ANÁLISIS DE CONGLOMERADOS JERÁRQUICO

El análisis de conglomerados es un procedimiento estadístico de clasificación que pretende identificar grupos relativamente homogéneos de casos (o de variables) basándose en las características seleccionadas. Dentro del análisis de conglomerados están los procedimientos jerárquicos y los no jerárquicos. Aquí estudiaremos los procedimientos jerárquicos.

Para realizar una clasificación mediante un análisis de conglomerados, primero debemos tener los datos que vamos a analizar en el Editor de datos. Para ello, nos situamos en dicha ventana y en el menú Archivo elegimos Abrir, y seleccionamos el fichero **Datos_CCAA.sav**, que contiene información relativa a ciertas características de las Comunidades Autónomas españolas, extraída de la edición de 2005 de los Indicadores Sociales que elabora el INE. Si, una vez abierto el fichero de datos, miramos en la vista de variables, se encuentra la siguiente información:

- **categoria:** Tipo de Entidad Territorial. Esta variable presenta las siguientes categorías:
    - **0:** Total Nacional.
    - **1:** Comunidad Autónoma.
    - **2:** Provincia.
- **entidad:** Denominación geográfica.
- **pibpcpm:** PIB per cápita a precios de mercado.
- **tactiv:** Tasa de actividad.
- **tparo:** Tasa de paro.
- **agri:** Porcentaje de personas ocupadas en el sector de agricultura.
- **ind:** Porcentaje de personas ocupadas en el sector de industria.
- **cons:** Porcentaje de personas ocupadas en el sector de la construcción.
- **serv:** Porcentaje de personas ocupadas en el sector servicios.
- **menor_15:** Porcentaje de personas menores de 15 años.
- **entre_15_64:** Porcentaje de personas entre 15 y 64 años.
- **mayor_64:** Porcentaje de personas mayores de 64 años.
- **densidad:** Densidad poblacional (habitantes/km^2).
- **crec:** Crecimiento natural (por cada 1000 habitantes).
- **pibsup:** PIB per cápita a precios de mercado superior al nacional. Esta variable presenta las siguientes categorías:
    - **0:** Inferior.
    - **1:** Superior.

Como se aprecia, hay 19 casos en el fichero de datos. Por motivos expositivos, vamos a seleccionar un subconjunto de casos para clasificar, de manera que el número de casos sea lo suficientemente reducido como para poder analizar los resultados alcanzados. En concreto vamos a realizar la clasificación de únicamente aquéllas Comunidades Autónomas cuyo PIB per cápita a precios de mercado sea superior al nacional. Por lo tanto, debemos seleccionar aquéllos casos para los que la variable **pibsup** toma el valor igual a 1.

Para seleccionar un subconjunto de casos del fichero de análisis, es necesario ejecutar en SPSS el procedimiento **Seleccionar casos** dentro del menú de **Datos**. En concreto, en el procedimiento hay que indicarle al sistema que se quede con aquéllos casos para los cuales la variable **pibsup** toma el valor 1, como se muestra en la Figura 1:

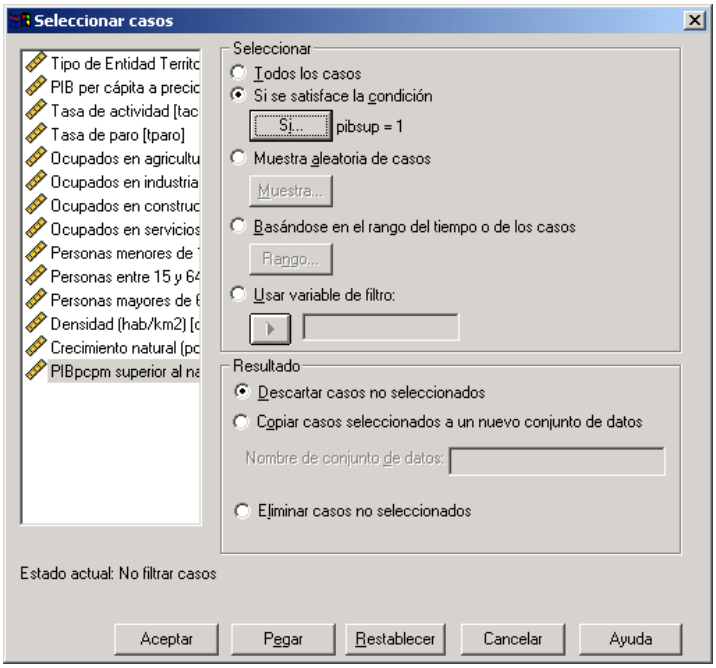


Figura 1: Procedimiento **Seleccionar casos**.

De esta forma, el conjunto de entidades territoriales que se pretenden clasificar mediante algún método de análisis de conglomerados jerárquico se reduce a 7 Comunidades Autónomas, que son, por orden alfabético, las siguientes:

- Aragón.
- Baleares (Illes).
- Cataluña.
- Madrid (Comunidad de).
- Navarra (Comunidad Foral).
- País Vasco
- Rioja (La)

Dentro de SPSS, el procedimiento que permite realizar el análisis de conglomerados jerárquico se encuentra en el submenú **Clasificar** del menú **Analizar**.

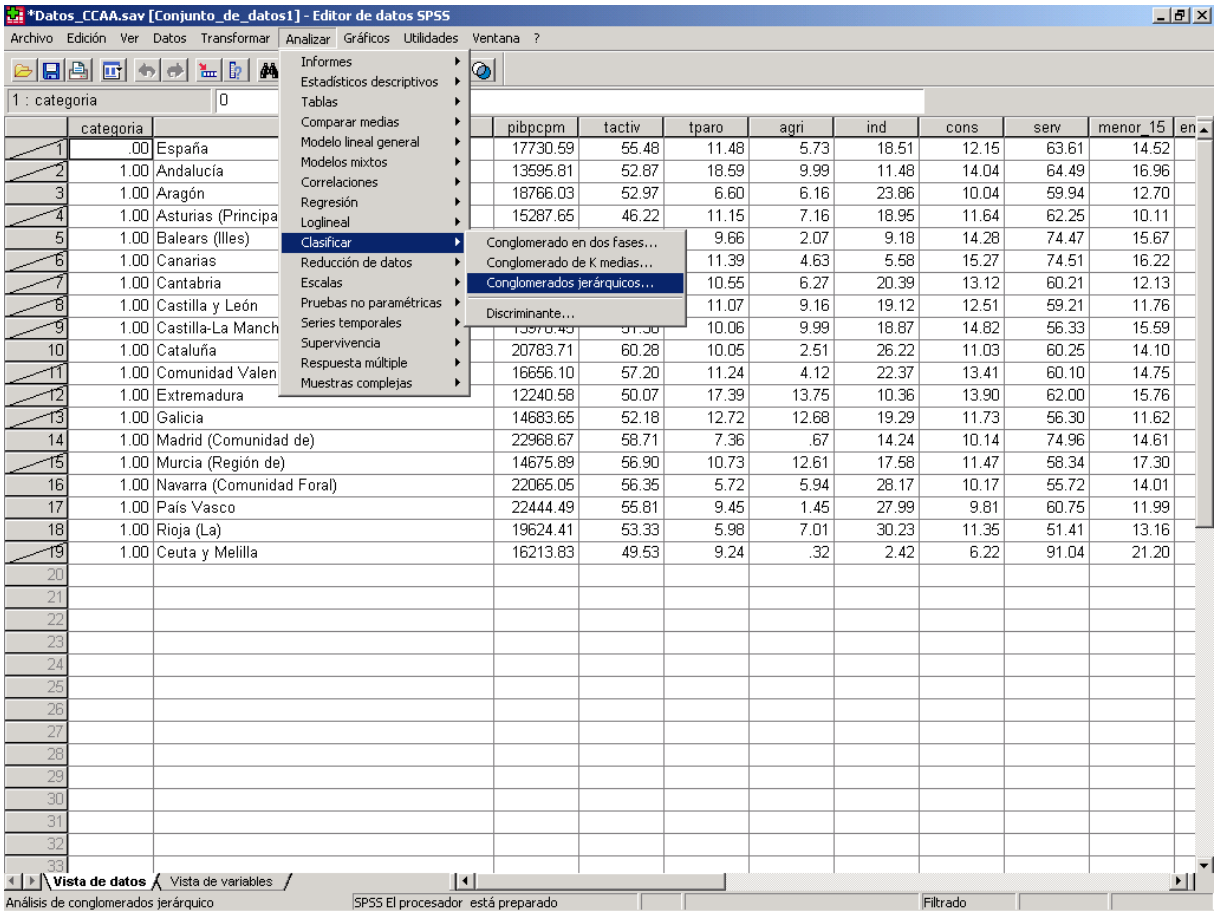


Figura 2: Selección del procedimiento **Conglomerados jerárquicos**.

Al pulsar en dicha opción, el cuadro de diálogo que aparece tiene el aspecto de la Figura 3, en la cual se pueden apreciar todas las opciones que permite SPSS en este procedimiento.

Figura 3: Cuadro de diálogo del procedimiento **Análisis de conglomerados jerárquico**.

En primer lugar, es necesario indicar al procedimiento cuáles son las variables sobre las cuales se va a elaborar la clasificación jerárquica. En este caso, las variables que se deben incorporar al cuadro de destino son las siguientes:

- **tactiv.**
- **tparo.**
- **agri.**
- **ind.**
- **serv.**
- **menor_15.**
- **mayor_64.**

Es conveniente, si se va a realizar una clasificación de los casos del fichero de análisis, que se indique una variable que represente a los casos. En nuestro ejemplo, vamos a indicar a SPSS que etiquete los casos mediante su **Denominación geográfica** (variable **entidad**).

Hay que recordar que el análisis de conglomerados jerárquico se basa en la matriz de distancias entre el conjunto de casos o variables analizadas. La matriz de distancias se calcula a partir de las observaciones (tanto para casos como para variables) utilizando cualquiera de las distancias disponibles en SPSS (ver práctica sobre medidas de similaridad y disimilaridad), que pueden ser elegidas en la opción del método y que veremos más adelante.

Los estadísticos que se pueden solicitar en el procedimiento de Análisis de conglomerados jerárquico aparecen en el siguiente cuadro de diálogo:

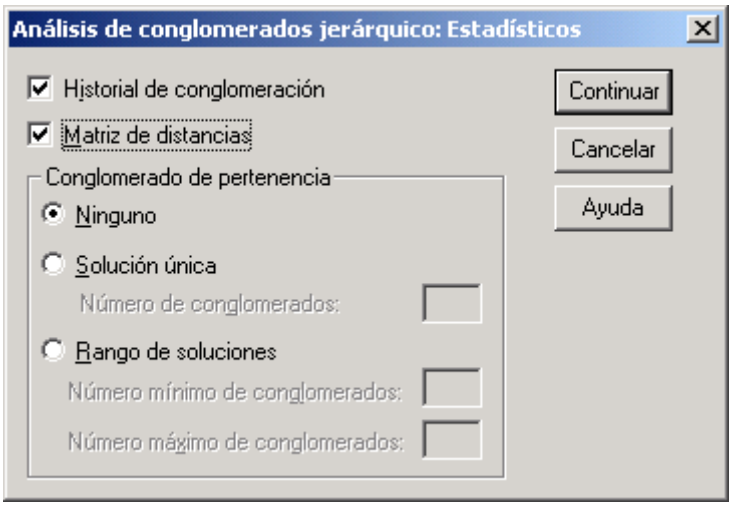


Figura 4: Opciones de **Estadísticos** del análisis de conglomerados.

Las opciones gráficas que permite el análisis de conglomerados son principalmente dos: el dendrograma y el diagrama de témpanos o de carámbanos. El dendrograma es un árbol lógico que permite representar una clasificación jerárquica, mientras que el diagrama de carámbanos es otra representación que permite seguir la clasificación realizada paso a paso. Veremos posteriormente su interpretación.

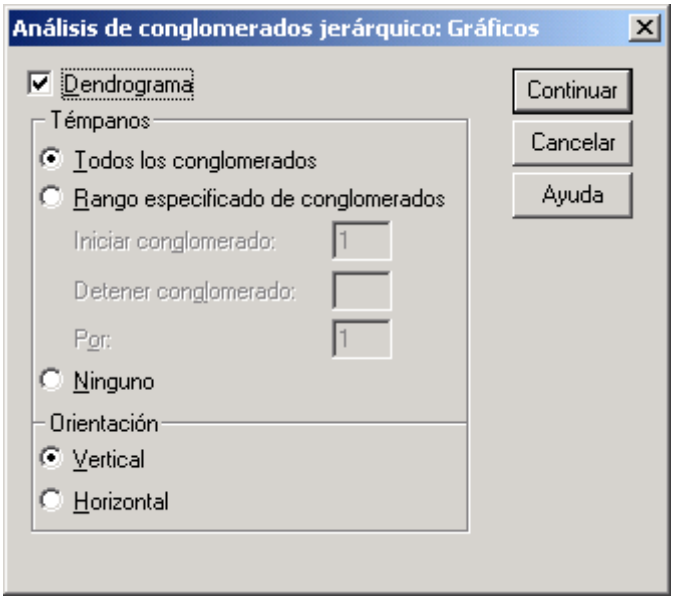


Figura 5: Opciones de **Gráficos** del análisis de conglomerados.

Al pulsar el botón de **Método** del cuadro de diálogo del Análisis de conglomerados jerárquico, aparece el siguiente cuadro de diálogo en el que se muestran las principales opciones que se han de seleccionar para realizar el análisis, en cuanto al algoritmo que se desea usar y la distancia que se desea considerar.

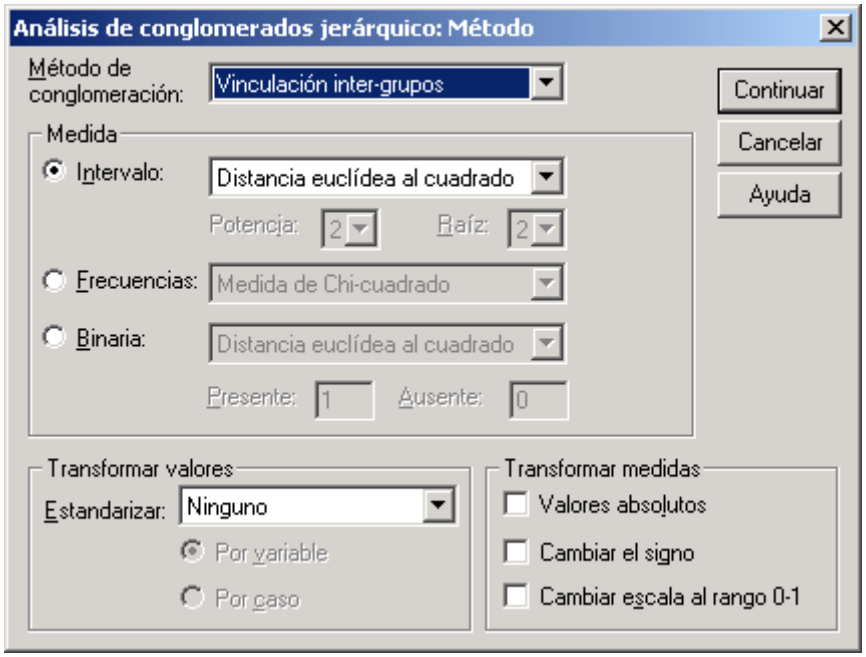


Figura 6: Opciones del **Método** del análisis de conglomerados.

En cuanto al **Método de conglomeración**, las opciones disponibles son la vinculación inter-grupos e intra-grupos, el vecino más próximo (single linkage o encaje simple), el vecino más lejano (complete linkage o encaje completo), la agrupación de centroides (método del centroide), la agrupación de medianas y el método de Ward (momento central de orden 2).

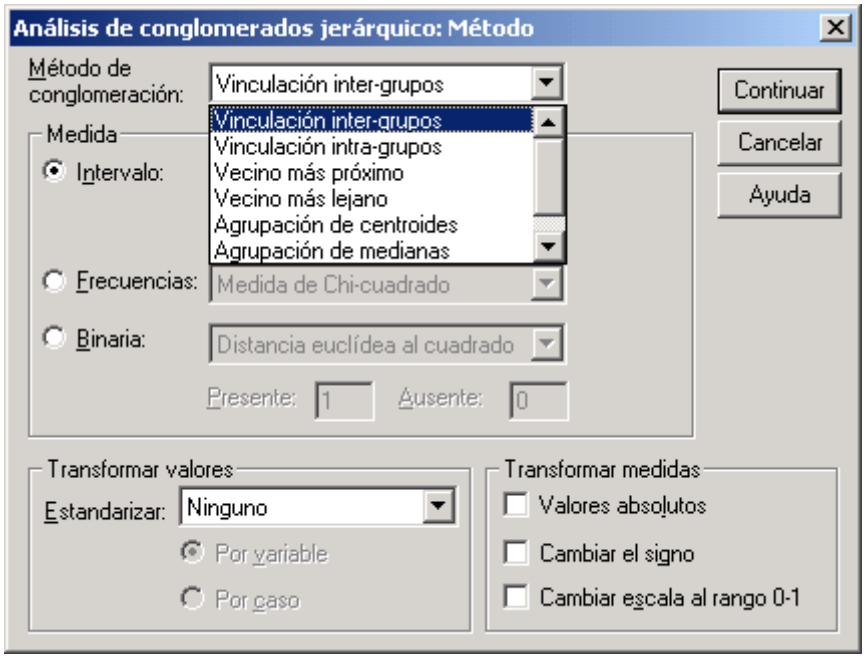


Figura 7: Métodos de conglomeración del análisis de conglomerados.

En las opciones de **Medida**, habrá que elegir la medida que se adapte a las características de nuestras observaciones. Las distancias entre las que podemos elegir son las que aparecen al desplegar los diferentes tipos de distancia:

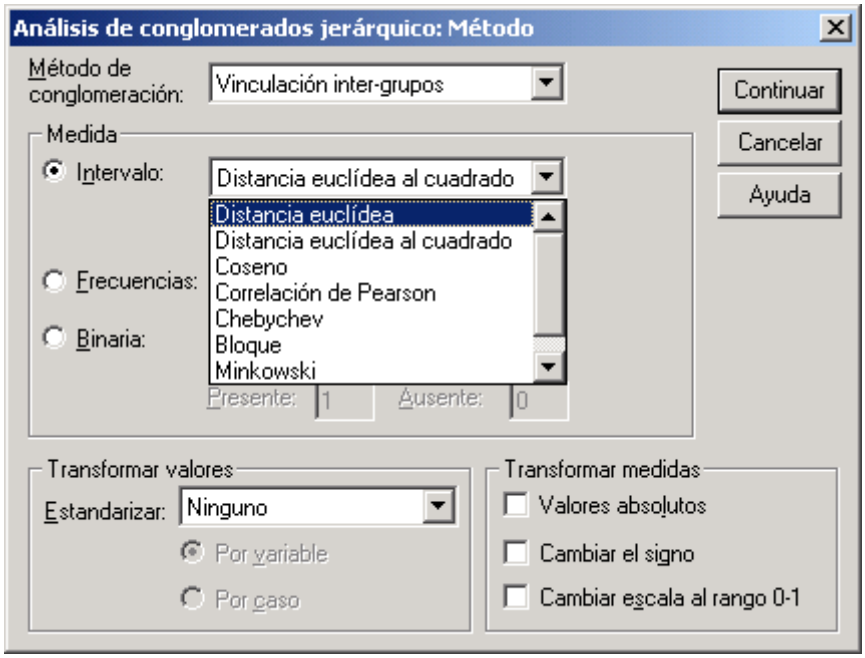


Figura 8: Elección de la distancia utilizada.

También se pueden estandarizar los valores, si tenemos unidades de medida muy distintas entre las variables, lo cual puede distorsionar los resultados.

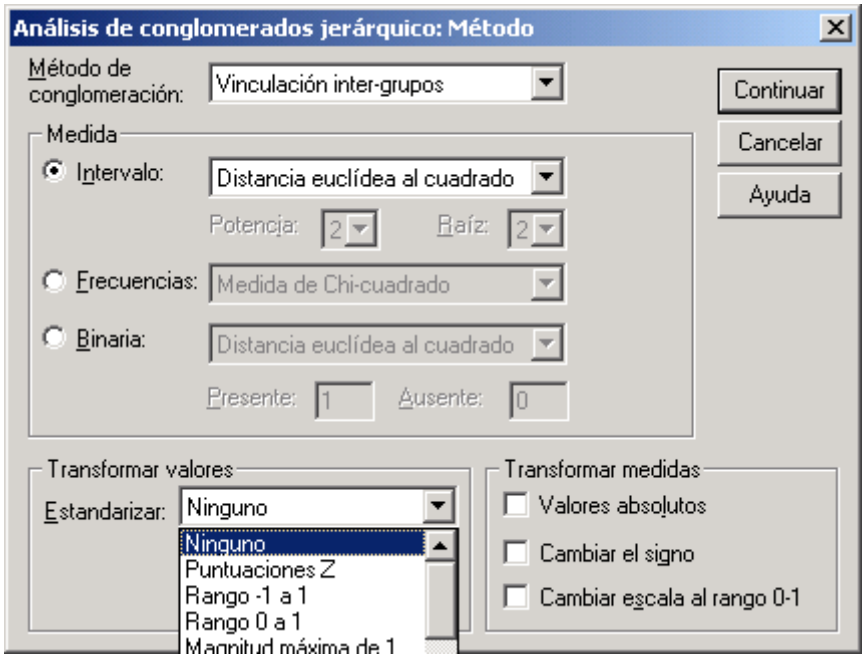


Figura 9: Estandarización previa de variables.

Para realizar la práctica correspondiente al análisis de conglomerados jerárquico, la medida de disimilaridad que vamos a emplear es la **distancia euclídea** sobre **datos estandarizados (puntuaciones Z)**, y el método de agrupamiento será el de **Vecino más próximo**, como se muestra en la Figura 10:

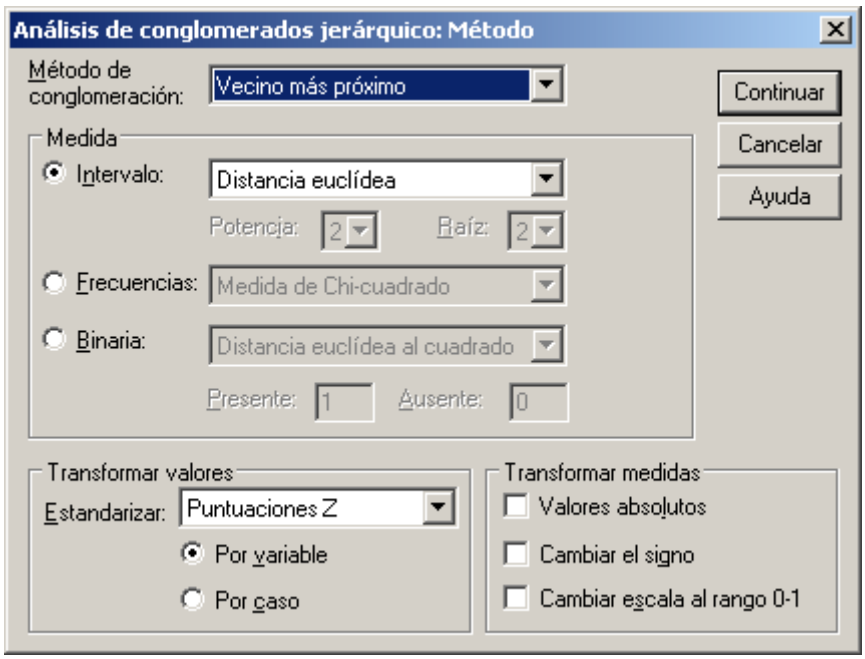


Figura 10: Selección del método empleado en el análisis.

La matriz de distancias euclídeas entre las diversas entidades comarcales, que servirá de apoyo a los diferentes métodos de clasificación que vamos a utilizar es:

Matriz de distancias

	distancia euclídea						
Caso	1:Aragón	2:Balears (Illes)	3:Cataluña	4:Madrid (Comunidad de	5:Navarra (Comunidad F	6:País Vasco	7:Rioja (La)
1:Aragón	.000	5.687	3.697	4.585	2.129	2.887	1.614
2:Balears (Illes)	5.687	.000	3.265	2.046	4.893	4.812	6.004
3:Cataluña	3.697	3.265	.000	3.005	2.976	2.236	3.753
4:Madrid (Comunidad de	4.585	2.046	3.005	.000	3.879	3.767	4.948
5:Navarra (Comunidad F	2.129	4.893	2.976	3.879	.000	3.178	1.387
6:País Vasco	2.887	4.812	2.236	3.767	3.178	.000	3.285
7:Rioja (La)	1.614	6.004	3.753	4.948	1.387	3.285	.000

Esta es una matriz de disimilaridades

Si la matriz de distancias euclídeas al cuadrado la hubiéramos calculado con los valores sin tipificar, el resultado habría sido:

Matriz de distancias

	distancia euclídea						
Caso	1:Aragón	2:Balears (Illes)	3:Cataluña	4:Madrid (Comunidad de	5:Navarra (Comunidad F	6:País Vasco	7:Rioja (La)
1:Aragón	.000	24.571	10.126	20.726	7.825	8.084	10.912
2:Balears (Illes)	24.571	.000	22.580	7.026	28.256	24.873	33.609
3:Cataluña	10.126	22.580	.000	19.495	8.416	5.518	13.567
4:Madrid (Comunidad de	20.726	7.026	19.495	.000	24.734	20.610	30.072
5:Navarra (Comunidad F	7.825	28.256	8.416	24.734	.000	7.993	5.949
6:País Vasco	8.084	24.873	5.518	20.610	7.993	.000	11.980
7:Rioja (La)	10.912	33.609	13.567	30.072	5.949	11.980	.000

Esta es una matriz de disimilaridades

En este caso, los resultados son prácticamente iguales, aunque existen ciertas discrepancias en las clasificaciones obtenidas usando uno u otra matriz de distancias, debido a ciertas diferencias en la relación existente en las distancias calculadas entre los casos.

Los resultados que se obtienen con el método de agrupación por el vecino más próximo (vinculación o encaje simple) son los siguientes:

Vinculación simple

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	5	7	1.387	0	0	2
2	1	5	1.614	0	1	5
3	2	4	2.046	0	0	6
4	3	6	2.236	0	0	5
5	1	3	2.887	2	4	6
6	1	2	3.005	5	3	0

Diagrama de témpanos vertical

Número de conglomerados	Caso												
	4:Madrid (Comunidad de		2:Balears (Illes)		6:País Vasco		3:Cataluña		7:Rioja (La)		5:Navarra (Comunidad F		1:Aragón
1	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
3	X	X	X		X	X	X		X	X	X	X	X
4	X	X	X		X		X		X	X	X	X	X
5	X		X		X		X		X	X	X	X	X
6	X		X		X		X		X	X	X		X

Dendrograma

* * * * * * H I E R A R C H I C A L C L U S T E R A N A L Y S I S * * * * * *

Dendrogram using Single Linkage

Rescaled Distance Cluster Combine

```
        C A S E              0         5        10        15        20        25
Label                Num     +---------+---------+---------+---------+---------+

Navarra (Comunidad F   5
Rioja (La)             7
Aragón                 1
Cataluña               3
País Vasco             6
Balears (Illes)        2
Madrid (Comunidad de   4
```

Si se utiliza el método del vecino más lejano (encaje completo):

Vinculación completa

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	5	7	1.387	0	0	3
2	2	4	2.046	0	0	6
3	1	5	2.129	0	1	5
4	3	6	2.236	0	0	5
5	1	3	3.753	3	4	6
6	1	2	6.004	5	2	0

Diagrama de témpanos vertical

Número de conglomerados	Caso												
	4:Madrid (Comunidad de		2:Balears (Illes)		6:País Vasco		3:Cataluña		7:Rioja (La)		5:Navarra (Comunidad F		1:Aragón
1	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
3	X	X	X		X	X	X		X	X	X	X	X
4	X	X	X		X		X		X	X	X	X	X
5	X	X	X		X		X		X	X	X		X
6	X		X		X		X		X	X	X		X

Dendrograma

```
* * * * * * H I E R A R C H I C A L  C L U S T E R  A N A L Y S I S * * * * * *

 Dendrogram using Complete Linkage

                                  Rescaled Distance Cluster Combine

        C A S E                0         5        10        15        20        25
  Label                  Num  +---------+---------+---------+---------+---------+

  Navarra (Comunidad F    5
  Rioja (La)              7
  Aragón                  1
  Cataluña                3
  País Vasco              6
  Balears (Illes)         2
  Madrid (Comunidad de    4
```

Si se utiliza el método de agrupación al centroide:

Vinculación de centroides

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	5	7	1.387	0	0	2
2	1	5	1.524	0	1	5
3	2	4	2.046	0	0	6
4	3	6	2.236	0	0	5
5	1	3	2.167	2	4	6
6	1	2	2.887	5	3	0

Diagrama de témpanos vertical

Número de conglomerados	Caso												
	4:Madrid (Comunidad de		2:Balears (Illes)		6:País Vasco		3:Cataluña		7:Rioja (La)		5:Navarra (Comunidad F		1:Aragón
1	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
3	X	X	X		X	X	X		X	X	X	X	X
4	X	X	X		X		X		X	X	X	X	X
5	X		X		X		X		X	X	X	X	X
6	X		X		X		X		X	X	X		X

Dendrograma

```
* * * * * * H I E R A R C H I C A L  C L U S T E R  A N A L Y S I S * * * * * *

 Dendrogram using Centroid Method

                              Rescaled Distance Cluster Combine

        C A S E           0         5        10        15        20        25
  Label                Num  +---------+---------+---------+---------+---------+

  Navarra (Comunidad F   5
  Rioja (La)             7
  Aragón                 1
  Cataluña               3
  País Vasco             6
  Balears (Illes)        2
  Madrid (Comunidad de   4
```

Se podrían hacer estos mismos análisis para clasificar variables, en cuyo caso sólo habría que seleccionar **Conglomerar Variables** dentro del cuadro de diálogo del **Análisis de conglomerados jerárquico**, y se obtendrían los siguientes resultados utilizando el método de agrupación al centroide, y como distancia la euclídea, con puntuaciones tipificadas:

Vinculación de centroides

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	1	6	1.472	0	0	2
2	1	5	1.566	1	0	5
3	4	7	1.658	0	0	4
4	3	4	1.601	0	3	6
5	1	2	1.905	2	0	6
6	1	3	3.064	5	4	0

Diagrama de témpanos vertical

Número de conglomerados	Caso												
	Personas mayores de 64 años (%)		Ocupados en industria (%)		Ocupados en agricultura (%)		Tasa de paro		Ocupados en servicios (%)		Personas menores de 15 años (%)		Tasa de actividad
1	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
3	X	X	X	X	X		X		X	X	X	X	X
4	X	X	X		X		X		X	X	X	X	X
5	X		X		X		X		X	X	X	X	X
6	X		X		X		X		X		X	X	X

Dendrograma

```
* * * * * * H I E R A R C H I C A L  C L U S T E R  A N A L Y S I S * * * * * *

 Dendrogram using Centroid Method

                      Rescaled Distance Cluster Combine

   C A S E      0         5        10        15        20        25
 Label     Num  +---------+---------+---------+---------+---------+

 tactiv      1
 menor_15    6
 serv        5
 tparo       2
 ind         4
 mayor_64    7
 agri        3
```

Podemos practicar también con otro fichero de datos (**Datos_Provinciales.sav**) en el que se encuentran datos relativos a las mismas variables, pero referidos a las 51 provincias españolas. Si se realiza el análisis mediante el vecino más próximo para las siete variables del ejemplo seleccionadas de este fichero, utilizando la distancia euclídea sobre datos tipificados, obtendríamos los siguientes resultados:

Conglomerados jerárquicos
Distancias

Resumen de procesamiento de los casos[a]

Casos					
Valid		Perdidos		Total	
N	Porcentaje	N	Porcentaje	N	Porcentaje
49	96.1%	2	3.9%	51	100.0%

a. Distancia euclídea usada

Matriz de distancias

Caso	Archivo matricial de entrada						
	Tasa de actividad	Tasa de paro	Ocupados en agricultura (%)	Ocupados en industria (%)	Ocupados en servicios (%)	Personas menores de 15 años (%)	Personas mayores de 64 años (%)
Tasa de actividad	.000	10.375	12.359	8.731	7.604	7.728	12.765
Tasa de paro	10.375	.000	9.371	11.864	8.166	7.012	11.569
Ocupados en agricultura (%)	12.359	9.371	.000	11.494	12.292	10.210	6.650
Ocupados en industria (%)	8.731	11.864	11.494	.000	11.922	11.586	8.946
Ocupados en servicios (%)	7.604	8.166	12.292	11.922	.000	7.748	12.580
Personas menores de 15 años (%)	7.728	7.012	10.210	11.586	7.748	.000	13.050
Personas mayores de 64 años (%)	12.765	11.569	6.650	8.946	12.580	13.050	.000

Vinculación simple

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	7	6.650	0	0	6
2	2	6	7.012	0	0	4
3	1	5	7.604	0	0	4
4	1	2	7.728	3	2	5
5	1	4	8.731	4	0	6
6	1	3	8.946	5	1	0

Diagrama de témpanos vertical

Número de conglomerados	Caso												
	Personas mayores de 64 años (%)		Ocupados en agricultura (%)		Ocupados en industria (%)		Personas menores de 15 años (%)		Tasa de paro		Ocupados en servicios (%)		Tasa de actividad
1	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
3	X	X	X		X		X	X	X	X	X	X	X
4	X	X	X		X		X	X	X		X	X	X
5	X	X	X		X		X	X	X		X		X
6	X	X	X		X		X		X		X		X

Si representamos el dendrograma correspondiente a la clasificación de estas variables, se obtiene:

Dendrograma

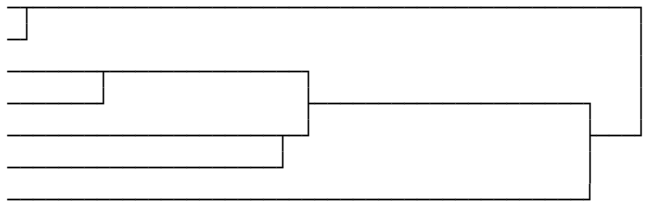
```
* * * * * * H I E R A R C H I C A L  C L U S T E R  A N A L Y S I S * * * * * *

 Dendrogram using Single Linkage

                      Rescaled Distance Cluster Combine

   C A S E      0         5        10        15        20        25
 Label     Num  +---------+---------+---------+---------+---------+

 Agri        3
 Mayor_64    7
 Tparo       2
 Menor_15    6
 Tactiv      1
 Serv        5
 Ind         4
```



Si hacemos el mismo análisis, pero sobre los casos, se obtendría el siguiente dendrograma: