un alto grado de homogeneidad interna y heterogeneidad externa.

- **Similitud con el Análisis Factorial: Mientras que el análisis cluster agrupa objetos, el análisis factorial se centra en la agrupación de variables.**
- **Inconvenientes de Análisis Cluster: Descriptivo, *ateórico* y no inferencial, se utiliza habitualmente como una técnica exploratoria. No ofrece soluciones únicas, a pesar de que existiera una estructura de clasificación "verdadera" en los datos, las soluciones dependen de las variables consideradas y del método de análisis cluster empleado.**

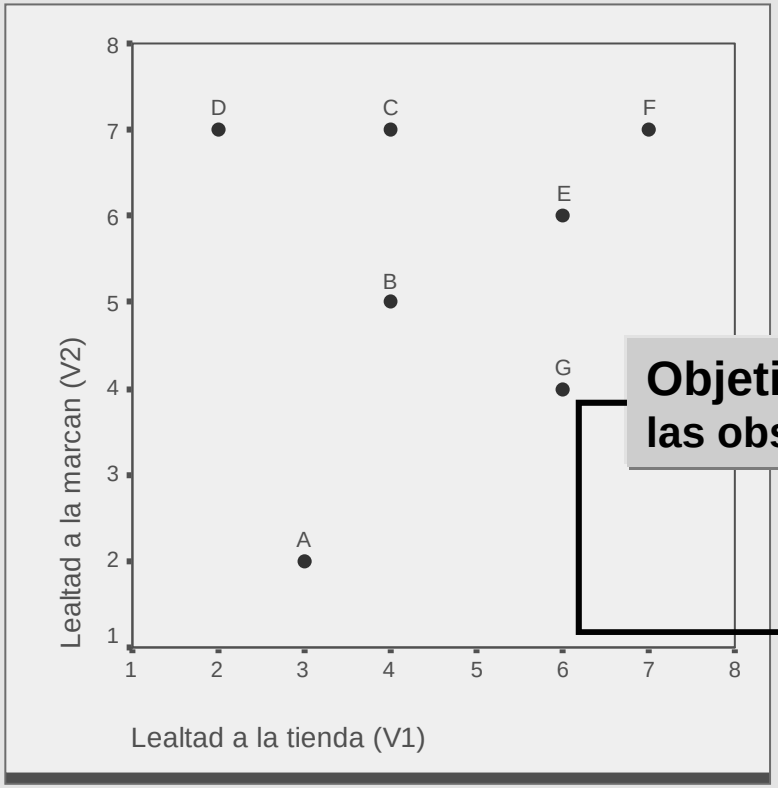
Funcionamiento del Análisis Cluster: Un ejemplo ilustrativo

Planteamiento *Determinar los segmentos de mercado en una comunidad reducida basándose en pautas de lealtad a marcas y tiendas.*

Datos Se selecciona una muestra de 7 encuestados sobre los que se miden dos variables:

V_1 (lealtad a la tienda) Escala de 0 a 10

V_2 (lealtad a la marca) Escala de 0 a 10



Encuestado / Variable	A	B	C	D	E	F	G
V_1	3	4	4	2	6	7	6
V_2	2	5	7	7	6	7	4

Objetivo: Definir la estructura de los datos colocando las observaciones más parecidas en grupos

1. **¿Cómo medimos la similitud?**
2. **¿Cómo formamos los conglomerados?**
3. **¿Cuántos formamos?**

1. Medición de la similitud

Distancia euclídea para cada par de observaciones

Matriz de distancias

Caso	distancia euclídea						
	A	B	C	D	E	F	G
A	,000	3,162	5,099	5,099	5,000	6,403	3,606
B	3,162	,000	2,000	2,828	2,236	3,606	2,236
C	5,099	2,000	,000	2,000	2,236	3,000	3,606
D	5,099	2,828	2,000	,000	4,123	5,000	5,000
E	5,000	2,236	2,236	4,123	,000	1,414	2,000
F	6,403	3,606	3,000	5,000	1,414	,000	3,162
G	3,606	2,236	3,606	5,000	2,000	3,162	,000

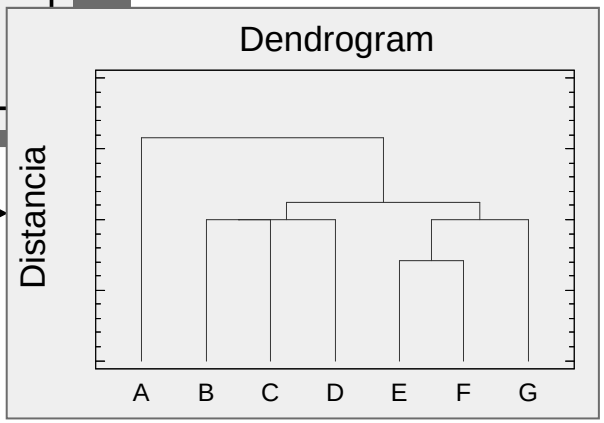
2. Formación de los conglomerados

Procedimiento jerárquico (método aglomerativo, *vecino más próximo*)

Historial de conglomeración

Etapa	Conglomerado que se combina		Coeficientes	Etapa en la que el conglomerado aparece por primera vez		Próxima etapa
	Conglomerado 1	Conglomerado 2		Conglomerado 1	Conglomerado 2	
1	E	F	1,414	0	0	2
2	E-F	G	2,000	1	0	5
3	C	D	2,000	0	0	4
4	B	C-D	2,000	0	3	5
5	B-C-D	E-F-G	2,236	4	2	6
6	A	B-C-D-E-F-G	3,162	0	5	0

Dendograma



3. Determinación del número de conglomerados en la solución final

Historial de conglomeración

Etapa	Conglomerado que se combina		Coeficientes	Etapa Congl
	Conglomerado 1	Conglomerado 2		
1	E	F	1,414	
2	E-F	G	2,000	
3	C	D	2,000	
4	B	C-D	2,000	
5	B-C-D	E-F-G	2,236	
6	A	B-C-D-E-F-G	3,162	

Medida de similitud intraconglomerado (Promedio)

1,414
2,192
2,144
2,234
2,896
3,420

Distancias euclídeas

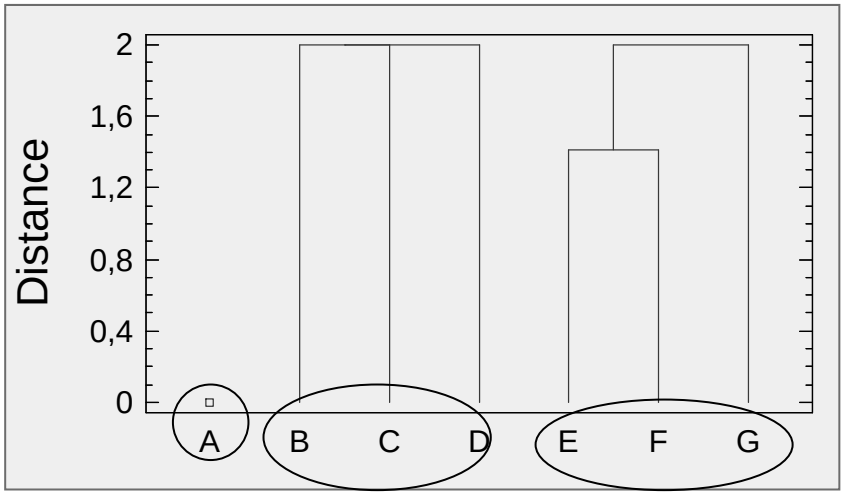
E-F (1,414)

E-G (2,000)

F-G (3,162)

Aumento mayor

La solución más apropiada parece la que ofrece el paso 4



Proceso de decisión en el Análisis Cluster

PASO 1. Objetivos del análisis

PASO 2. Diseño de investigación mediante análisis cluster

PASO 3. Supuestos del análisis cluster

PASO 4. Obtención de conglomerados y valoración del ajuste conjunto

PASO 5. Interpretación de los conglomerados

PASO 6. Validación y perfil de los grupos

PASO 1. Objetivos del análisis

1. **Descripción de una taxonomía** (una clasificación de objetos realizada empíricamente). Uso exploratorio o confirmatorio.
2. **Simplificación de los datos.** La estructura resultante permite simplificar el conjunto de observaciones.
3. **Identificación de la relación** entre las observaciones (relaciones que a priori están ocultas).

Un problema a resolver: Selección de variables del análisis cluster

La clasificación dependerá de las variables elegidas

Introducir variables irrelevantes aumenta la posibilidad de errores.

Criterio de selección:

- **Sólo aquellas variables que caracterizan los objetos que se van agrupando, y referentes a los objetivos del análisis cluster que se va a realizar**
- **Se puede realizar un ACP previamente y resumir el conjunto de variables.**

PASO 2. Diseño de la investigación mediante Análisis Cluster

1. Detección de atípicos y posible exclusión. El análisis cluster es muy sensible a la presencia de objetos muy diferentes del resto (***atípicos***).

- Métodos gráficos (caso univariante o bivariante): **Diagramas de perfil.**
- **Distancia de Mahalanobis.**

2. Medidas de similitud entre objetos.

*La **similitud** entre objetos es una medida de correspondencia, o parecido, entre objetos que van a ser agrupados.*

- **Medidas de correlación** } **Datos métricos**
- **Medidas de distancia** }
- **Medidas de asociación** **(Datos no métricos)**

3. Tipificación de los datos. El orden de las similitudes puede cambiar profundamente con sólo un cambio en la escala de una de las variables. Sólo se tipificará cuando resulte necesario.

Distancias y Similaridades en SPSS

Dependiendo de tipo de datos se elegirá la distancia o similaridad adecuada:

- **Datos de intervalo: Distancia euclídea, Distancia euclídea al cuadrado, Coseno, Correlación de Pearson, Chebychev, Bloque, Minkowski y Personalizada.**

- **Datos de frecuencias: Medida de chi-cuadrado y Medida de phi-cuadrado.**

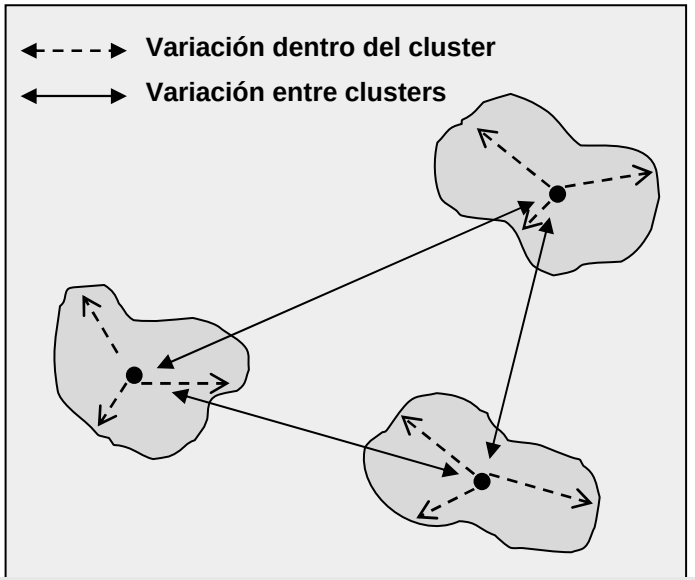
- **Datos binarios: Distancia euclídea, Distancia euclídea al cuadrado, Diferencia de tamaño, Diferencia de configuración, Varianza, Dispersión, Forma, Concordancia simple, Correlación phi de 4 puntos, Lambda, D de Anderberg, Dice, Hamann, Jaccard, Kulczynski 1, Kulczynski 2, Lance y Williams, Ochiai, Rogers y Tanimoto, Russel y Rao, Sokal y Sneath 1, Sokal y Sneath 2, Sokal y Sneath 3, Sokal y Sneath 4, Sokal y Sneath 5, Y de Yule y Q de Yule.**

PASO 3. Supuestos del Análisis Cluster

- ✓ **Representatividad de la muestra.** La bondad de los resultados depende directamente de la calidad (representatividad) de los datos considerados.
- ✓ **Multicolinealidad**. Las variables que son multicolineales están implícitamente ponderadas con más fuerza.

PASO 4. Obtención de los cluster y valoración del ajuste conjunto

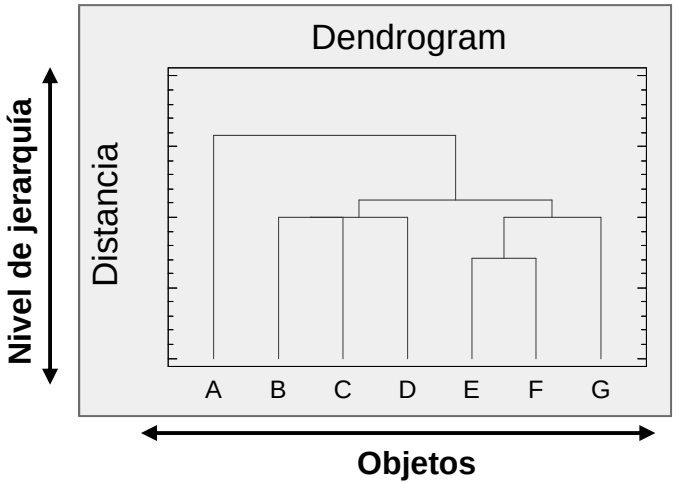
1. Algoritmo para la obtención de los cluster.
   - ➢ Procedimientos jerárquicos
   - ➢ Procedimientos no jerárquicos
2. Número de cluster: **Regla de parada.**
3. Adecuación del modelo. (Comprobar que el modelo no ha definido cluster con un solo objeto, o de tamaños muy desiguales, etc.)



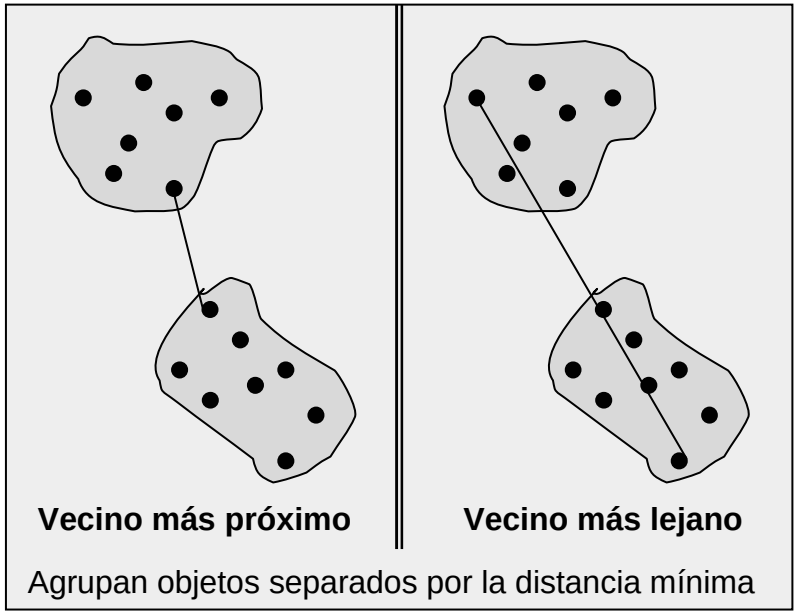
Objetivo: Maximizar las diferencias entre los cluster relativa a la variación dentro de los mismos.

Métodos jerárquicos en SPSS (aglomerativos)

- **Vinculación intergrupos**
- **Vinculación intragrupos**
- **Vecino más próximo (Encadenamiento simple)**
- **Vecino más lejano (Encadenamiento completo)**
- **Agrupación de centroides**
- **Vinculación de medianas**
- **Método de Ward**



Idea común: Construir una estructura en forma de árbol



Métodos no jerárquicos en SPSS: Método de las K medias

Características:

- **No implican la construcción de una estructura en árbol.**
- **Los objetos se asignan a los clusters una vez que se ha decidido cuántos se van a formar.**

Procedimiento general:

Primero se selecciona una semilla para cada conglomerado a formar (el primer individuo que constituirá cada conglomerado) y se van asignando nuevos objetos según el criterio de distancia considerado, buscando que la variabilidad dentro de los conglomerados sea lo menor posible, y la variabilidad entre grupos lo mayor posible (la minimización de una función objetivo).

Nota: Es común utilizar una combinación de los métodos no jerárquicos con los jerárquicos (constituirían el paso previo)

PASO 5. Interpretación de los conglomerados

Asignar una etiqueta precisa que describa la naturaleza de los cluster formados.

Herramientas:

1. **Examen de los centroides (sobre datos no tipificados, y sólo si no provienen de una reducción mediante ACP).**
2. **Si el objetivo del análisis era confirmatorio, contrastar la clasificación con los datos preconcebidos.**

PASO 6. Validación y perfil de los grupos

Confirmar que la solución es representativa de la población general.

Herramientas:

1. **Correlación cofenética** *(correlación entre las distancias iniciales y las finales)*
2. **Estabilidad de la solución desde distintos procedimientos dentro del análisis cluster.**

Un ejemplo ilustrativo: Análisis Cluster en SPSS

Fichero de datos: "jovenes.sav"

PROCESO DE DECISIÓN

Paso 1. Objetivos del análisis cluster:

Se trata de clasificar al conjunto de los 14 jóvenes encuestados por el número de veces que van anualmente al fútbol (*futbol*), la paga semanal que reciben (*paga*) y el número de horas semanales que ven la televisión (*tv*).

Paso 2. Diseño de investigación del análisis cluster:

- **Elegir la medida de distancia o similitud: Distancia euclídea al cuadrado (se trata de datos métricos en escala de intervalo)**
- **Se deberán tipificar los datos dado que las variables se miden en distinta escala (tipificación de cada variable, *puntuaciones z*).**
- **Puede ser de utilidad representar gráficamente los datos para ver si se percibe alguna forma de clasificación (Diagrama de dispersión tridimensional de las variables identificando los individuos con la variable *id*)**

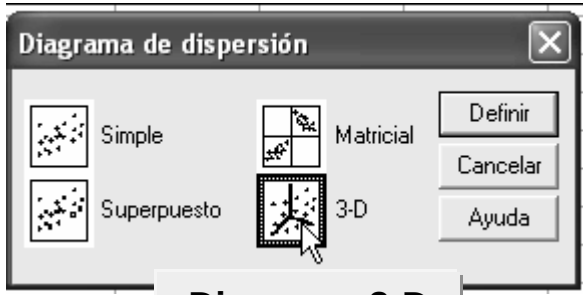
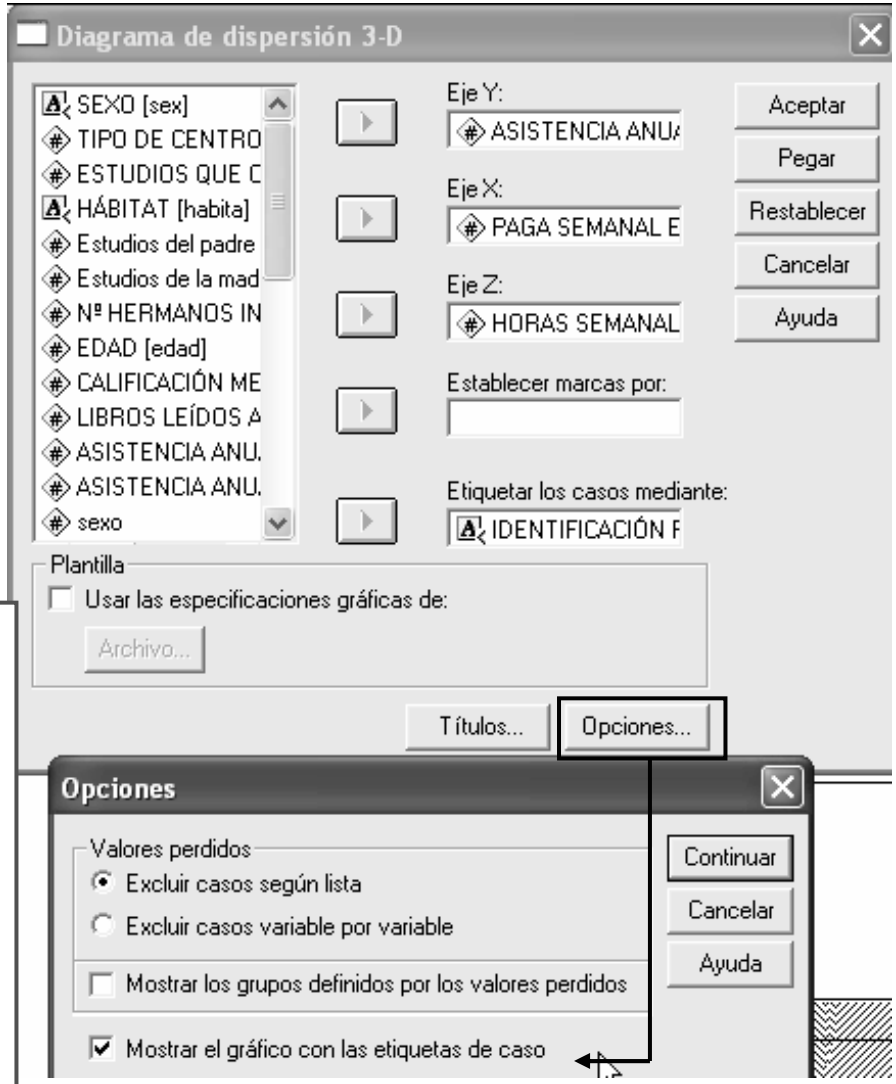


Diagrama 3-D

Eje Y: futbol

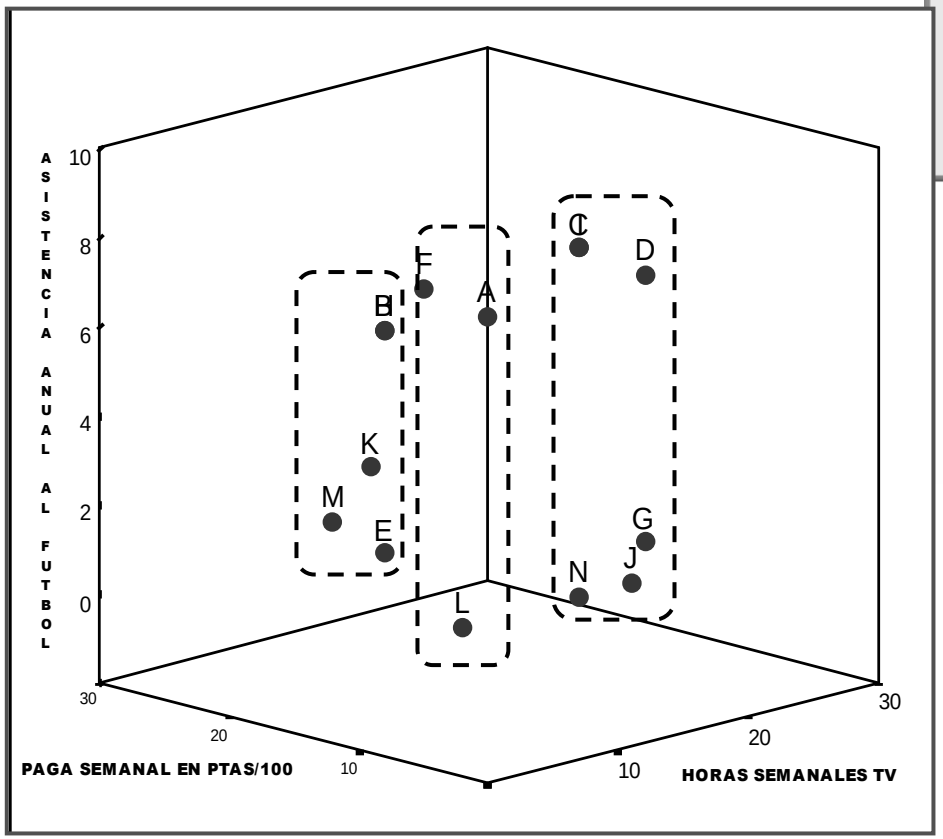
Eje X: tv

Eje Z: paga



Opciones:

✓ Mostrar el gráfico con las etiquetas de caso



Paso 3. Supuestos del análisis cluster:

- **¿La muestra se considera representativa? Es muy pequeña (n=14), tan sólo nos servirá como ilustración.**
- **Estudio de atípicos y valores aislados (medidas de influencia y distancia de Mahalanobis).**

Paso 4. Obtención de los clusters y valoración del ajuste conjunto:

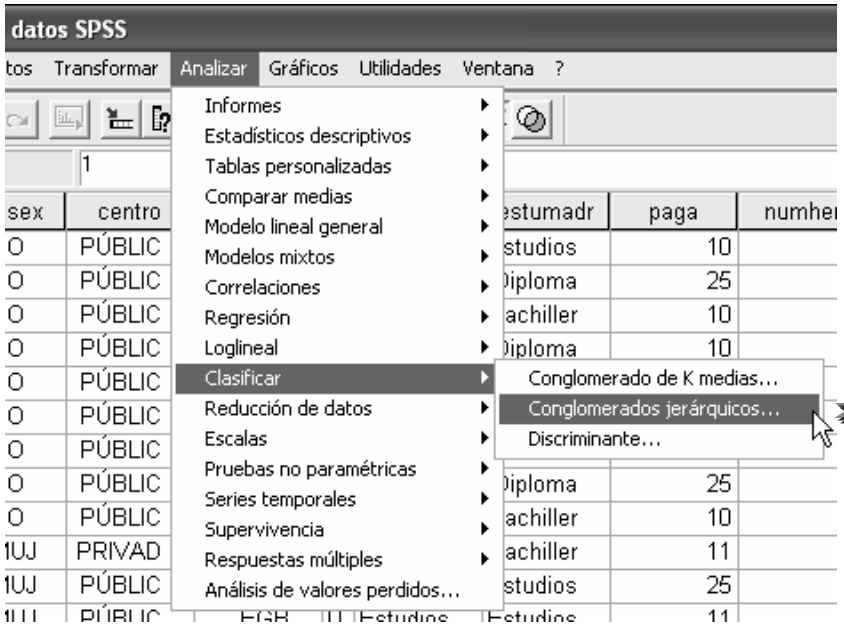
A. Análisis cluster jerárquico:

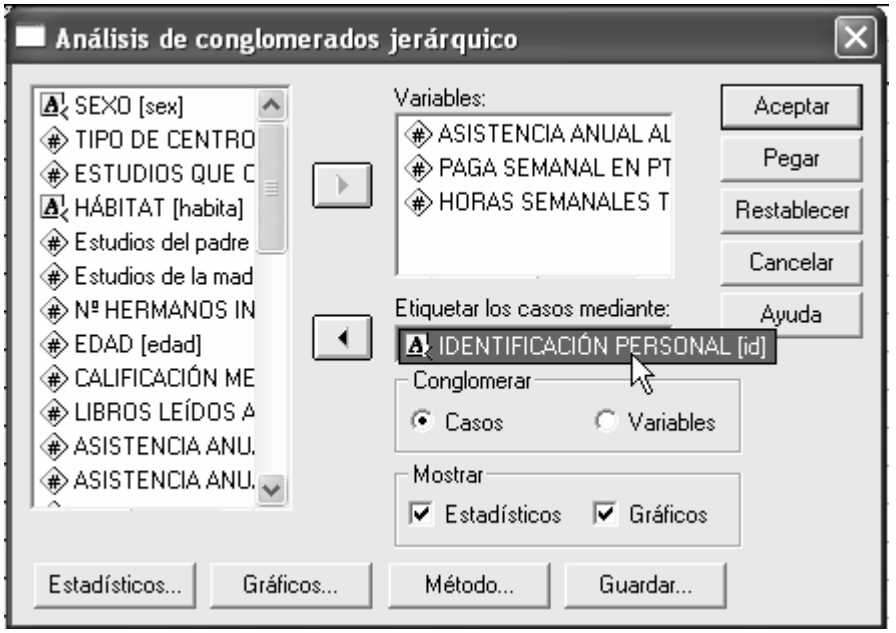
i. **Encadenamiento completo**
ii. **Método de Ward**
iii. **Vinculación de medianas**

} **Verificar la estabilidad de las tres soluciones cluster obtenidas**

B. Análisis cluster no jerárquico: Fijaremos previamente el número de cluster a formar

(A.i.) Análisis cluster jerárquico mediante encadenamiento completo





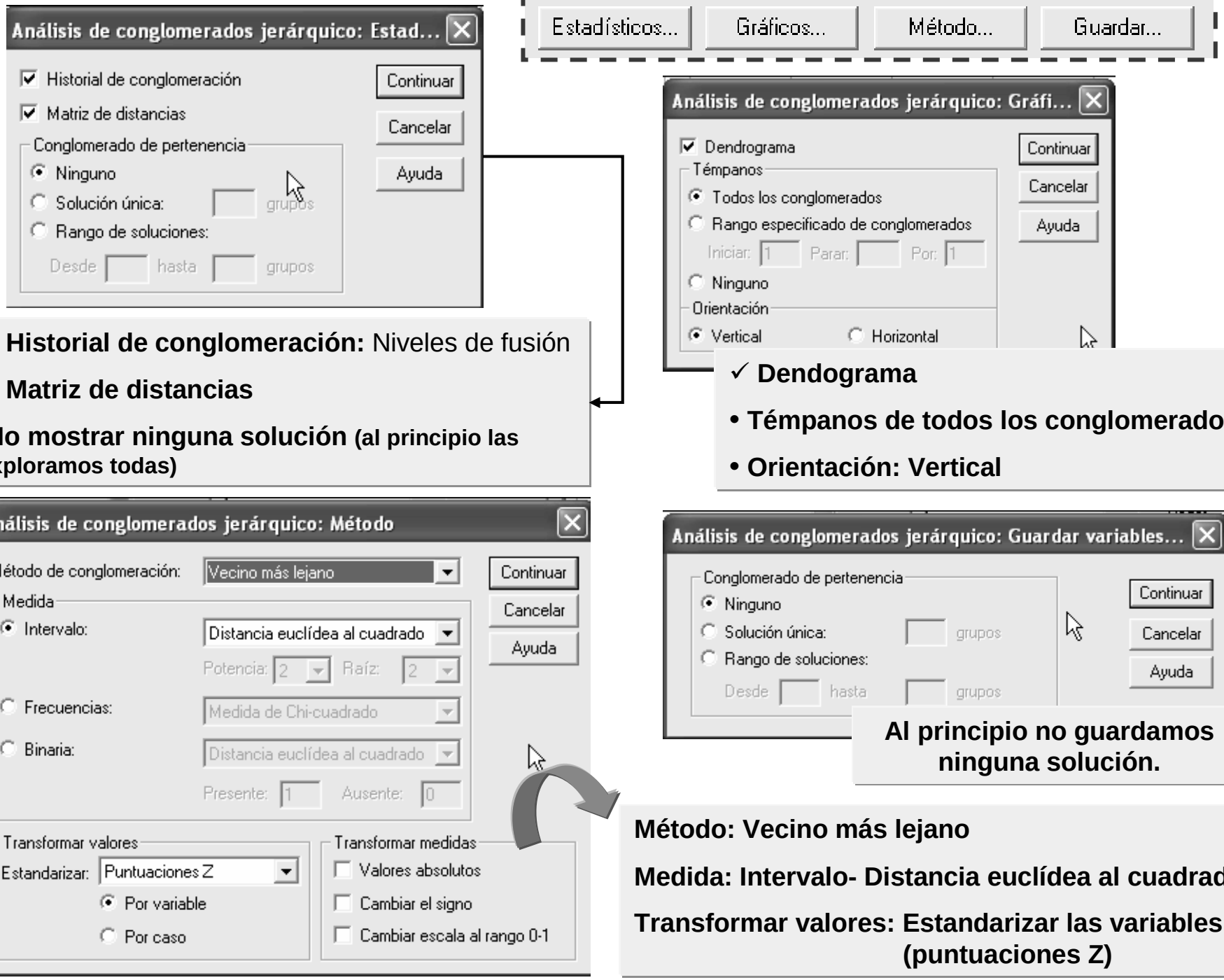
Variables: Introducimos las variables en las que se basará la clasificación (futbol, paga, tv)

Etiquetar los casos mediante: Si tenemos alguna variable de cadena que permita la identificación de los casos (id)

Conglomerar:

- **Casos (clasificar individuos)**
- **Variables (clasificar variables, similitud con análisis factorial)**

Mostrar: Estadísticos (resultados de texto) y/o Gráficos


Estadísticos...
Gráficos...
Método...
Guardar...
Análisis de conglomerados jerárquico: Estad...
Historial de conglomeración
Matriz de distancias
Conglomerado de pertenencia
Ninguno
Solución única:
grupos
Rango de soluciones:
Desde
hasta
grupos
Continuar
Cancelar
Ayuda
✓ Historial de conglomeración: Niveles de fusión
✓ Matriz de distancias
No mostrar ninguna solución (al principio las exploramos todas)
Análisis de conglomerados jerárquico: Gráfi...
Dendrograma
Témpanos
Todos los conglomerados
Rango especificado de conglomerados
Iniciar:
Parar:
Por:
Ninguno
Orientación
Vertical
Horizontal
✓ Dendograma
• Témpanos de todos los conglomerados
• Orientación: Vertical
Análisis de conglomerados jerárquico: Método
Método de conglomeración:
Vecino más lejano
Medida
Intervalo:
Distancia euclídea al cuadrado
Potencia:
Raíz:
Frecuencias:
Medida de Chi-cuadrado
Binaria:
Distancia euclídea al cuadrado
Presente:
Ausente:
Transformar valores
Estandarizar:
Puntuaciones Z
Por variable
Por caso
Transformar medidas
Valores absolutos
Cambiar el signo
Cambiar escala al rango 0-1
Método: Vecino más lejano
Medida: Intervalo- Distancia euclídea al cuadrado
Transformar valores: Estandarizar las variables (puntuaciones Z)
Análisis de conglomerados jerárquico: Guardar variables...
Conglomerado de pertenencia
Ninguno
Solución única:
grupos
Rango de soluciones:
Desde
hasta
grupos
Al principio no guardamos ninguna solución.

- **La matriz de las distancias (euclídeas al cuadrado) entre cada dos individuos del fichero de datos antes de comenzar la clasificación.**

Matriz de distancias

Caso	distancia euclídea al cuadrado													
	1: A	2: B	3: C	4: D	5: E	6: F	7: G	8: H	9: I	10: J	11: K	12: L	13: M	14: N
1: A	,000	6,479	2,005	5,643	10,306	1,065	8,705	6,479	2,005	9,829	7,755	4,225	7,633	6,694
2: B	6,479	,000	4,984	5,538	2,126	10,627	6,559	,000	4,984	6,780	,805	8,309	1,988	5,840
3: C	2,005	4,984	,000	1,065	9,661	5,643	5,147	4,984	,000	6,441	7,319	7,970	9,013	5,501
4: D	5,643	5,538	1,065	,000	9,365	11,411	3,062	5,538	1,065	4,186	7,755	10,809	10,454	4,813
5: E	10,306	2,126	9,661	9,365	,000	15,305	5,283	2,126	9,661	4,654	,379	6,183	,712	3,714
6: F	1,065	10,627	5,643	11,411	15,305	,000	15,493	10,627	5,643	16,788	12,022	6,089	10,894	12,085
7: G	8,705	6,559	5,147	3,062	5,283	15,493	,000	6,559	5,147	,104	5,714	6,727	7,393	,731
8: H	6,479	,000	4,984	5,538	2,126	10,627	6,559	,000	4,984	6,780	,805	8,309	1,988	5,840
9: I	2,005	4,984	,000	1,065	9,661	5,643	5,147	4,984	,000	6,441	7,319	7,970	9,013	5,501
10: J	9,829	6,780	6,441	4,186	4,654	16,788	,104	6,780	6,441	,000	5,426	6,623	6,934	,627
11: K	7,755	,805	7,319	7,755	,379	12,022	5,714	,805	7,319	5,426	,000	5,935	,438	4,171
12: L	4,225	8,309	7,970	10,809	6,183	6,089	6,727	8,309	7,970	6,623	5,935	,000	4,387	3,174
13: M	7,633	1,988	9,013	10,454	,712	10,894	7,393	1,988	9,013	6,934	,438	4,387	,000	4,739
14: N	6,694	5,840	5,501	4,813	3,714	12,085	,731	5,840	5,501	,627	4,171	3,174	4,739	,000

Esta es una matriz de disimilaridades

Los individuos que guardan menor distancia son el tercero y el noveno (0.000), serán los primeros que se unan en un mismo cluster. Los siguentes serán el segundo y el noveno (guardan aproximadamente la misma distancia, 0.000)

- **El historial de conglomeración muestra los niveles de fusión (coeficientes) al que se van uniendo los individuos en los conglomerados.**

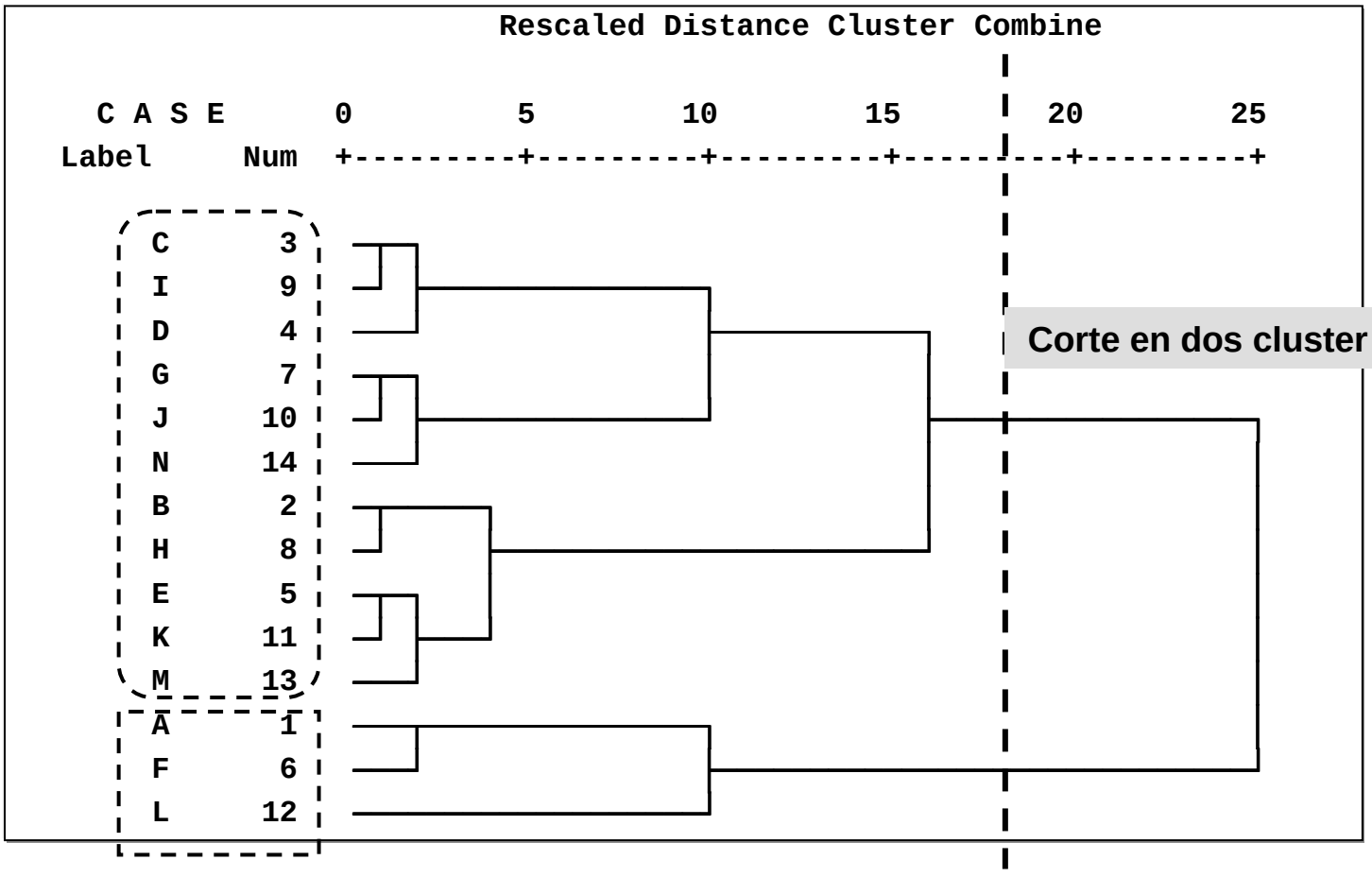
Historial de conglomeración

Etapa	Conglomerado que se combina		Coeficientes	Etapa en la que el conglomerado aparece por primera vez		Próxima etapa
	Conglomerado 1	Conglomerado 2		Conglomerado 1	Conglomerado 2	
1	3	9	,000	0	0	7
2	2	8	,000	0	0	9
3	7	10	,104	0	0	6
4	5	11	,379	0	0	5
5	5	13	,712	4	0	9
6	7	14	,731	3	0	11
7	3	4	1,065	1	0	11
8	1	6	1,065	0	0	10
9	2	5	2,126	2	5	12
10	1	12	6,089	8	0	13
11	3	7	6,441	7	6	12
12	2	3	10,454	9	11	13
13	1	2	16,788	10	12	0

Los coeficientes (niveles de fusión) se calculan mediante el método del vecino más lejano y empleando como distancia la euclídea al cuadrado). Podemos observar cómo va aumentando la variabilidad dentro de los conglomerados conforme vamos agradándolos. En la primera etapa habría 13 clúster y en la última etapa un cluster que engloba a los catorce jóvenes.

En la primera etapa se combinan los individuos tercero y noveno que eran los menos distantes (coeficiente=distancia=0.000). El cluster formado por estos dos individuos se modifica (se añaden individuos) en la séptima etapa)

➢ DENDOGRAMA.



Muestra cómo se va formando la clasificación jerárquica de los individuos. Por ejemplo, si se consideraran 2 cluster, la clasificación sería:

Cluster 1: N, J, G, D, I, C, M, K, E, H, B

Cluster 2: L, F, A

➢ El diagrama de témpanos.

Diagrama de témpanos vertical

Número de conglomerados	Caso																											
	14: N		10: J		7: G		4: D		9: I		3: C		13: M		11: K		5: E		8: H		2: B		12: L		6: F		1: A
1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
2	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X
3	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X		X	X	X	X	X	X
4	X	X	X	X	X		X	X	X	X	X		X	X	X	X	X	X	X	X	X		X	X	X	X	X	X
5	X	X	X	X	X		X	X	X	X	X		X	X	X	X	X	X	X	X	X		X		X	X	X	X
6	X	X	X	X	X		X	X	X	X	X		X	X	X	X	X		X	X	X		X		X	X	X	X
7	X	X	X	X	X		X	X	X	X	X		X	X	X	X	X		X	X	X		X		X		X	X
8	X	X	X	X	X		X		X	X	X		X	X	X	X	X		X	X	X		X		X		X	X
9	X		X	X	X		X		X	X	X		X	X	X	X	X		X	X	X		X		X		X	X
10	X		X	X	X		X		X	X	X		X		X	X	X		X	X	X		X		X		X	X
11	X		X	X	X		X		X	X	X		X		X		X		X	X	X		X		X		X	X
12	X		X		X		X		X	X	X		X		X		X		X	X	X		X		X		X	X
13	X		X		X		X		X	X	X		X		X		X		X		X		X		X		X	X

Muestra cómo quedaría la clasificación de individuos dependiendo del número de conglomerados que consideremos (cada fila de la tabla). Por filas, se van pintando X's y se deja un hueco cuando cambiamos de cluster. Por ejemplo, si se consideran 3 cluster, la clasificación sería:

Cluster 1: N, J, G, D, I, C

Cluster 2: M, K, E, H, B

Cluster 3: L, F, A

¿Qué número de cluster vamos a considerar?

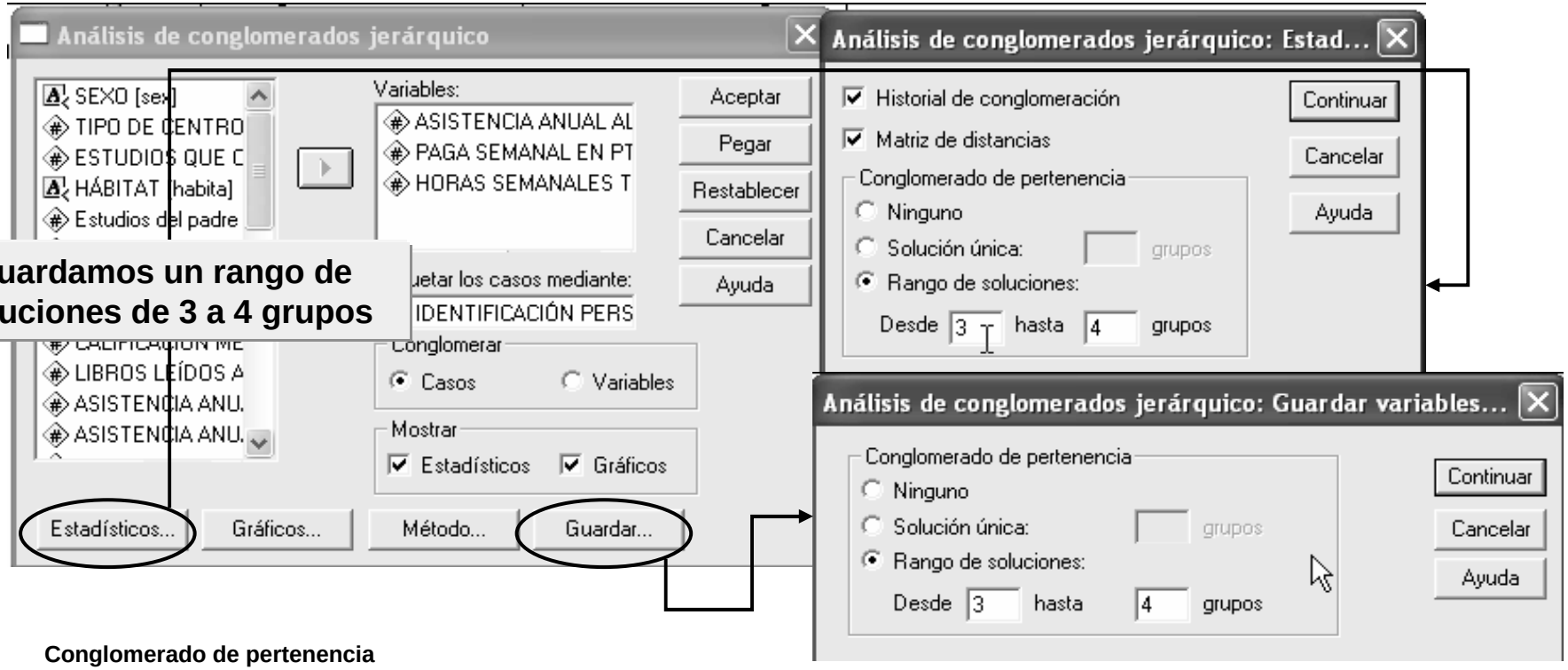
CRITERIO: Elegir el número de cluster observando los niveles de fusión, y teniendo en cuenta el diagrama de dispersión de los individuos

Historial de conglomeración

Etapa	Conglomerado que se combina		Coeficientes	Etapa en la que el conglomerado aparece por primera vez		Próxima etapa
	Conglomerado 1	Conglomerado 2		Conglomerado 1	Conglomerado 2	
1	3	9	,000	0	0	7
2	2	8	,000	0	0	9
3	7	10	,104	0	0	6
4	5	11	,379	0	0	5
5	5	13	,712	4	0	9
6	7	14	,731	3	0	11
7	3	4	1,065	1	0	11
8	1	6	1,065	0	0	10
9	2	5	2,126	2	5	12
10	1	12	6,089	8	0	13
11	3	7	6,441	7	6	12
12	2	3	10,454	9	11	13
13	1	2	16,788	10	12	0

Rango de soluciones:

3 o 4 cluster



Guardamos un rango de soluciones de 3 a 4 grupos

Conglomerado de pertenencia

Caso	4 conglomerados	3 conglomerados
1: A	1	1
2: B	2	2
3: C	3	3
4: D	3	3
5: E	2	2
6: F	1	1
7: G	4	3
8: H	2	2
9: I	3	3
10: J	4	3
11: K	2	2
12: L	1	1
13: M	2	2
14: N	4	3

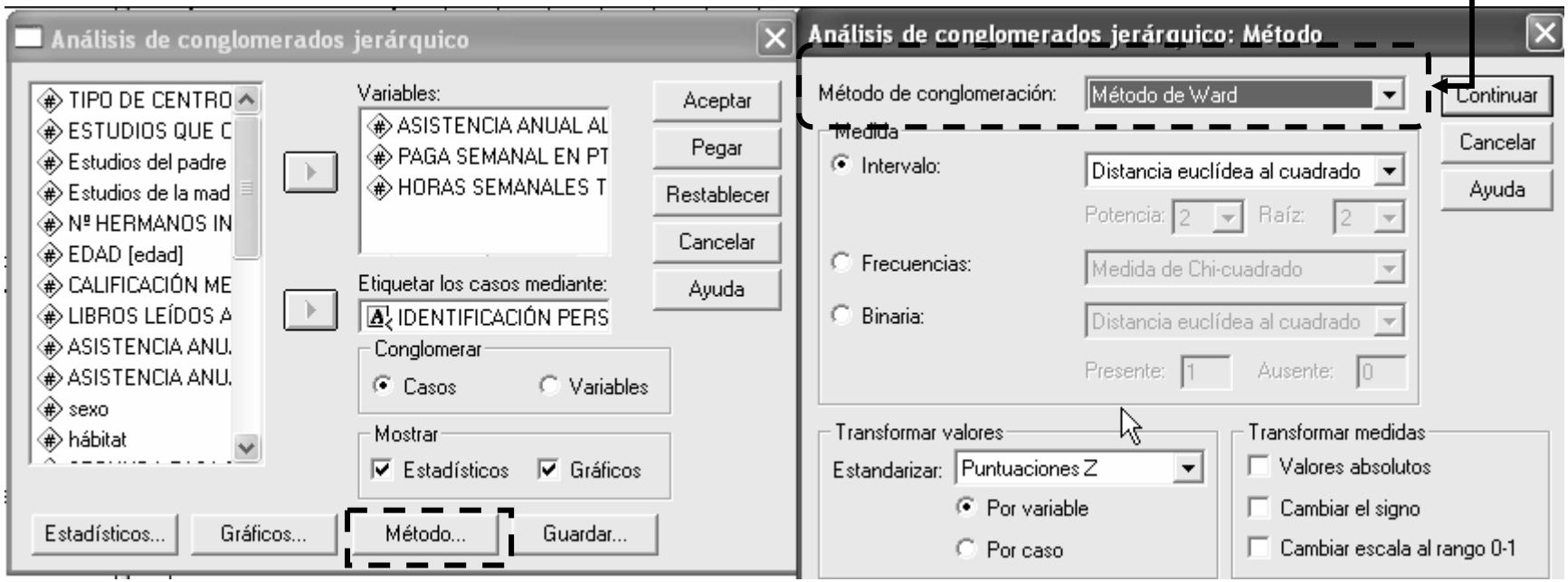
- **Aparece la clasificación de los 14 individuos para los dos casos 3 y 4 cluster elegidos**

clu4_1	clu3_1
1	1
2	2
3	3
3	3
2	2
1	1
4	3
2	2
3	3
4	3
2	2
1	1
2	2
4	3

- **Se añaden dos variables al fichero de datos:**
  - **clu4_1: define 4 grupos**
  - **clu3_1: define 3 grupos**

(A.ii.) Análisis cluster jerárquico mediante el método de WARD

Se sigue el mismo proceso anterior, tan sólo habría que elegir como método de conglomeración: Método de Ward



Historial de conglomeración

Etapa	Conglomerado que se combina		Coeficientes	Etapa en la que el conglomerado aparece por primera vez		Próxima etapa
	Conglomerado 1	Conglomerado 2		Conglomerado 1	Conglomerado 2	
1	3	9	,000	0	0	8
2	2	8	,000	0	0	9
3	7	10	,052	0	0	6
4	5	11	,242	0	0	5
5	5	13	,562	4	0	9
6	7	14	,997	3	0	11
7	1	6	1,529	0	0	10
8	3	4	2,239	1	0	11
9	2	5	4,003	2	5	13
10	1	12	7,263	7	0	12
11	3	7	14,371	8	6	12
12	1	3	25,210	10	11	13
13	1	2	39,000	12	9	0

Conglomerado de pertenencia

Caso	4 conglomerados	3 conglomerados
1: A	1	1
2: B	2	2
3: C	3	3
4: D	3	3
5: E	2	2
6: F	1	1
7: G	4	3
8: H	2	2
9: I	3	3
10: J	4	3
11: K	2	2
12: L	1	1
13: M	2	2
14: N	4	3

Rango de soluciones: 3 o 4 cluster

- **Se añaden dos variables al fichero de datos:**
  - **clu4_2: define 4 grupos**
  - **clu3_2: define 3 grupos**
  - **(ahora mediante el método de Ward)**

(A.iii.) Análisis cluster jerárquico mediante el método de la mediana

Se sigue el mismo proceso anterior, tan sólo habría que elegir como método de conglomeración: Método de medianas

Historial de conglomeración

Etapa	Conglomerado que se combina: Conglomerado 1	Conglomerado que se combina: Conglomerado 2	Coeficientes	Etapa en la que el conglomerado aparece por primera vez: Conglomerado 1	Etapa en la que el conglomerado aparece por primera vez: Conglomerado 2	Próxima etapa
1	3	9	,000	0	0	7
2	2	8	,000	0	0	9
3	7	10	,104	0	0	6
4	5	11	,379	0	0	5
5	5	13	,480	4	0	9
6	7	14	,653	3	0	10
7	3	4	1,065	1	0	10
8	1	6	1,065	0	0	13
9	2	5	1,559	2	5	11
10	3	7	4,491	7	6	11
11	2	3	4,663	9	10	12
12	2	12	4,887	11	0	13
13	1	2	4,563	8	12	0

Conglomerado de pertenencia

Caso	4 conglomerados	3 conglomerados
1: A	1	1
2: B	2	2
3: C	3	2
4: D	3	2
5: E	2	2
6: F	1	1
7: G	3	2
8: H	2	2
9: I	3	2
10: J	3	2
11: K	2	2
12: L	4	3
13: M	2	2
14: N	3	2

Rango de soluciones: 3 o 4 cluster

- **Se añaden dos variables al fichero de datos:**
  - **clu4_3: define 4 grupos**
  - **clu3_3: define 3 grupos**
  - **(ahora mediante el método de viculación de medianas)**

Vecino más lejano

- **Caso de 3 cluster:**
  - **Cluster 1: A, F, L**
  - **Cluster 2: B, E, H, K, M**
  - **Cluster 3: C, D, G, I, J, N**
- **Caso de 4 cluster:**
  - **Cluster 1: A, F, L**
  - **Cluster 2: B, E, H, K, M**
  - **Cluster 3: C, D, I**
  - **Cluster 4: G, J, N**

Método de WARD

- **Caso de 3 cluster:**
  - **Cluster 1: A, F, L**
  - **Cluster 2: B, E, H, K, M**
  - **Cluster 3: C, D, G, I, J, N**
- **Caso de 4 cluster:**
  - **Cluster 1: A, F, L**
  - **Cluster 2: B, E, H, K, M**
  - **Cluster 3: C, D, I**
  - **Cluster 4: G, J, N**

Vinculación medianas

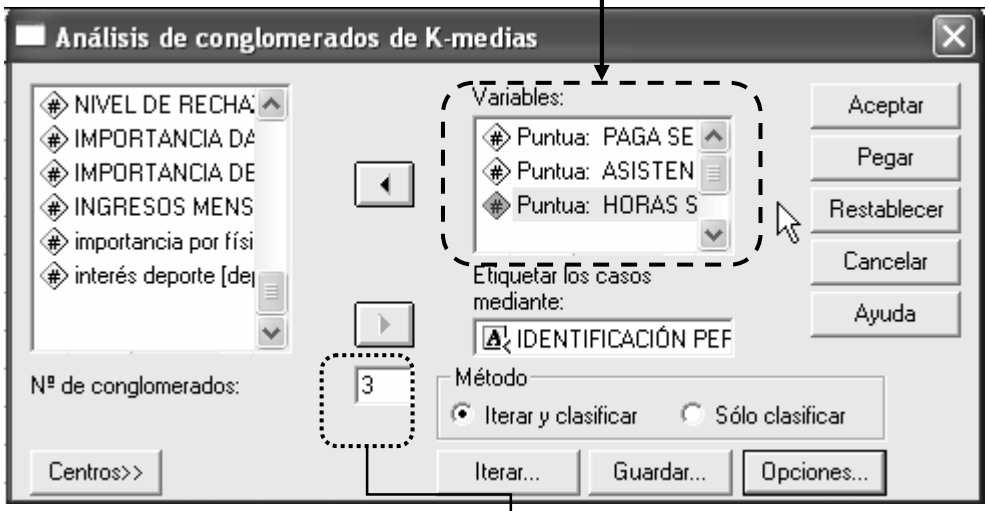
- **Caso de 3 cluster:**
  - **Cluster 1: A, F**
  - **Cluster 2: B, C, D, E, G, H, I, J, K, M, N**
  - **Cluster 3: L**
- **Caso de 4 cluster:**
  - **Cluster 1: A, F**
  - **Cluster 2: B, E, H, K, M**
  - **Cluster 3: C, D, G, I, J, N**
  - **Cluster 4: L**

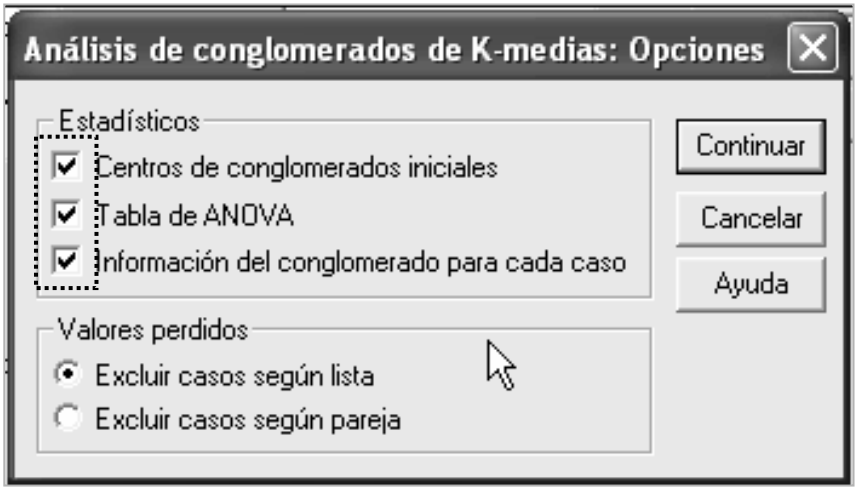
Resumen de los resultados mediante el análisis cluster jerárquico:

- **Los dos primeros métodos proporcionan resultados idénticos, el método de la mediana parece representar peor los datos observados.**
- **Nos quedamos con la solución de tres cluster y pasamos a realizar un análisis no jerárquico para compara la clasificación.**

(B) Análisis cluster no jerárquico: Método de las K medias

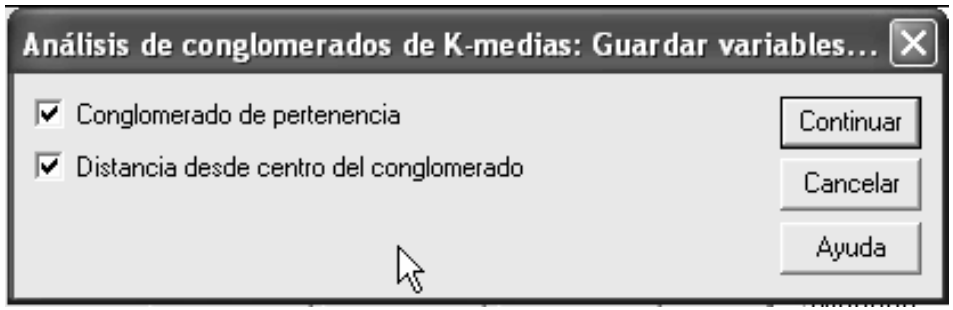
Advertencia: SPSS no permite especificar la tipificación de las variables dentro del cuadro de diálogo, por tanto pasamos las varariables tipificadas (*zpaga, zfutbol, ztv*)

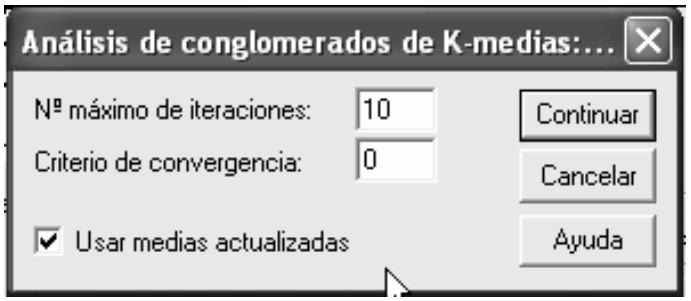




Elegimos una clasificación en 3 conglomerados

Centros>> Se podrían elegir los primeros individuos que formarán los 3 conglomerados





Iterar… Solicitamos que vaya actualizando cada vez las medias de los cluster

Centros iniciales de los conglomerados

	Conglomerado		
	1	2	3
Puntua: PAGA SEMANAL EN PTAS	-,76285	1,29097	-,76285
Puntua: ASISTENCIA ANUAL AL FUTBOL	1,24983	-,79156	,95821
Puntua: HORAS SEMANALES TV	-2,14934	-,56562	1,21607

Medias de los cluster iniciales (un individuo cada uno). Por defecto se selecciona entre los datos un número de casos debidamente espaciados igual al número de conglomerados.

Historial de iteraciones[a]

	Cambio en los centros de los conglomerados		
Iteración	1	2	3
1	,794	,629	,849
2	,198	,105	,121
3	4,961E-02	1,748E-02	1,732E-02
4	1,240E-02	2,914E-03	2,475E-03
5	3,101E-03	4,856E-04	3,536E-04
6	7,752E-04	8,094E-05	5,051E-05
7	1,938E-04	1,349E-05	7,216E-06
8	4,845E-05	2,248E-06	1,031E-06
9	1,211E-05	3,747E-07	1,473E-07
10	3,028E-06	6,245E-08	2,104E-08

a. Las iteraciones se han detenido porque se ha llevado a cabo el número máximo de iteraciones. Las iteraciones no han convergido. La distancia máxima en la que han cambiado los centros es 2,503E-06. La iteración actual es 10. La distancia mínima entre los centros iniciales es 3,233.

Historial de iteraciones: Muestra las medias (centros) de los cluster en cada paso. El método para en 10 pasos sin alcanzar el criterio de convergencia.

Pertenencia a los conglomerados

Número de caso	IDENTIFICACIÓN PERSONAL	Conglomerado	Distancia
1	A	1	,706
2	B	2	,727
3	C	3	1,281
4	D	3	,990
5	E	2	,784
6	F	1	1,058
7	G	3	,990
8	H	2	,727
9	I	3	1,281
10	J	3	1,258
11	K	2	,175
12	L	1	1,474
13	M	2	,755
14	N	3	1,216

qcl_1	qcl_2
1	,70642
2	,72736
3	1,28122
3	,99039
2	,78365
1	1,05840
3	,99039
2	,72736
3	1,28122
3	1,25795
2	,17498
1	1,47435
2	,75527
3	1,21571

Pertenencia a los conglomerados: Muestra la solución final de la clasificación en 3 grupos.

- **Además se crean dos variables en el editor de datos:**

  qcl_1 codificación que indica la pertenencia a cada cluster

  qcl_2 distancia euclídea entre cada caso y el centro del cluster utilizado para clasificar ese caso.

Resultados para valorar la solución de 3 cluster obtenida
(No tiene mucho sentido sobre los datos tipificados)

Centros de los conglomerados finales

	Conglomerado		
	1	2	3
Puntua: PAGA SEMANAL EN PTAS	-,71721	1,29097	-,71721
Puntua: ASISTENCIA ANUAL AL FUTBOL	,37495	-,32496	,08332
Puntua: HORAS SEMANALES TV	-1,55544	,02828	,75415

Distancias entre los centros de los conglomerados finales

Conglomerado	1	2	3
1		2,652	2,328
2	2,652		2,174
3	2,328	2,174	

Centros de los conglomerados finales: Valorar si difieren las medias de cada cluster en cada variable.

Distancias entre los centros de los conglomerados finales: Valorar si distan entre sí los cluster formados lo suficiente.

Número de casos en cada conglomerado: Valorar si hay muchos cluster con pocas observaciones, o si difieren mucho en tamaño

ANOVA: Para cada variable se contrasta la igualdad de medias de los cluster. No obstante los niveles de significacion no se deben interpretar salvo a nivel descriptivo ya que no aparecen corregidos.

ANOVA

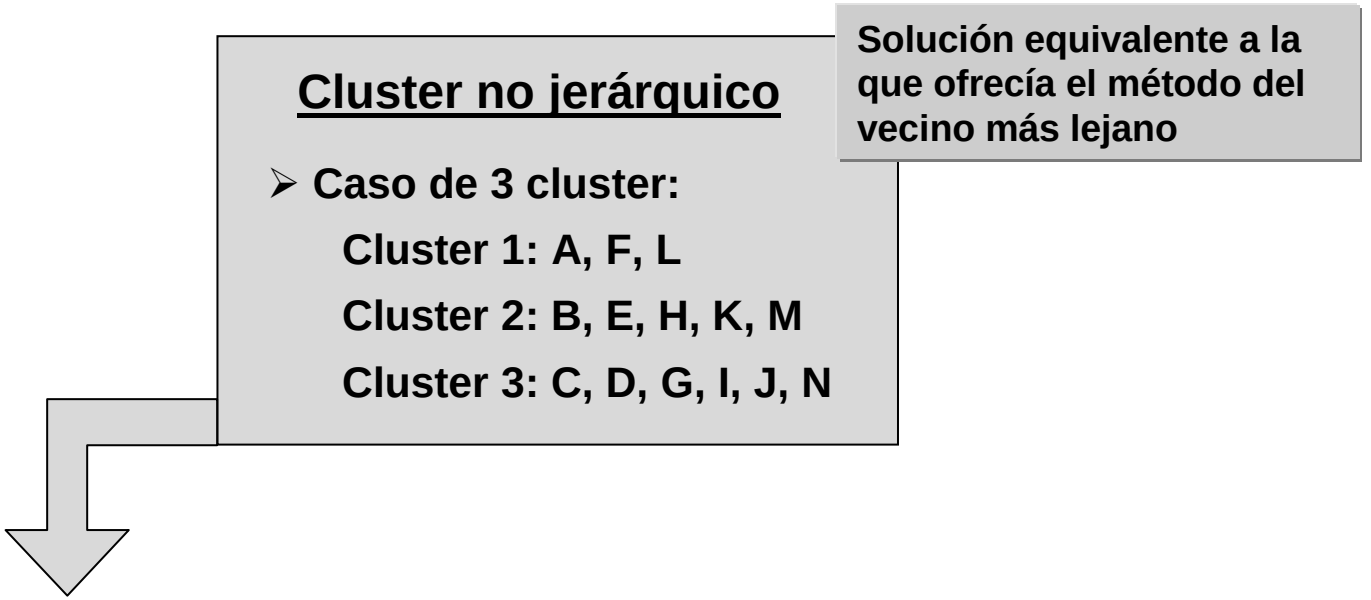
	Conglomerado		Error			
	Media cuadrática	gl	Media cuadrática	gl	F	Sig.
Puntua: PAGA SEMANAL EN PTAS	6,481	2	,003	11	1901,429	,000
Puntua: ASISTENCIA ANUAL AL FUTBOL	,496	2	1,092	11	,454	,646
Puntua: HORAS SEMANALES TV	5,337	2	,211	11	25,249	,000

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.

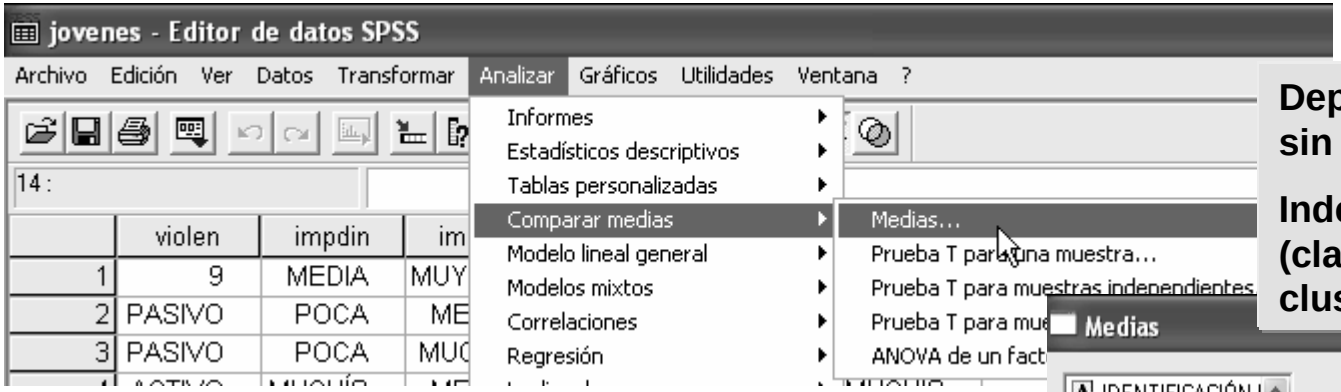
Número de casos en cada conglomerado

Conglomerado	1	3,000
	2	5,000
	3	6,000
Válidos		14,000
Perdidos		,000

Resumen de los resultados del análisis cluster no jerárquico

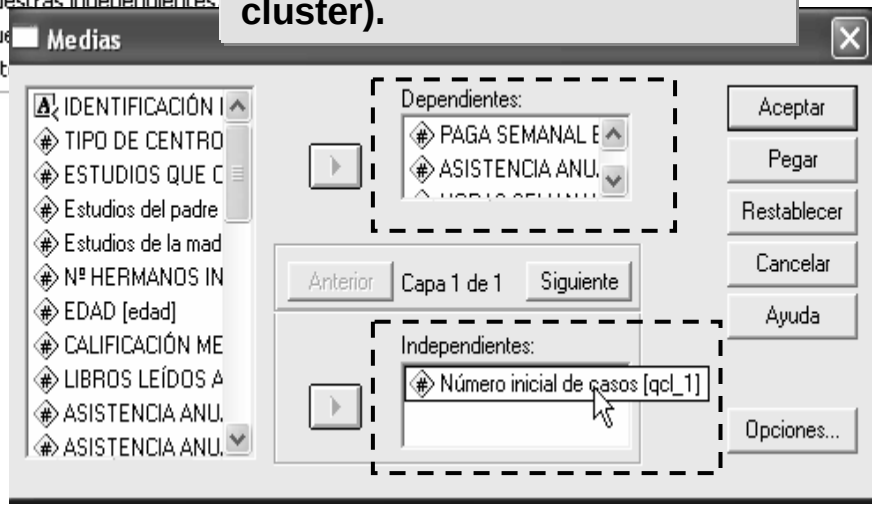


Pasos 5 y 6: Interpretaremos la solución (Nos quedamos con esta última que parece la más estable) y se discutirá la validez de los resultados.



Dependientes: Las variables sin tipificar (paga, futbol, tv)

Independientes: Variable qcl1 (clasificacion en los 3 cluster).



Informe

Número inicial de casos		PAGA SEMANAL EN PTAS/100	ASISTENCIA ANUAL AL FUTBOL	HORAS SEMANALES TV
1	Media	10,33	5,00	8,00
	N	3	3	3
	Desv. típ.	,577	4,359	2,646
2	Media	25,00	2,60	16,00
	N	5	5	5
	Desv. típ.	,000	2,302	1,732
3	Media	10,33	4,00	19,67
	N	6	6	6
	Desv. típ.	,516	4,050	2,582
Total	Media	15,57	3,71	15,86
	N	14	14	14
	Desv. típ.	7,303	3,429	5,051

Interpretación de los cluster

Observando las medias en cada variable, de los tres grupos…….Asignarle una etiqueta a cada grupo de jóvenes.