ANÁLISIS DE CONGLOMERADOS NO 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. En esta práctica estudiaremos los procedimientos no jerárquicos, concretamente el método de las K-medias de MacQueen.

El análisis de conglomerados de las K-medias sólo permite clasificar a los casos de la matriz de datos, no a las variables.

Para esta práctica, se va a utilizar el fichero de datos **Datos_Provinciales.sav**, cuyo contenido se presenta en la práctica de Análisis de Conglomerados.

Dentro de SPSS, el procedimiento que permite realizar el análisis de conglomerados de las K-medias de MacQueen se encuentra en el submenú **Clasificar** del menú **Analizar**, justo encima del procedimiento de los conglomerados jerárquicos, como se aprecia en la Figura 1:

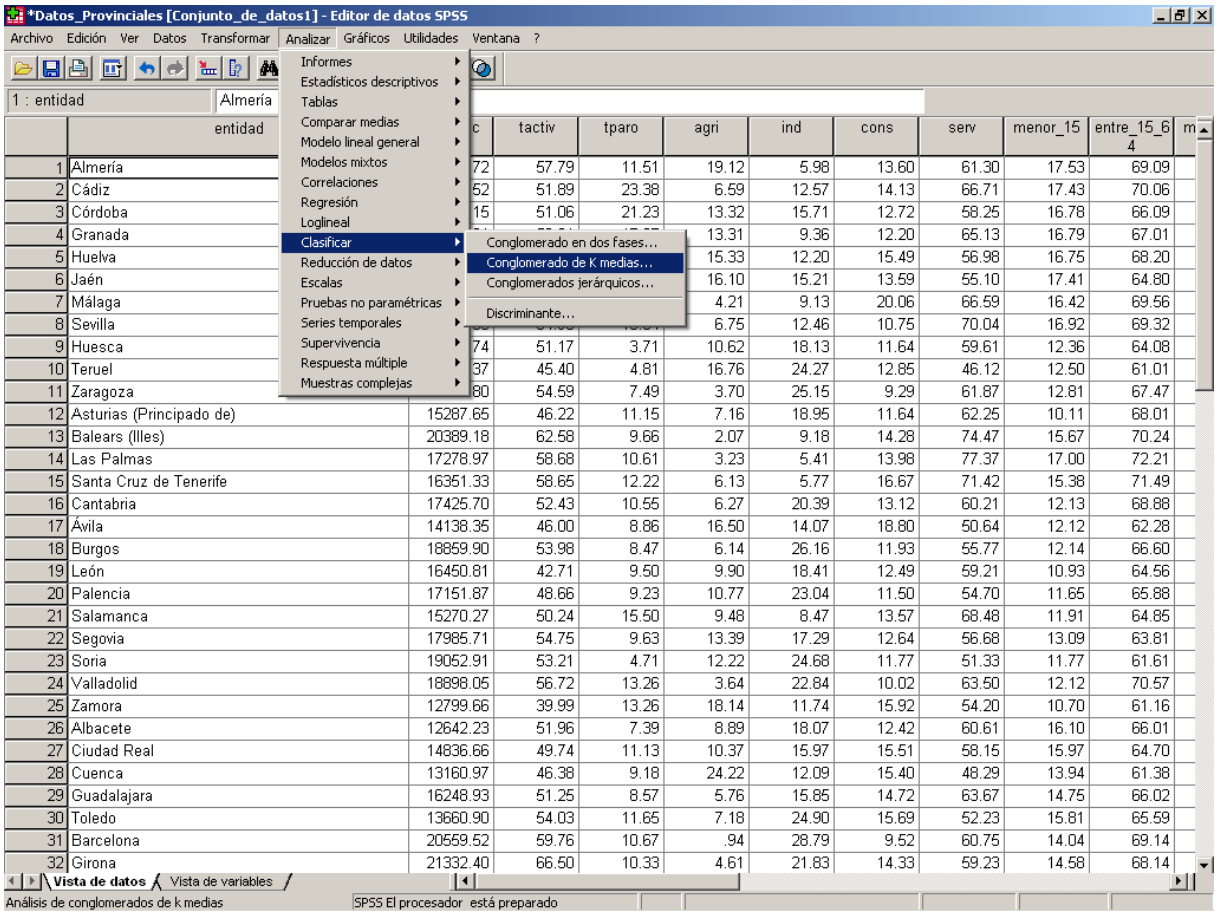

Figura 1: Selección del procedimiento **Conglomerados de K-medias**.

Al pulsar en dicha opción, el cuadro de diálogo que aparece tiene el aspecto de la Figura 2, en la cual se pueden apreciar todas las opciones que permite SPSS en este procedimiento.

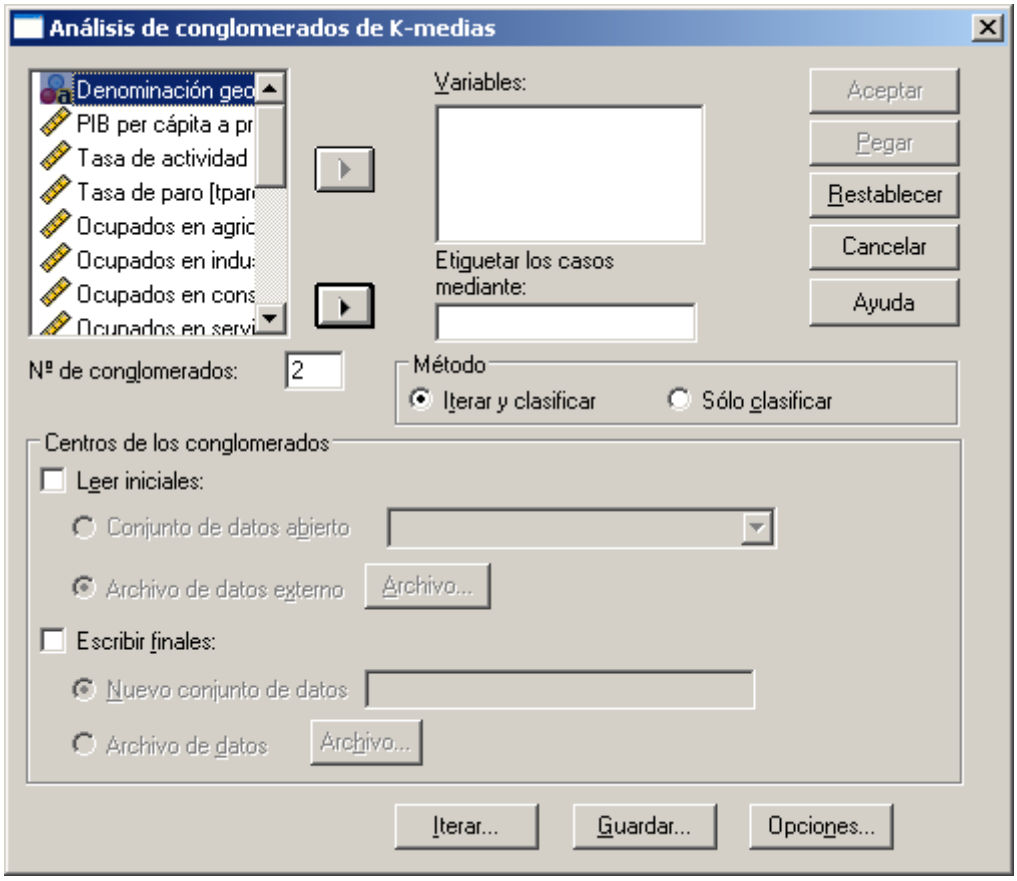

Figura 2: Cuadro de diálogo del procedimiento **Análisis de conglomerados de K-medias**.

Las variables deben ser cuantitativas (escala de medida de intervalo o razón). En caso de tener variables cualitativas, no se puede aplicar este procedimiento, ya que las distancias se calculan utilizando la distancia euclídea. Si las variables utilizan unidades de medida muy diferentes, los resultados podrían ser equívocos y sería conveniente estandarizar las variables antes de realizar el análisis de conglomerados de K-medias (esto se puede hacer mediante el procedimiento **Descriptivos**).

En el apartado de **Centros de los conglomerados**, se muestran las opciones de especificación de los archivos en los que se encuentran los centroides de partida, y en los que se quieran escribir los centroides resultantes.

Si no se indica un fichero en el que se especifiquen los centros iniciales de los conglomerados, se selecciona entre los datos un número de casos debidamente espaciados igual al número de conglomerados fijado.

En el procedimiento de análisis de conglomerados de las K-medias, podemos elegir la opción de realizar la clasificación en torno a los centros iniciales (**Sólo clasificar**) que utiliza los centros iniciales de los conglomerados para clasificar todos los casos, y los centros de los conglomerados no se actualizan, o de conseguir la mejor clasificación (**Iterar y clasificar**) mediante un proceso iterativo de reasignación de los casos al grupo cuyo centroide esté más cercano (los centros iniciales de los conglomerados se utilizan como criterio para una primera clasificación y, a partir de ahí, se van actualizando). Lo normal es utilizar siempre el criterio iterativo a partir de unos centros iniciales.

Si se pulsa en el botón **Iterar...**, se obtiene el cuadro de diálogo que se presenta en la Figura 3 con las siguientes opciones:

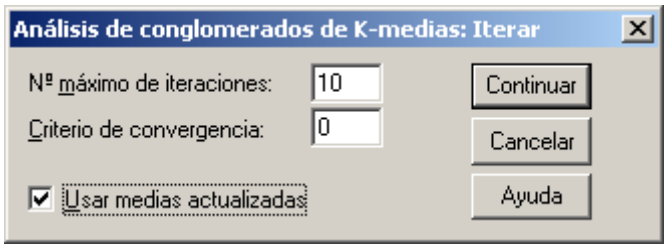

Figura 3: Opciones de Iteración para el algoritmo de las K-medias.

Estas opciones sólo están disponibles si se selecciona el método Iterar y clasificar en el cuadro de diálogo principal. En el **N° máximo de iteraciones** se limita el número de iteraciones en el algoritmo de K-medias. La iteración se detiene después de este número de iteraciones, incluso si no se ha satisfecho el criterio de convergencia. Este número debe estar entre el 1 y el 999. Por defecto está fijado en 10. El **Criterio de convergencia** sirve para determinar cuándo se detiene el proceso de iteración. Representa una proporción de la distancia mínima entre los centros iniciales de los conglomerados, por lo que debe ser mayor que 0 pero no mayor que 1. Por ejemplo, si el criterio es igual a 0,02, la iteración cesará si una iteración completa no mueve ninguno de los centros de los conglomerados en una distancia superior al dos por ciento de la distancia menor entre cualquiera de los centros iniciales. Inicialmente, fijaremos su valor en 0. Seleccionando la opción **Usar medias actualizadas** se permite la actualización de los centros de los conglomerados tras la asignación de cada caso. Si no seleccionáramos esta opción, los nuevos centros de los conglomerados se calcularían después de la asignación de todos los casos.

Se puede guardar información sobre la solución como nuevas variables para que puedan ser utilizadas en análisis subsiguientes, como se aprecia si pulsamos el botón de **Guardar** (Figura 4).

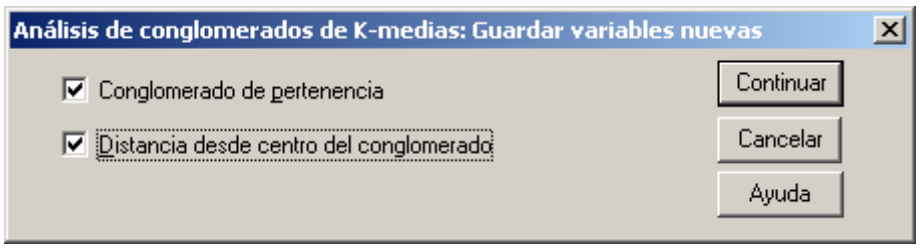

Figura 4: Opciones de Guardar variables nuevas.

Si marcamos la opción **Conglomerado de pertenencia** se crea una nueva variable en el fichero de datos que indica el conglomerado final al que pertenece cada caso del fichero de análisis. Los valores de la nueva variable van desde el 1 hasta el número de conglomerados fijado en el procedimiento. Si marcamos la opción **Distancia desde centro del conglomerado** se crea una nueva variable que indica la distancia euclídea entre cada caso y su centro de clasificación.

En cuanto a las **Opciones** del análisis, son las que se presentan en la Figura 5:

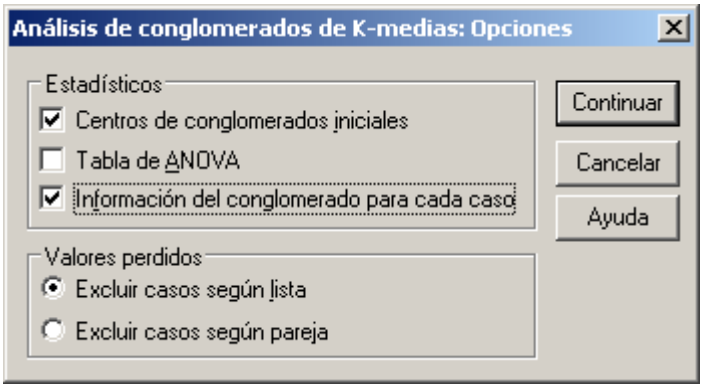

Figura 5: Opciones del procedimiento **Análisis de conglomerados de K-medias**.

Si se solicitan los **Centros de conglomerados iniciales**, se mostrará en los resultados la primera estimación de las medias de las variables para cada uno de los conglomerados. La **Tabla de ANOVA** se corresponde con las pruebas de análisis de la varianza para cada variable de aglomeración. La tabla de ANOVA no se mostrará si se asignan todos los casos a un único conglomerado. Por último, se puede solicitar, mediante la **Información del conglomerado para cada caso**, el conglomerado final asignado y la distancia euclídea entre el caso y el centro del conglomerado utilizado para clasificarlo. También se mostrará la distancia euclídea entre los centros de los conglomerados finales. En caso de que hubiera valores perdidos en el fichero, se puede decidir si excluirlos o no del análisis.

Lo primero que habrá que hacer es estandarizar variables que se van a incluir en el análisis, a través del procedimiento **Descriptivos** dentro del submenú **Estadísticos descriptivos** del menú **Analizar**, como se muestra en la Figura 6:

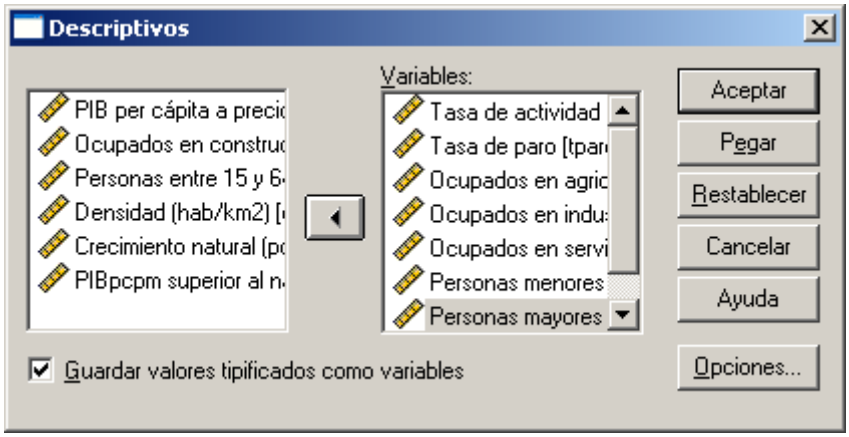

Figura 6: Cuadro de diálogo del procedimiento **Descriptivos**.

Si seleccionamos las siete variables a incluir en el análisis y marcamos la opción de **Guardar valores tipificados como variables**, se almacenarán en el fichero de datos unas nuevas variables, que comienzan con la letra Z, de tal forma que representan los valores tipificados de las variables para cada caso, y sobre esas nuevas variables tipificadas es sobre las que realizaremos el análisis, de la forma que se muestra en la Figura 7:

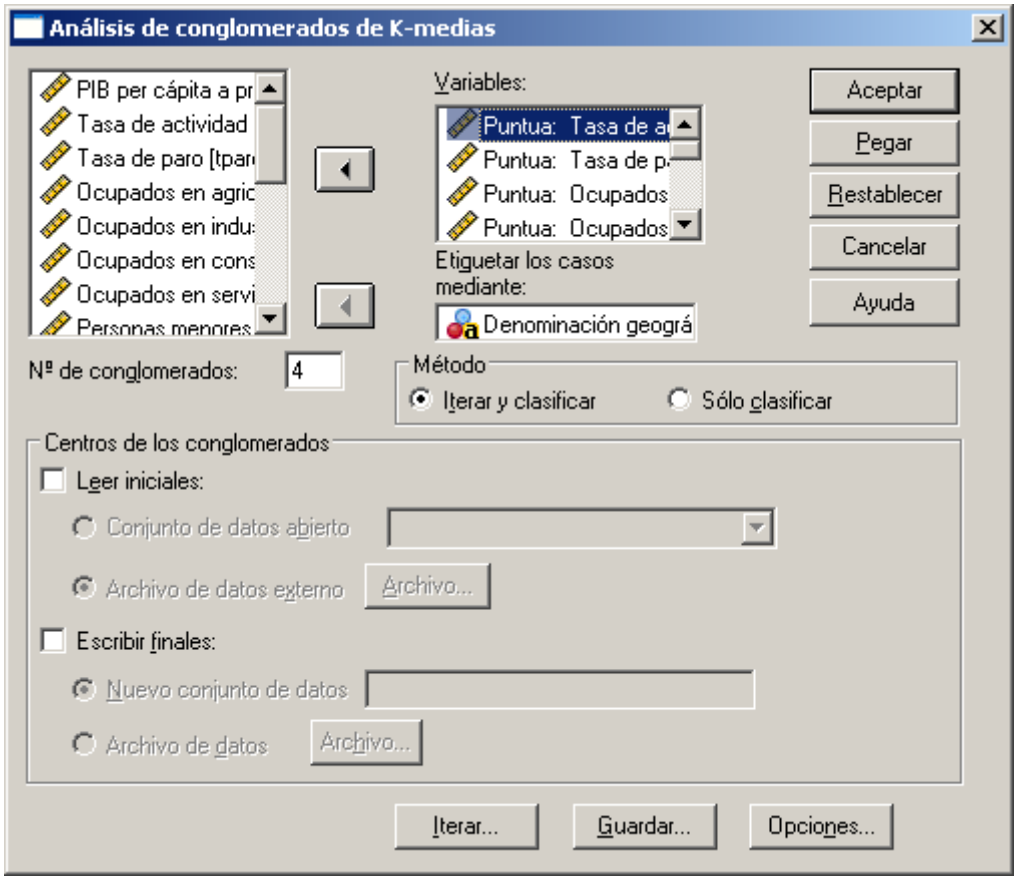

Figura 7: Inclusión de las variables para el análisis propuesto.

En las demás opciones del análisis marcamos las propuestas en la introducción al procedimiento de análisis de conglomerados de K-medias, de tal forma que se obtienen los siguientes resultados:

Análisis de conglomerados de K medias

Los centros iniciales, generados por el propio procedimiento, son los siguientes:

Centros iniciales de los conglomerados

	Conglomerado			
	1	2	3	4
Puntua: Tasa de actividad	1.03325	-.26723	.69192	-2.53641
Puntua: Tasa de paro	-.86460	2.64305	-.99598	.42723
Puntua: Ocupados en agricultura (%)	-1.42652	-.45462	-.42015	1.44156
Puntua: Ocupados en industria (%)	-.50797	-.74540	1.87203	-.86341
Puntua: Ocupados en servicios (%)	2.22491	1.01645	-1.48982	-.81601
Puntua: Personas menores de 15 años (%)	.12721	1.19770	.11961	-1.35706
Puntua: Personas mayores de 64 años (%)	-.92657	-1.36591	-.33347	2.06752

En este procedimiento, el historial de iteraciones se limita a registrar los distintos cambios producidos entre los centros de los grupos:

Historial de iteraciones[a]

	Cambio en los centros de los conglomerados			
Iteración	1	2	3	4
1	.945	1.705	1.889	1.202
2	.189	.213	.288	.627
3	.038	.019	.010	.070
4	.008	.002	.000	.008
5	.002	.000	1.31E-005	.001
6	.000	1.45E-005	4.69E-007	9.55E-005
7	6.05E-005	1.32E-006	1.67E-008	1.06E-005
8	1.21E-005	1.20E-007	5.98E-010	1.18E-006
9	2.42E-006	1.09E-008	2.13E-011	1.31E-007
10	4.84E-007	9.92E-010	7.62E-013	1.46E-008

a. Se han detenido las iteraciones debido a que se ha alcanzado el número máximo de iteraciones. Las iteraciones no han logrado la convergencia. El cambio máximo de coordenadas absolutas para cualquier centro es de 3.26E-007. La iteración actual es 10. La distancia mínima entre los centros iniciales es de 4.218.

Naturalmente, no se ha producido la convergencia del procedimiento debido a que la exigencia que habíamos impuesto es muy fuerte, en el sentido de que solo admitiremos que el algoritmo ha llegado al final cuando no se producen más cambios.

Los centros resultantes de los conglomerados han resultado ser, finalmente:

Centros de los conglomerados finales

	Conglomerado			
	1	2	3	4
Puntua: Tasa de actividad	1.21345	-.24192	.38751	-1.32004
Puntua: Tasa de paro	-.29477	1.47702	-.37879	-.62205
Puntua: Ocupados en agricultura (%)	-1.03989	.42265	-.44851	1.19147
Puntua: Ocupados en industria (%)	-1.30272	-.96642	.63460	-.00262
Puntua: Ocupados en servicios (%)	2.16558	.45850	-.14754	-1.11304
Puntua: Personas menores de 15 años (%)	.52769	.75839	-.24678	-1.06265
Puntua: Personas mayores de 64 años (%)	-1.22972	-.48504	-.03190	1.59645

Por último, se presentan las distancias entre los centroides de los conglomerados resultantes:

Distancias entre los centros de los conglomerados finales

Conglomerado	1	2	3	4
1		3.321	3.490	5.870
2	3.321		2.957	4.148
3	3.490	2.957		3.212
4	5.870	4.148	3.212	

y se indica el número de casos que han resultado incluidos en cada conglomerado:

Número de casos en cada conglomerado

Conglomerado	1	4.000
	2	11.000
	3	25.000
	4	9.000
Válidos		49.000
Perdidos		2.000

La clasificación individual de cada caso en el conjunto general de observaciones se almacena en una nueva variable, llamada **QCL_1** (se puede comprobar su aparición en la hoja de datos), y su distancia al centro del conglomerado al que pertenece en otra nueva variable, llamada **QCL_2**. En el visor de resultados aparece la siguiente tabla:

Pertenencia a los conglomerados

Número de caso	Denominación geográfica	Conglomerado	Distancia
1	Almería	2	2.373
2	Cádiz	2	1.770
3	Córdoba	2	1.153
4	Granada	2	.590
5	Huelva	2	1.050
6	Jaén	2	1.495
7	Málaga	2	1.589
8	Sevilla	2	1.607
9	Huesca	3	2.039
10	Teruel	4	1.590
11	Zaragoza	3	1.018
12	Asturias (Principado de)	3	2.300
13	Balears (Illes)	1	.625
14	Las Palmas	1	.979
15	Santa Cruz de Tenerife	1	.965
16	Cantabria	3	.773
17	Ávila	4	.592
18	Burgos	3	1.124
19	León	4	1.839
20	Palencia	4	1.755
21	Salamanca	2	2.763
22	Segovia	3	1.592
23	Soria	4	2.091
24	Valladolid	3	1.355
25	Zamora	4	1.855
26	Albacete	3	1.437
27	Ciudad Real	3	1.725
28	Cuenca	4	1.771
29	Guadalajara	3	1.417
30	Toledo	3	1.465
31	Barcelona	3	1.779
32	Girona	3	2.314
33	Lleida	3	1.694
34	Tarragona	3	1.614
35	Alicante/Alacant	3	1.218
36	Castellón/Castelló	3	2.120
37	Valencia/València	3	1.333
38	Badajoz	2	.753
39	Cáceres	2	1.600
40	A Coruña	3	1.441
41	Lugo	4	2.242
42	Ourense	4	1.336
43	Pontevedra	3	1.274
44	Madrid (Comunidad de)	1	1.181
45	Murcia (Región de)	3	2.150
46	Navarra (Comunidad Foral)	3	1.427
47	Álava	3	2.248
48	Guipúzcoa	3	1.996
49	Vizcaya	3	1.529
50	Ceuta	.	.
51	Melilla	.	.

Si se repite ahora el análisis de conglomerados de K-medias para los datos de las Comunidades Autónomas, fichero **Datos_CCAA.sav** (recordemos que es necesario seleccionar aquéllas Comunidades cuyo PIB per cápita a precios de mercado es superior al nacional, y que hay que tipificar las variables para evitar los efectos de las distintas unidades de medida de cada una de ellas). Si solicitamos tres grupos, los resultados serán los siguientes:

Análisis de conglomerados de K medias

Los centros iniciales son los siguientes (están generados por el propio programa):

Centros iniciales de los conglomerados

	Conglomerado		
	1	2	3
Puntua: Tasa de actividad	-1.07101	1.52434	.87901
Puntua: Tasa de paro	-1.00138	.98902	1.19996
Puntua: Ocupados en agricultura (%)	1.28149	-.62367	-.45398
Puntua: Ocupados en industria (%)	.92530	-1.71088	.42311
Puntua: Ocupados en servicios (%)	-1.23711	1.33528	-.25099
Puntua: Personas menores de 15 años (%)	-.47624	1.55471	.28436
Puntua: Personas mayores de 64 años (%)	.66476	-1.34216	-.12353

Historial de iteraciones[a]

	Cambio en los centros de los conglomerados		
Iteración	1	2	3
1	.532	.682	.745
2	.133	.227	.248
3	.033	.076	.083
4	.008	.025	.028
5	.002	.008	.009
6	.001	.003	.003
7	.000	.001	.001
8	3.25E-005	.000	.000
9	8.11E-006	.000	.000
10	2.03E-006	3.47E-005	3.79E-005

a. Se han detenido las iteraciones debido a que se ha alcanzado el número máximo de iteraciones. Las iteraciones no han logrado la convergencia. El cambio máximo de coordenadas absolutas para cualquier centro es de 2.89E-005. La iteración actual es 10. La distancia mínima entre los centros iniciales es de 3.265.

Los tres grupos solicitados son los siguientes:

Pertenencia a los conglomerados

Número	Denominación geográfica	Conglomerado	Distancia
1	Aragón	1	1.171
2	Balears (Illes)	2	1.023
3	Cataluña	3	1.118
4	Madrid (Comunidad de)	2	1.023
5	Navarra (Comunidad Foral)	1	1.070
6	País Vasco	3	1.118
7	Rioja (La)	1	.709

Como se observa en esta tabla, los grupos son:

- Aragón, Navarra y Rioja.
- Baleares y Madrid.
- Cataluña y País Vasco.

Estos resultados concuerdan con los que se obtuvieron en el análisis de conglomerados jerárquico por los métodos del vecino más próximo, del vecino más lejano y del centroide, como se vio en la práctica correspondiente al análisis de conglomerados jerárquico.

Otros resultados que también se presentan al realizar este procedimiento son:

Centros de los conglomerados finales

	Conglomerado		
	1	2	3
Puntua: Tasa de actividad	-.82223	.98142	.25192
Puntua: Tasa de paro	-.93648	.36702	1.03770
Puntua: Ocupados en agricultura (%)	1.03467	-.89363	-.65838
Puntua: Ocupados en industria (%)	.57339	-1.39404	.53395
Puntua: Ocupados en servicios (%)	-.75967	1.36261	-.22310
Puntua: Personas menores de 15 años (%)	-.37105	1.12587	-.56929
Puntua: Personas mayores de 64 años (%)	.77737	-1.25770	.09164

Distancias entre los centros de los conglomerados finales

Conglomerado	1	2	3
1		4.840	2.952
2	4.840		3.459
3	2.952	3.459	

Número de casos en cada conglomerado

Conglomerado	1	3.000
	2	2.000
	3	2.000
Válidos		7.000
Perdidos		.000

Veamos qué ocurre si repetimos el análisis solicitando ahora cuatro grupos en lugar de tres:

Análisis de conglomerados de K medias

Los centroides iniciales ahora son cuatro:

Centros iniciales de los conglomerados

	Conglomerado			
	1	2	3	4
Puntua: Tasa de actividad	-.37517	1.52434	.87901	-1.07101
Puntua: Tasa de paro	.87543	.98902	1.19996	-1.00138
Puntua: Ocupados en agricultura (%)	-.86278	-.62367	-.45398	1.28149
Puntua: Ocupados en industria (%)	.64478	-1.71088	.42311	.92530
Puntua: Ocupados en servicios (%)	-.19522	1.33528	-.25099	-1.23711
Puntua: Personas menores de 15 años (%)	-1.42294	1.55471	.28436	-.47624
Puntua: Personas mayores de 64 años (%)	.30681	-1.34216	-.12353	.66476

Historial de iteraciones[a]

	Cambio en los centros de los conglomerados			
Iteración	1	2	3	4
1	.000	.682	.000	.532
2	.000	.227	.000	.133
3	.000	.076	.000	.033
4	.000	.025	.000	.008
5	.000	.008	.000	.002
6	.000	.003	.000	.001
7	.000	.001	.000	.000
8	.000	.000	.000	3.25E-005
9	.000	.000	.000	8.11E-006
10	.000	3.47E-005	.000	2.03E-006

a. Se han detenido las iteraciones debido a que se ha alcanzado el número máximo de iteraciones. Las iteraciones no han logrado la convergencia. El cambio máximo de coordenadas absolutas para cualquier centro es de 2.11E-005. La iteración actual es 10. La distancia mínima entre los centros iniciales es de 2.236.

Pertenencia a los conglomerados

Número de caso	Denominación geográfica	Conglomerado	Distancia
1	Aragón	4	1.171
2	Balears (Illes)	2	1.023
3	Cataluña	3	.000
4	Madrid (Comunidad de)	2	1.023
5	Navarra (Comunidad Foral)	4	1.070
6	País Vasco	1	.000
7	Rioja (La)	4	.709

Centros de los conglomerados finales

	Conglomerado			
	1	2	3	4
Puntua: Tasa de actividad	-.37517	.98142	.87901	-.82223
Puntua: Tasa de paro	.87543	.36702	1.19996	-.93648
Puntua: Ocupados en agricultura (%)	-.86278	-.89363	-.45398	1.03467
Puntua: Ocupados en industria (%)	.64478	-1.39404	.42311	.57339
Puntua: Ocupados en servicios (%)	-.19522	1.36261	-.25099	-.75967
Puntua: Personas menores de 15 años (%)	-1.42294	1.12587	.28436	-.37105
Puntua: Personas mayores de 64 años (%)	.30681	-1.25770	-.12353	.77737

Distancias entre los centros de los conglomerados finales

Conglomerado	1	2	3	4
1		4.199	2.236	2.955
2	4.199		2.966	4.840
3	2.236	2.966		3.346
4	2.955	4.840	3.346	

Número de casos en cada conglomerado

Conglomerado	1	1.000
	2	2.000
	3	1.000
	4	3.000
Válidos		7.000
Perdidos		.000

Como se aprecia, la única diferencia que se ha producido es que se ha dividido uno de los conglomerados anteriores (concretamente el tercero) en otros dos conglomerados de la forma:

- País Vasco.
- Baleares y Madrid.
- Cataluña.
- Aragón, Navarra y Rioja.

Estos resultados son equivalentes a los que se podían observar en la práctica de análisis de conglomerados jerárquicos, en los métodos del vecino más próximo, del vecino más lejano y de la agrupación al centroide.