Análisis multivariable

- Las diferentes técnicas de análisis multivariante cabe agruparlas en tres categorías:

  - ***«Análisis de dependencia»*** tratan de explicar la variable considerada independiente a través de otras consideradas independientes o explicativas

  - ***«Análisis de interdependencia»*** otorgan la misma consideración a todas las variables, tienden a descubrir las interrelaciones y estructura subyacente entre ellas. Son técnicas de clasificación

  - ***«Otras técnicas»*** Intentan superar el enfoque monocriterio de las anteriores intentando explicar procesos complejos

Técnicas de análisis de dependencia

Técnica	Variable dependiente	Variables independientes
Análisis de la varianza y la covarianza	Métrica	No métrica
Análisis discriminante	No métrica	Métricas
Regresión lineal múltiple idem de variables ficticias	Métrica Métrica	Métricas No métricas
Modelos de elección discreta idem de variables ficticias	No métrica No métrica	Métricas No métricas
Análisis conjunto	Métrica o no métrica	No métricas
Segmentación jerárquica	No métrica o métrica	No métricas
Análisis de ecuaciones estructurales	Métrica	Métrica o no métrica
Análisis con clases latentes	No métrica latente	No métricas observables

Técnicas de análisis de interdependencia

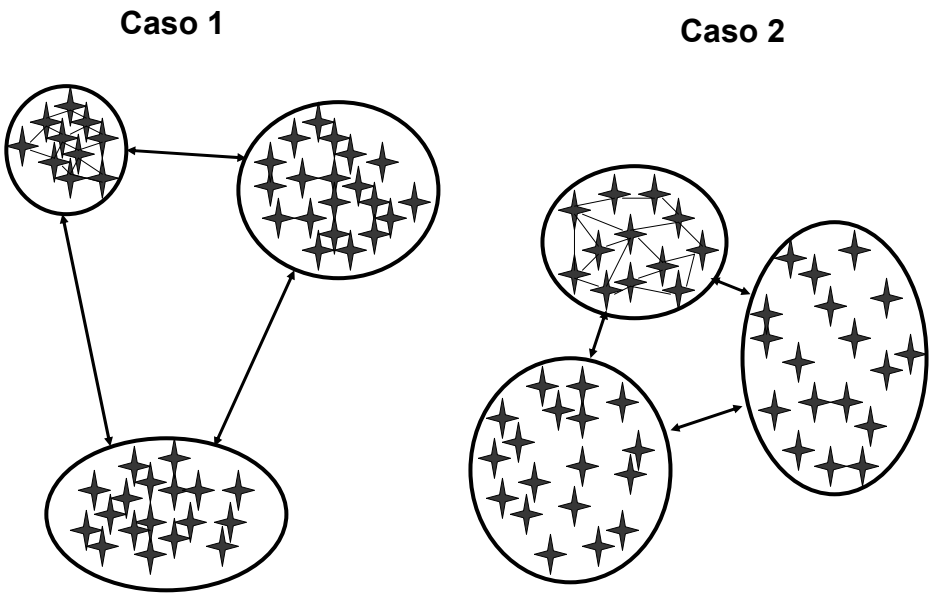
Técnica	Variables	Forma grupos de
Análisis factorial y por componentes principales	Métricas	Variables
Análisis de correspondencias	No métricas	Categorías de variables
Análisis de conglomerados	Métricas y no métricas	Objetos
Escalonamiento multidimensional	Métricas y no métricas	Objetos
Análisis con clases latentes	No métricas	Objetos y categorías de variables

La clasificación y segmentación del universo (*)

- En los mercados actuales resulta imposible satisfacer a todos los consumidores. Sus gustos y conductas son muy dispares.
  - ➡ Las organizaciones empresariales, las agencias publicitarias necesitan conocer mejor sus mercados, buscar perfiles más homogéneos para responder a sus necesidades.
  - ➡ Clasificar y segmentar los mercados posibilita responder mejor a las necesidades de los grupos.
- Tradicionalmente han existido dos formas de hacerlo:
  - ➡ ***«Segmentación a priori»*** se fija de antemano la variable dependiente (condición que nos interesa, como consumir un determinado producto), y unos criterios mínimos para los grupos (> del 10%) incluso el número de segmentos a realizar, se buscan grupos lo más homogéneos posibles con respecto a las restantes variables
  - ➡ ***«Segmentación a post hoc»*** Se desconocen las características del mercado y nos interesa encontrar grupos similares en todas las variables sin que se considere a alguna de ellas dependiente. Su número, tamaño y descripción se conocerán después. Se establecen tipologías tras realizar el análisis
  - ➡ A la *segmentación a priori* se les conoce como *«segmentación jerárquica»* y entraría dentro de las *técnicas de dependencia,* mientras que a la *segmentación post hoc* se les denomina *«Tipologías» o «análisis de conglomerados»* y son de interdependencia

LÉVIN, J-P y VALETA, J. [Director] (2003): *Análisis Multivariante para Ciencias Sociales*, Pearson, Madrid.

La Segmentación, finalidad

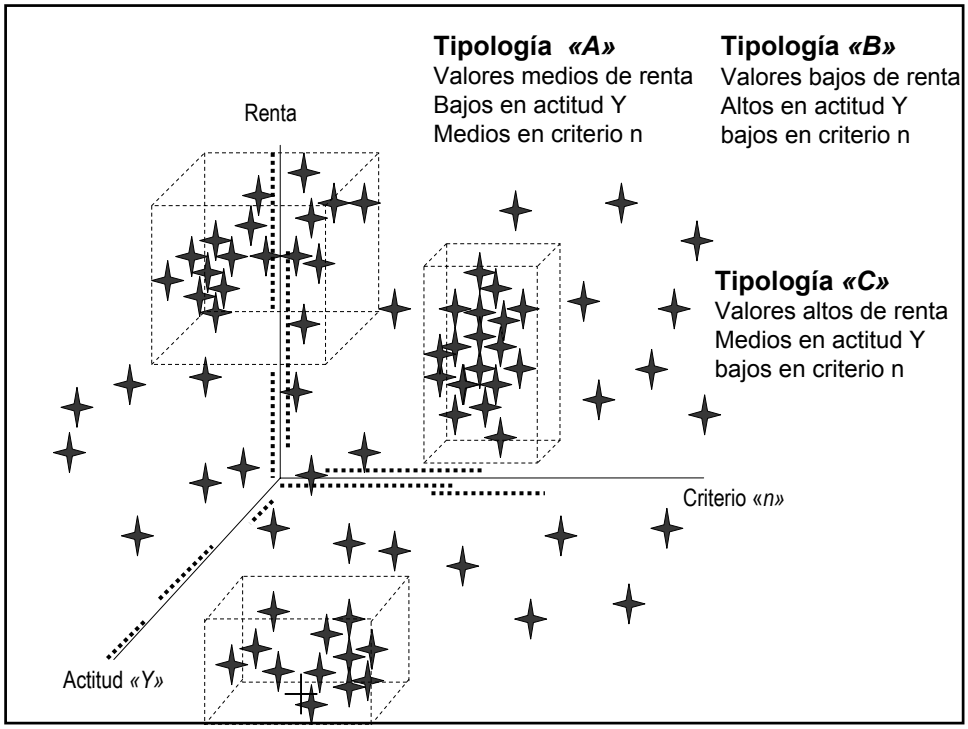


Logro de una buena segmentación (a priori)

- El sentido de la segmentación es siempre establecer segmentos cuyos elementos sean lo más homogéneos posible internamente y lo más heterogéneos entre dichos segmentos
- Se valora positivamente una segmentación cuando
  - ⇒ Cuando los segmentos son lo suficientemente cuantiosos, a medida que tienen mayor número de elementos son más interesantes
  - ⇒ El número resultante de segmentos no es elevado, lo ideal es que se acerque a la unidad lo más posible. Su número viene condicionado por:
    1. El grado de explicación o significación de cada uno de los segmentos
    2. Los objetivos y propósitos de la segmentación

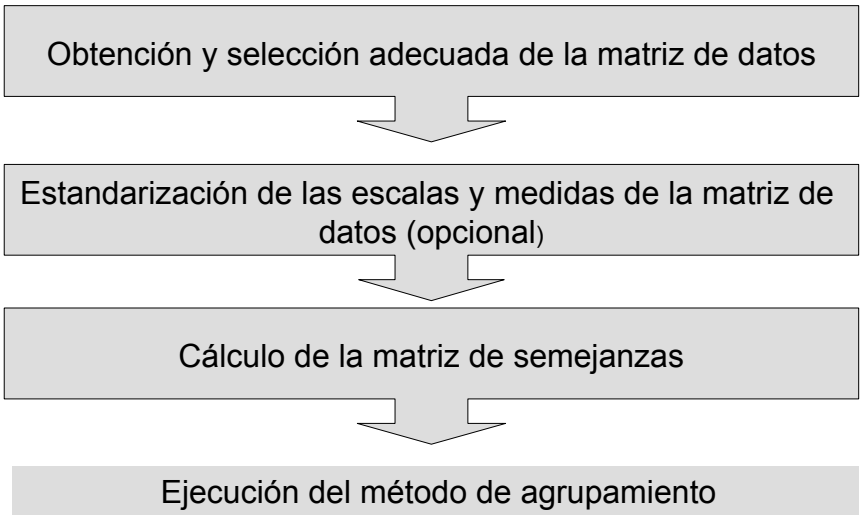
Segmentación post hoc. *Análisis de conglomerados*

- La tipología, igual que la segmentación, tiene como finalidad la constitución de grupos homogéneos que identifiquen y describan la población; pero presenta con aquella las siguientes diferencias:
  - La segmentación busca grupos homogéneos respecto de una variable que se considera dependiente a explicar (por lo general ser comprador, consumidor, simpatizante de algo o alguien). La tipología busca tipos con características similares y comunes, por lo general, en más de un ámbito.
  - En la segmentación las variables o criterios utilizados se suponen explican una variable, mientras que la tipología no distingue entre ellas dando un tratamiento igual a todas ellas.
  - El proceso seguido en la segmentación es descendente, parte a la población en grupos cada vez más reducidos. En la tipología se obra a la inversa, se parte de categorías y segmentos de la población y se buscan tipos mediante la agrupación de aquellos que poseen características similares o menos dispares
- La tipología es una técnica de obtención de grupos que quiere explicar en función de su similitud y comportamiento en varios criterios



Segmentación Post hoc. Análisis de conglomerados

- Para realizar un buen análisis de conglomerados y obtener resultados congruentes en los grupos o tipologías finales es necesario:



Obtención y selección de la matriz de datos

- Los datos básicos de los que se parten son un conjunto de variables en las que se recogen las puntuaciones en cada una de ellas de n sujetos (edad, nivel de estudios, valores, etc.)
  - ➡ La primera pregunta que debemos hacernos es si hemos seleccionado las variables correctas y si estas son relevantes teóricamente para buscar una tipología social concreta. En otras palabras, si son relevantes para la agrupación que buscamos
  - ➡ Para estudiar actitudes ante el voto puede que los estudios, el sexo, ingresos y religión sean más interesantes que el tipo de hábitat donde vive. Este último puede serlo en otros casos
- Otra cuestión a valorar es el número de variables a tener en cuenta en el análisis. Muchas variables complican el cálculo de las similitudes muy pocas aportan a la solución final poco conocimiento nuevo del que obtendríamos con otras técnicas.
  - ➡ Una solución es aplicar primero un análisis de componentes principales y escoger los factores que expliquen mayor varianza

Estandarización de la matriz de datos

- Normalmente nos encontramos con variables medidas en diferentes escalas y unidades. Su estandarización convierte las medidas en unidades adimensionales y, por tanto, comparables
  - ➡ Variables con una misma escala (razón) pueden tener ámbitos muy distintos de medida. Ejemplo: Ingresos en miles de pesetas y edad en años. En estos casos conviene la estandarización.
  - ➡ No siempre la estandarización es necesaria, a veces los resultados obtenidos sin estandarizar son idénticos a los que se obtienen estandarizando.
  - ➡ La estandarización nos permite dar a todas las variables el mismo peso, pero puede que en nuestro caso no todas tengan la misma importancia.
  - ➡ El problema de las escalas siempre es solucionable por conversión en otras escalas
- En cualquier caso estandarizar o no es una decisión que el investigador debe tomar en cada caso, no hay reglas fijas.

Cálculo de la matriz de semejanzas

- La tipología supone la búsqueda, en una población determinada, de una estructura latente, es decir, de que en su seno existen grupos con un cierto comportamiento homogéneo en torno a unos criterios
- Para localizar estas homogeneidades internas a los grupos y las disimilitudes entre estos, se recurre a las medidas estadísticas de distancia:
  - ➡ ***«Distancia cuadrática»*** Para dos sujetos cualquiera sería igual a la suma al cuadrado de las diferencias entre los valores alcanzados por estos en los *«n»* criterios o variables considerados. Es más adecuada cuando las variables son de intervalo o razón.

$$d^2_{(1,2)} = \sum_{i=1}^{n} (y_i^1 - y_i^2)^2$$

Ejemplo: Supongamos los dos siguientes sujetos:
S^1=[70,102, 43, 7]
S^2=[68, 99, 42, 5]
La distancia cuadrática $d^2 = (2+3+1+2)^2 = 8^2 = 64$; d=8

  - ➡ ***«Distancia informática»*** Se utiliza con criterios o variables que son o dicotómicos o pueden reducirse a dos opciones. Es igual a la suma de los criterios dispares. Ejemplo:
    Sujeto 1= [01010010]
    Sujeto 2= [10110011] $d_{(1,2)} = 1+1+1+0+0+0+0+1= 4$

➡ ***«Medidas de similaridad»*** Se utiliza igualmente en categorías binarias existen dos indicadores:

En donde:
a = Número de coincidencias positivas
b = Número de dispares (1,0)
c= Número de disparidad (0,1)
d = Número de coincidencias negativas.
P = Total de características analizadas

Concordancia simple $s_s = \frac{a+d}{P}$

Sorenson $S_{so} = \frac{2a}{2a + b + c}$

Jaccard $S_j = \frac{a}{a + b + c}$

Sokal y Sneath $S_o = \frac{2(a+d)}{2(a+d)+b+c}$

Si el sujeto 1[110001] y el sujeto 2[010111] la distancia sería:

$s_{1.2} = \frac{2+1}{2+1+3} = \frac{2+1}{6} = 0{,}5$ O bien $S_{1.2} = \frac{2}{2+3} = \frac{2}{5} = 0{,}4$

- Con las distancias entre sujetos medidas con alguno de estos sistemas se van comparando las distancias entre los componentes de una población hasta configurar tipologías lo más homogéneas posible

➡ El SPSS recoge veintisiete tipos diferentes de indicadores para evaluar la distancia, ya que no hay criterios fiables de cómo evaluar las disimilaridades, ya que es difícil comparar éstas: El más utilizado es el de concordancias simple y de los que excluyen la d el de Jaccard.

- Para el cálculo de la matriz de distancias con ***variables de intervalo o razón*** es conveniente utilizar la *distancia euclídea*
- Con ***variables ordinales*** la mayoría de autores coinciden en tratarlas como de intervalo
- Con ***variables categóricas*** siempre hay el recurso de convertirlas en variables binarias o dicotómicas, de forma que cada categoría constituya una nueva variable binaria dicotómica.

  Ejemplo: la variable estado civil con las categorías Soltero, Casado, Separado Viudo se puede convertir en cuatro variables dicotómicas según la categoría de pertenencia

- Cuando trabajamos con variables en distintas escalas de medida caben varias posibilidades.

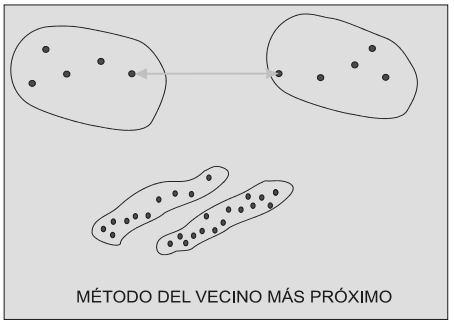
  ➡ Llevar a cabo diferentes análisis agrupando las que son de igual escala de medida

  ➡ Convertir las variables de mayor escala en las de menor, intervalo a ordinal, ordinal a categórica binaria, etc.

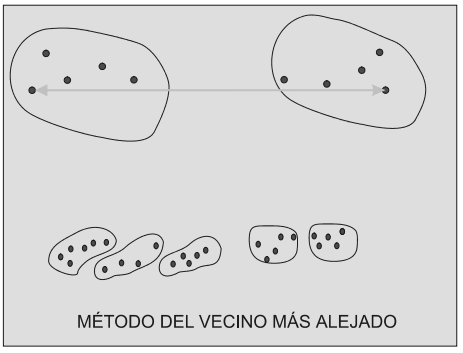
Ejecución del método de agrupamiento

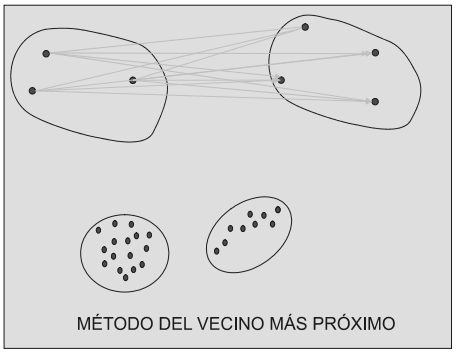
- Una vez que disponemos de la matriz de distancias calculadas por alguno de los procedimientos anteriores, debemos agrupar los sujetos en grupos homogéneos para ello hay que utilizar un método:
- Básicamente se utilizan varios procedimientos:

➡ ***«Vecino más próximo (SLINK)»*** (se comparan pares de sujetos incorporando a un mismo tipo los más cercanos). Este procedimiento no es adecuado en aquellos casos que hay grupos o conglomerados cercanos, ya que sujetos de diferentes grupos pasan a estar en el mismo encadenando todos los casos



➡ *«Vecino más alejado»* Opuesto al anterior utiliza la disimilaridad más grande para establecer los grupos





➡ «*Vinculación intergrupos»* se forman grupos cuya disimilaridad es la media existente entre sus sujetos, da como resultado grupos de forma más o menos esférica

➡ «Método de Ward» También llamado método de la varianza mínima, porque busca separar conglomerados cuya unión conlleve el menor incremento de la varianza. Esto supone que a cada paso debe calcular el valor de la varianza y su variación

Recomendaciones para su aplicación

- Seleccionar variables que realmente estén relacionadas con la solución final que buscamos, ya que las que no discriminan para esa solución constituyen más estorbo que otra cosa.
- Mas que el indicador de distancia o similitud es el método empleado el que ocasiona soluciones diferentes.
- En el método de K-medias su rendimiento disminuye enormemente si se deja al computador (o investigador) seleccionar de forma aleatoria el número de agrupaciones y los centros iniciales de los grupos.
- Por todo ello en general se trabaja siguiendo estas tres etapas.

Selección cuidadosa de las variables (relación)

↓

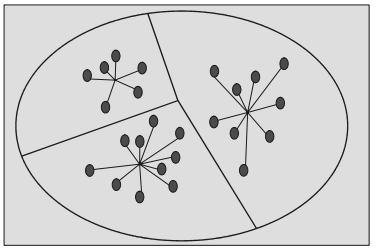
Aplicación de un método jerárquico (Ward o vinculación intergrupos)

↓

Aplicación de un método de optimización K-medias

El procedimiento de *«K-medias»*

- En definitiva es un procedimiento de optimización que divide al universo en k grupos que solicita el investigador y lo hace buscando el centroide más próximo a cada sujeto como muestra la figura en un caso de K=3



Los centroides de un conglomerado ***«v»*** se definen como un punto p-dimensional resultado de promediar en cada variable los valores de las entidades integrantes en el conglomerado

$$\bar{x}(v) = \left(\bar{x}_1(v), \bar{x}_2(v), \bar{x}_3(v), ..., \bar{x}_p(v)\right)$$

⇒ El método trata de buscar los centroides para los K grupos que minimicen las varianzas de las distancias con sus respectivos centros, para ello utiliza la suma errores calculada por la suma de las distancias euclídeas al cuadrado entre las entidades de un conglomerado y su centroide:

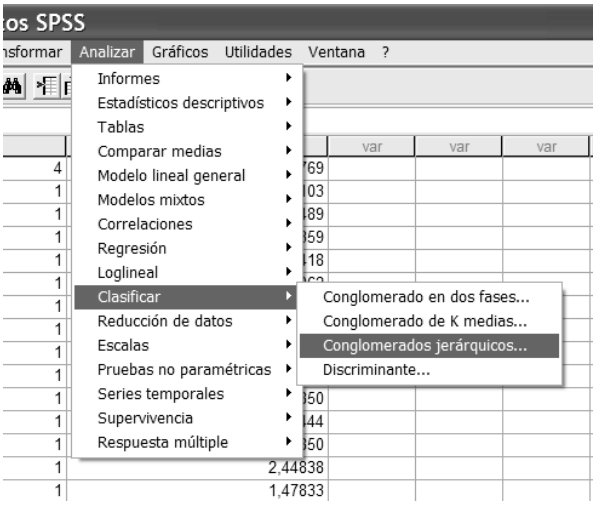
$$SEC(Cv) = \sum_{i \in Cv} \sum_{f=1}^{p} \left(x_{if} - \bar{x}_f(v)\right)^2$$

La suma de los errores para todos los conglomerados sería:

$$SEC = \sum_{V=1}^{k} SEC(Cv)$$

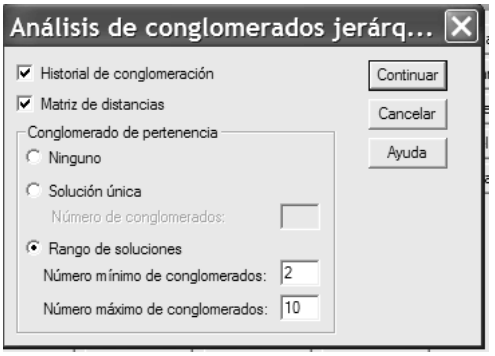
Pasos para realizar una segmentación con SPSS

- Para determinar el numero de K conglomerados y sus centroides procedemos primero a hacer un estudio de segmentación jerárquica. Mediante: ***«Analizar→ Clasificar→ Conglomerados→ jerárquicos»***

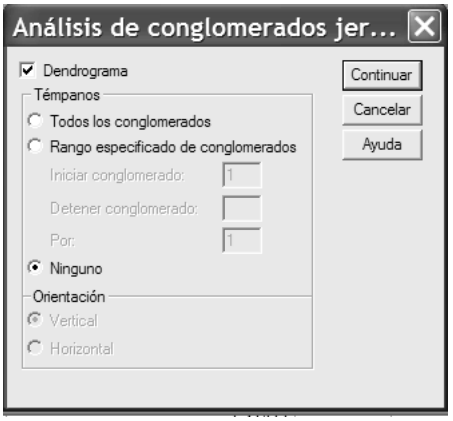


- Incorporamos las variables que queremos formen parte de la segmentación a realizar en el menú correspondiente:

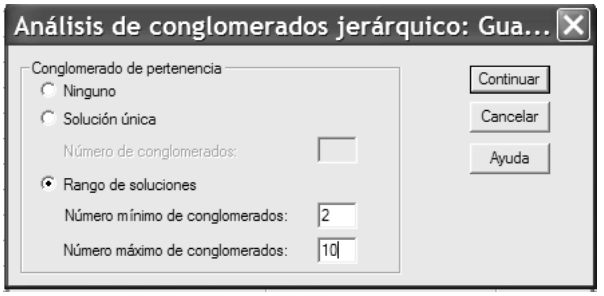
- En ***«Estadísticos»*** solicitamos el ***«Historial de conglomeración»*** y ***«Matriz de distancias»*** con un ***«Rango de soluciones»*** de entre 2 y 10



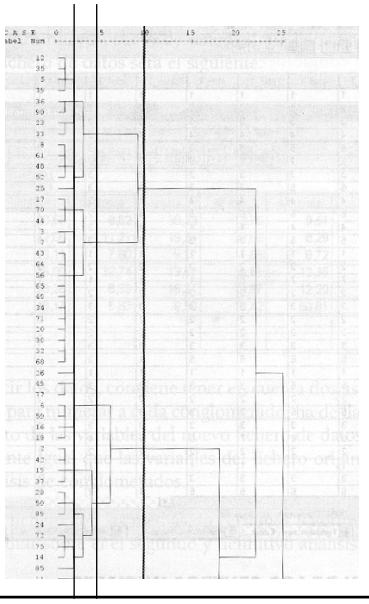
- En ***«Gráficos»*** solicitamos solamente el ***«Dendograma»*** activando en la opción ***«Tempanos» «Ninguno»***



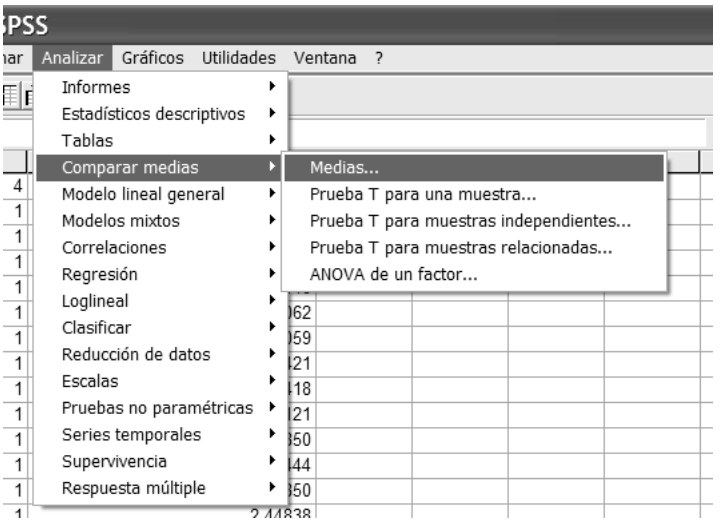
- Finalmente en la opción ***«Guardar»*** pedimos al programa que nos genere en el fichero de datos, 9 nuevas variables, cada una de ellas conteniendo el conglomerado al que asigna a los sujetos. Para ello activamos ***«Rango de soluciones»*** y ponemos de 2 a 10



- Sobre el dendograma de salida elegimos un número de conglomerados apropiado (para ello trazamos una línea vertical en el dendograma y buscamos el punto donde sin muchos conglomerados estos sean homogéneos, para ello la línea tiene que estar lo más cerca posible a la izquierda

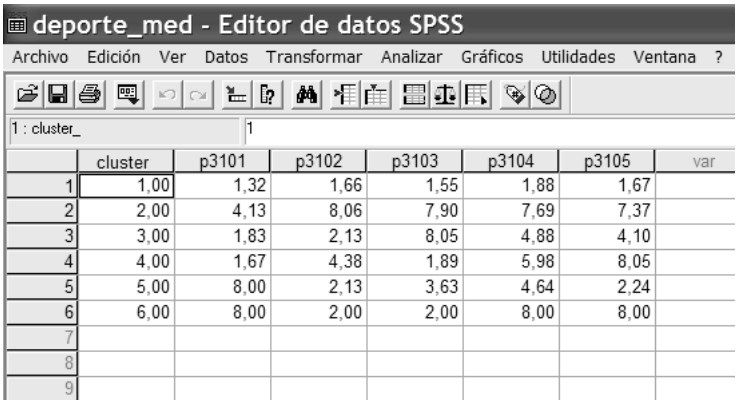


- Una vez que hemos elegido el número K de conglomerados calculamos sus centroides para ello vamos a ***«Analizar→ Comparar medias→Medias»***



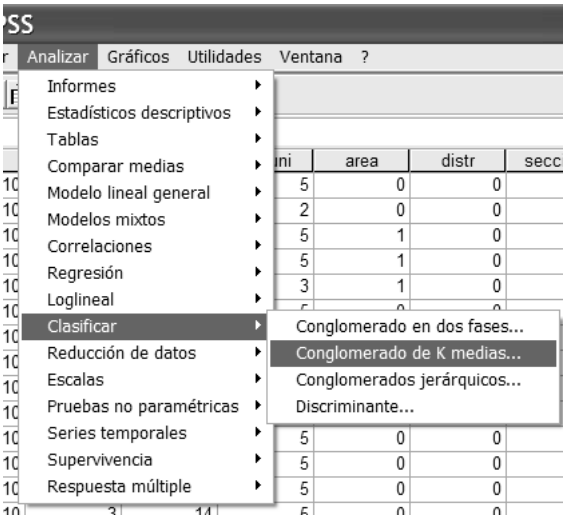
- Seleccionamos de nuevo las variables que hemos utilizado en la segmentación jerárquica y las introducimos colocando como independiente la ***«Average Linkage»*** que corresponda al nº de conglomerados elegidos

- A partir de la salida del paso anterior abrimos un nuevo archivo de datos en blanco en SPSS y a ala primera variable le denominamos exactamente ***«Cluster_»*** y al resto de variables, donde introduciremos las medias que aparecen en la salida, el mismo nombre que poseían las variables en el fichero original, Guardándolo a continuación con el nombre que queramos (centroides.sav)



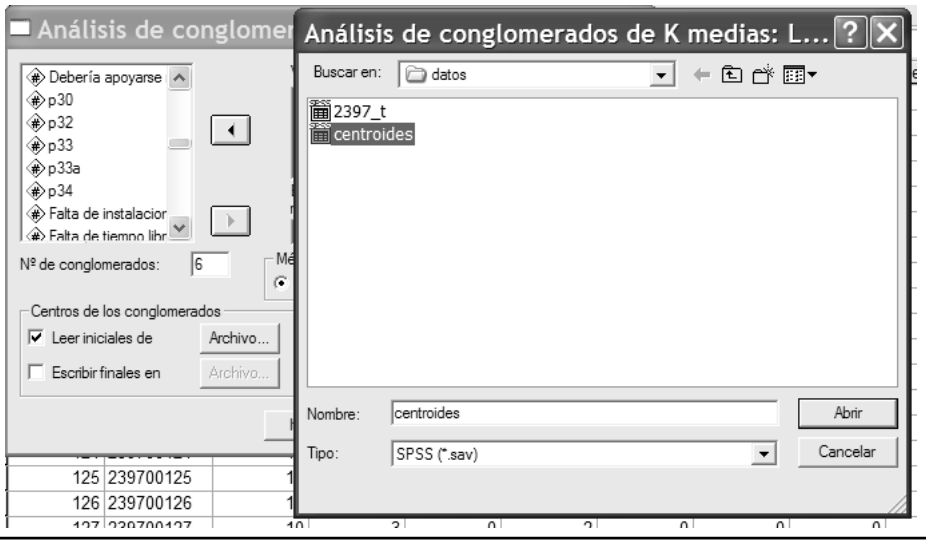
	cluster	p3101	p3102	p3103	p3104	p3105	var
1	1,00	1,32	1,66	1,55	1,88	1,67	
2	2,00	4,13	8,06	7,90	7,69	7,37	
3	3,00	1,83	2,13	8,05	4,88	4,10	
4	4,00	1,67	4,38	1,89	5,98	8,05	
5	5,00	8,00	2,13	3,63	4,64	2,24	
6	6,00	8,00	2,00	2,00	8,00	8,00	
7							
8							
9							

- Ahora estamos en condiciones de iniciar el ***K-medias*** para ello, con el fichero de datos original abierto, nos vamos a ***«Analizar→ Conglomerados→ K medias»***

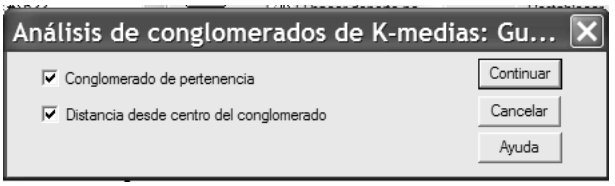


- Volvemos a introducir las variables que queremos formen parte de la segmentación ***«Variables»*** introduciendo el número de K media elegido en ***«Nº de conglomerados»***

- A continuación activamos en ***«Centros de los conglomerados» l***a opción ***«Leer iniciales de»*** pulsando ***«Archivo»*** una vez desplegado el menú buscamos el archivo que guardamos con las medias de los centroides (centroides.sav)



- Pedimos en la opción ***«Guardar»*** que nos guarde tanto el ***«Conglomerado de pertenencia»*** como ***«La distancia desde centro del conglomerado»*** activando ambas opciones



- Pedimos en ***«Estadísticos»*** las opciones que deseemos