Análisis de Clúster con SPSS. Método de las K-Medias

El análisis de cluster es un tipo de clasificación de datos que se lleva a cabo mediante la agrupación de los elementos analizados. El objetivo fundamental de este tipo de análisis es el de clasificar *n* objetos en *k* (*k*>1) grupos, llamados *clusters*, mediante la utilización de *p* (*p*>0) variables. Como muchos otros tipos de análisis estadísticos el de *cluster* posee muchas variantes, cada una de las cuales tiene su propio proceso de clasificación.

Los procedimientos del análisis de *cluster* se dividen principalmente en dos sub grupos. En el primero de ellos el número de *clusters* está predefinido. Se conoce como el método de las **K-Medias**. Sin embargo, cuando el número de *clusters* no está predeterminado usaremos el **análisis jerárquico de *clusters***.

La gran variedad de procedimientos de clasificación deriva de las métricas que se utilizan entre los diferentes elementos. Las métricas más habituales que se usan son la Euclídea, la Manhattan, la Chebyshev y otras. Hay además diferentes normas para la creación de los *clusters*. Algunas permiten que los elementos compartan diferentes grupos, mientras que otros sólo admiten 'socios exclusivos'.

Análisis de *Clusters* por el método de las K-Medias

Desde el menú principal del SPSS presiona la opción **Analyze→ Classify → K-Means Cluster**.

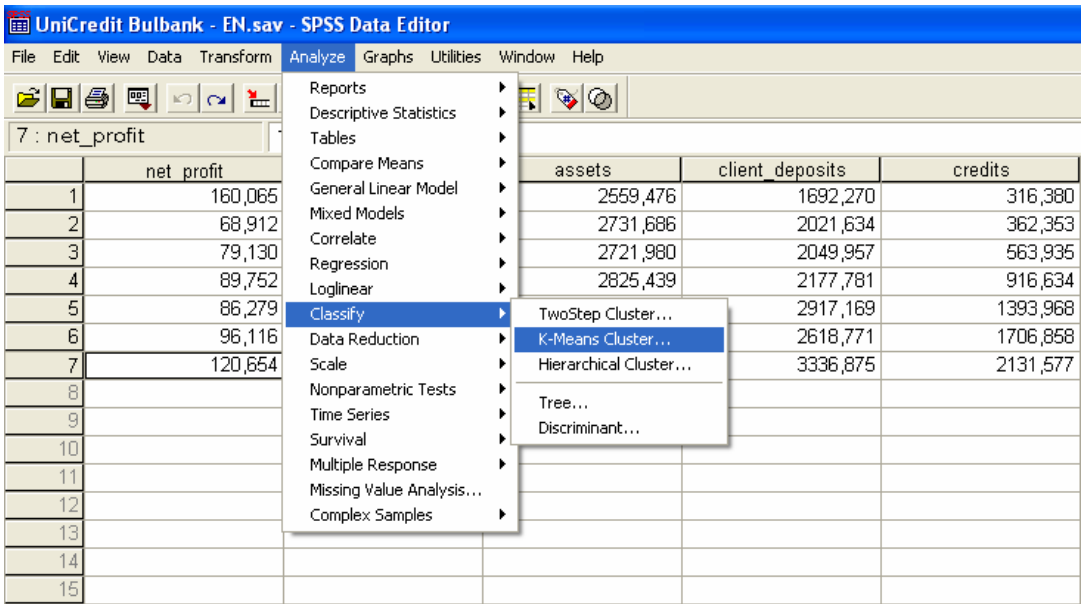

Figura 1.

Señala las variables sobre las que van a construirse los *clusters* y las envía a la opción de **Variables** para incluirlas como tales. La opción de Nombres de Etiquetas (**Label Cases)** se usa para introducir una sucesión de variables que identifique las unidades. Después indicaremos el número de *clusters* que deseamos en la opción Número de Clusters (**Number of Clusters**). En nuestro caso en la opción **Método** (**Method)** escogeremos **Iterate and Classify**. A diferencia del método alternativo (**Solo Clasificar** -que define centros de *cluster* rígidos), este otro método determina las sucesivas iteraciones e indica cómo llevar a cabo la clasificación final.

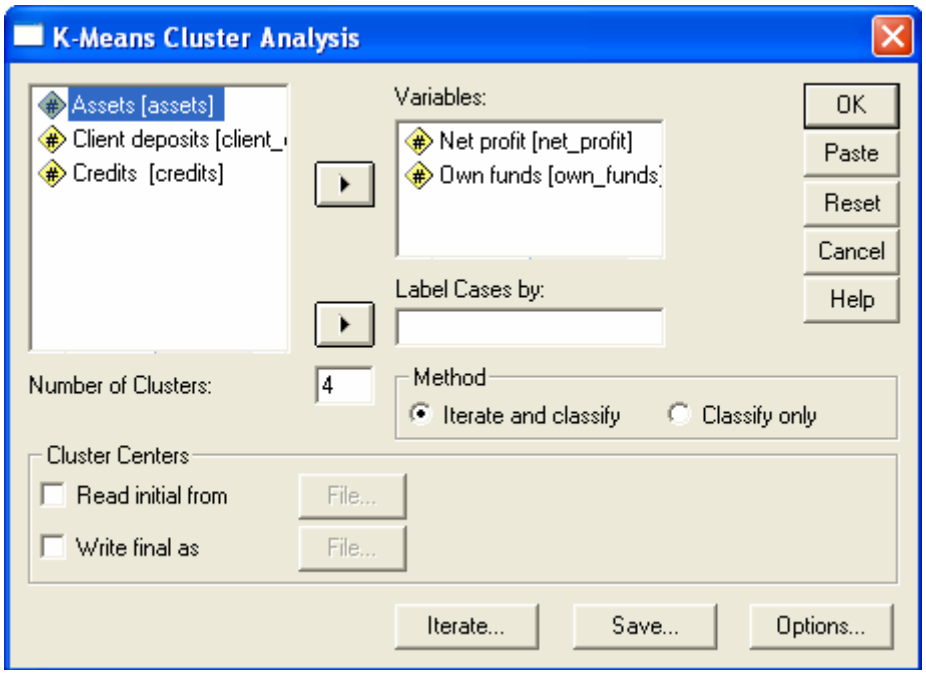

Figura 2.

En la opción Centros de *Cluster* (**Cluster Centers)** especificaremos la carpeta (si es una), que contiene el centro de *cluster* inicial y, si se requiere, la que tiene el centro de *cluster* final. En la opción '**Read initial from'** indicaremos el archivo que tenga el centro de *cluster* inicial, y en la opción '**Write final as'** especificaremos el que contenga el centro de *cluster* final. Con el botón de iterar '**Iterate'** podemos determinar el criterio de actualización de los centros de *cluster*, en iteraciones máximas (**Maximum Iterations)** indicaremos el número máximo de iteraciones posibles (no más de 999), y en la opción 'Criterio de Convergencia' (**Convergence Criterion)** decidiremos qué regla parará el proceso de iteración. Por defecto el proceso se realiza con 10 iteraciones y 0 convergencias. Además, es posible seleccionar la opción **Use running means**. Si se marca, los centros de *cluster* cambiarán tras la suma de cada uno de los elementos. Si en cambio, no se selecciona, calculará los centros de *cluster* después de que todos los elementos hayan sido ubicados en cada *cluster*. En ambos casos obtendremos resultados diferentes y se deberá especificar la vía por la que se ha logrado la clasificación. Avancemos seleccionando 'Continue'.

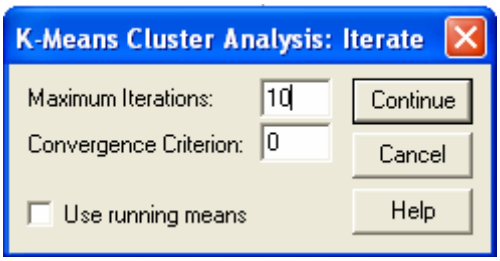

Figura 3.

Si usamos el botón '**Save'** podemos guardar nuevas variables en archivos de datos en los que podremos determinar la clase de *cluster* de cada objeto (**Cluster Membership**) y la distancia al centro del cluster de cada uno de los objetos (**Distance from Cluster Center)**.

Figura 4.

El botón de opciones '**Options'** ofrece estadísticas adicionales -centros de cluster iniciales (**Initial cluster centers**), tablas de análisis de dispersión (**ANOVA table**) e información sobre las clases de *cluster* de cada objeto(**Cluster information in each case**). Lo aconsejable es que las tres opciones estén seleccionadas. Obtendremos el resultado final haciendo click sobre el OK.

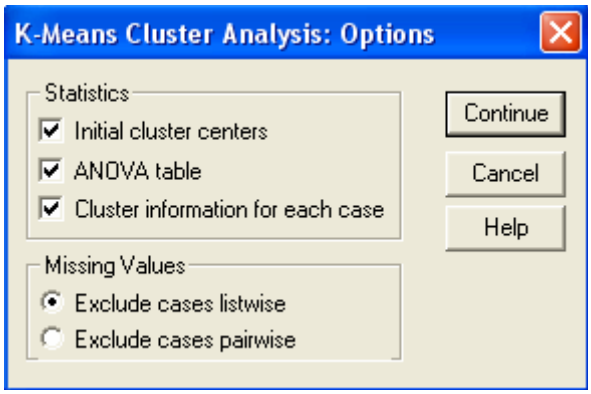

Figura 5.

Veamos un ejemplo de análisis de cluster con el método de las K-medias usando los datos que aparecen en el ejemplo UniCredit Bulbank (Tabla 1ª del capítulo *Primeros pasos en* SPSS). En primer lugar indicamos que el número de cluster es 4 y que los centros inciales de cluster son calculados usando los datos del archivo. Utilizaremos la distancia euclídea al cuadrado para mediar la divergencia entre las unidades. Igualmente escogeremos los centros de cluster sean calculados después de que se hayan clasificado todos los objetos en cada uno de los cluster definidos, i.e. no marcaremos la casilla '**Use running means'.**

Los centros de *cluster* iniciales aparecen en la Tabla 1 (**Initial Cluster Centers**). Son vectores cuyos valores se basan en las cinco variables: la referida al 2000 (primer *cluster*), 2005 (segundo), 2006 (tercer *cluster*) y 2003 (cuarto *cluster*) Estos cuatro años están en la distancia de índice máximo con los otros.

Tabla 1.

Centros de *cluster* iniciales

	Cluster			
	1	2	3	4
Net profit	160,065	96,116	120,654	89,752
Own funds	602,776	609,609	630,781	550,026
Assets	2559,476	3474,829	4346,594	2825,439
Client deposits	1692,270	2618,771	3336,875	2177,781
Loans	316,380	1706,858	2131,577	916,634

En la Tabla 2 podemos observar el número de iteraciones y cambios que se producen en los centros de *cluster*. En la primera iteración el año 2001 se une al 2000 y el centro de

cluster se actualiza. El año 2004 se une en el segundo *cluster* al año 2003, mientras que el 2005 lo hace en el cuarto cluster al *2002*. El tercer *cluster* no sufre ninguna modificación. En la segunda iteración el proceso de redistribución de las unidades se paraliza porque los centro de *cluster* no experimentan cambios.

Tabla 2.

Historia de las Iteraciones (a)

Iteration	Change in Cluster Centers			
	1	2	3	4
1	200,730	227,959	,000	195,515
2	,000	,000	,000	,000

(a) La convergencia se consigue gracias a que o bien no existe o hay un pequeño cambio en los centros de los *cluster*. El máximo cambio absoluto de coordenadas para cada centro de cluster es ,000. La actual iteración es la 2ª. El máximo de distancia entre los centros es de 821,273.

Los resultados se resumen en la Tabla 3, i.e. qué *cluster* pertenece a cada unidad y los nuevos centros de *cluster*. El primer *cluster* se forma en los años 200 y 2001, el segundo en el 2004 y 2005, en el tercero está el 2006 y en el cuarto se incluyen los años 2002 y 2003. En la Tabla 4 podemos ver los centros de *cluster* finales y en la Tabla 5 la distancia entre esos centros finales de *cluster*.

Tabla 3.

Socios del *Cluster*

Case Number	Cluster	Distance
1: 2000	1	200,730
2: 2001	1	200,730
3: 2002	4	195,515
4: 2003	4	195,515
5: 2004	2	227,959
6: 2005	2	227,959
7: 2006	3	,000

Tabla 4.

Centros Finales de *Cluster*

	Cluster			
	1	2	3	4
Net profit	114,489	91,198	120,654	84,441
Own funds	546,628	591,861	630,781	531,638
Assets	2645,581	3544,763	4346,594	2773,710
Client deposits	1856,952	2767,970	3336,875	2113,869
Loans	339,367	1550,413	2131,577	740,285

Tabla 5.

Distancias entre los Centros Finales de *Cluster*

Cluster	1	2	3	4
1		1762,868	2881,450	494,253
2	1762,868		1143,119	1297,055
3	2881,450	1143,119		2432,395
4	494,253	1297,055	2432,395	

Si comparamos estos resultados con los de la Tabla 1 y la Tabla 4 observaremos que el centro *cluster* del tercer *cluster* no cambia.

Los grupos se han formado deliberadamente de acuerdo con la distancia entre ellos en el espacio multidimensional, i.e. la condición para la aleatoriedad de las observaciones en los diferentes grupos no se halla, los resultados del análisis de dispersión son puramente descriptivos. Es decir, no podemos usar el nivel relevante (columna 'Sign' de la Tabla ANOVA - análisis de dispersión de los resultados de clasificación) para verificar la hipótesis de las principales variables. Sin embargo, las diferencias entre las F-ratios (columna F en la Tabla ANOVA) hace posible extraer conclusiones generales sobre el comportamiento de las principales variables en la formación de los *clusters*.

En la Tabla 6 tenemos los resultados del análisis de dispersión. Muestra que los activos ejercen la mayor influencia en la formación de los clusters mientras que las ganancias netas son las que menos influyen.

Tabla 6.

ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Net profit	495,145	3	1419,744	3	,349	,795
Own funds	2878,202	3	2537,200	3	1,134	,460
Assets	842788,443	3	9987,138	3	84,387	,002
Client deposits	634017,636	3	35643,498	3	17,788	,021
Loans	957411,333	3	37401,709	3	25,598	,012

El test F debe usarse solo para describir propósitos ya que los *clusters* han sido escogidos para maximizar las diferencias entre los casos en los diferentes clusters. Por esto el nivel de relevancia observado no está corregido y por lo tanto no podemos interpretarlos como test de las hipótesis que indican que los clusters principales son iguales.

Tabla 7.

Número de Casos en cada *Cluster*

Cluster	1	2,000
	2	2,000
	3	1,000
	4	2,000
Valid		7,000
Missing		,000

La Tabla 7 presenta los datos del número de unidades que hay en cada cluster así como el número total y los elementos perdidos (si hay alguno).

Ahora presentaremos los resultados con el mismo procedimiento de clasificación, con la diferencia de que escogeremos que los centros de cluster sean modificados después de que se una cada elemento a un cluster dado. Seleccionamos la opción '**Use running means'**.

Tabla 8.

Historial de las iteraciones (a)

	Change in Cluster Centers			
Iteration	1	2	3	4
1	215,142	151,973	,000	,000
2	53,786	50,658	,000	,000
3	13,446	16,886	,000	,000
4	3,362	5,629	,000	,000
5	,840	,1876	,000	,000
6	,210	,625	,000	,000
7	,053	,208	,000	,000
8	,013	,069	,000	,000
9	,003	,023	,000	,000
10	,001	,008	,000	,000

Tabla 9.

Socios *Cluster*

Case Number	Cluster	Distance
1: 2000	1	286,856
2: 2001	1	140,021
3: 2002	1	206,434
4: 2003	4	,000
5: 2004	2	227,963
6: 2005	2	227,955
7: 2006	3	,000

(a) Las iteraciones se paran porque se ha establecido el número máximo de ellas. Las iteraciones fallan al converger**.** El máximo cambio absoluto para cada centro es ,005. La actual iteración es 10. La distancia minima entre los centros iniciales es de 821,273.

Tabla 10.

Centros Finales de *Clusters*

	Cluster			
	1	2	3	4
Net profit	102,702	91,198	120,654	89,752
Own funds	535,501	591,861	630,781	550,026
Assets	2671,047	3544,763	4346,594	2825,439
Client deposits	1921,287	2767,970	3336,875	2177,781
Loans	414,223	1550,413	2131,577	916,634

Tabla 11.

Distancias entre los Centros Finales de *Clusters*

Cluster	1	2	3	4
1		1665,679	2787,481	585,168
2	1665,679		1143,119	1126,578
3	2787,481	1143,119		2267,372
4	585,168	1126,578	2267,372	

Tabla 12.

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Net profit	236,122	3	1678,767	3	,141	,929
Own funds	2856,043	3	2559,359	3	1,116	,465
Assets	843275,336	3	9500,245	3	88,764	,002
Client deposits	628462,814	3	41198,320	3	15,255	,025
Loans	966937,206	3	27875,836	3	34,687	,008

El test F debe usarse solo para describir propósitos ya que los *clusters* han sido escogidos para maximizar las diferencias entre los casos en los diferentes clusters. Por esto el nivel de relevancia observado no está corregido y por lo tanto no podemos interpretarlos como test de las hipótesis que indican que los clusters principales son iguales.

Tabla 13.

Número de Casos en cada *Cluster*

Cluster	1	3,000
	2	2,000
	3	1,000
	4	1,000
Valid		7,000
Missing		,000

De los datos presentados en la Tabla 9 podemos observar que ahora el primer *cluster* se forma en los años 2000, 2001 y 2002, el segundo en el 2004 y 2005, el tercero en el 2006 y el cuarto en el 2003. Según los datos que aparecen en la tabla ANOVA los activos vuelven a ser los que ejercen la mayor influencia para la formación de los *clusters* y las ganancias netas las que menos.

Autora: Dessislava Vojnikova,
Plovdiv University,
4th year Bachelor program in Applied Mathematics

Supervisora: Snezhana Gocheva-Ilieva