Análisis Cluster

1. Objetivos y Competencias a alcanzar

- Saber obtener matrices de distancias de cualquier tipo de datos originales.
- Conocer los fundamentos de los algoritmos de Análisis Clúster Jerárquicos: Método de Ward, del vecino más cercano, del vecino más alejado, del centroide y de la media.
- Conocer los fundamentos de los algoritmos de Análisis Cluster No Jerárquico.
- Conocimiento de técnicas que permitan justificar el número de agrupaciones finalmente obtenidas así como la no presencia de cluster en los datos, si así fuera.
- Obtención e interpretación de las agrupaciones generadas por cualquiera de los dos métodos.

2. Introducción

El Análisis Cluster (o de conglomerados) tiene como objetivo formar grupos de individuos con características similares con respecto a determinadas variables. Para ello, se cuenta con una matriz de datos $\boldsymbol{X}$ de dimensión $n \times m$ cuyas filas y columnas representan las observaciones y las variables, respectivamente. La diferencia fundamental con respecto al análisis discriminante es que no se conocen de antemano el número de grupos en los que se divide a la población, ni el valor de la variable que identifica cada grupo. Es por ello que el análisis cluster también recibe el nombre de "clasificación no supervisada" al no existir una muestra de elementos previamente clasificados en grupos que sirva de pauta como en el caso del análisis discriminante (también llamado "clasificación supervisada").

La idea básica es, a partir de un conjunto de individuos, crear grupos excluyentes y exhaustivos tales que:

- Los individuos de cada grupo deben ser lo más parecidos que sea posible (homogeneidad interna).
- Los grupos deben ser lo más diferentes que sea posible (heterogeneidad entre grupos).

A la hora de preparar los datos, es frecuente que las variables vengan en diferentes unidades de medida. En ese caso conviene normalizarlas, aunque no se debe abusar de esta técnica puesto que al homogeneizar la varianza de todas las variables podemos mermar la "capacidad clasificatoria" de alguna variable con gran variabilidad natural.

Otra circunstancia no deseable es que las variables se encuentren correladas. En ese caso será preferible recurrir previamente a alguna técnica como el análisis factorial o el análisis de componentes principales que sintetice la información, proporcionándonos variables incorreladas. Además es conveniente corregir el problema de los atípicos ya que distorsionarían la generación de cluster.

Existen dos tipos de análisis cluster: jerárquico y no jerárquico. En el análisis no jerárquico se ha de definir el número de grupos a crear mientras que en el jerárquico se construye una especie de jerarquía en función de la similitud/distancia de los datos y se obtiene una posible clasificación para cualquier número de grupos entre uno y el tamaño de la muestra.

3. Análisis Cluster jerárquico

A la hora de realizar análisis cluster jerárquico hay que definir cuidadosamente la forma de medir las distancias entre las observaciones. Cuando se trabaja con variables contínuas se suele trabajar con distancias euclídeas, pero también es posible recurrir a otras distancias como la de Mahalanobis, la de Minkoswski o la de Tchebychev. Cuando las variables que se utilizan son binarias u ordinales, se puede recurrir a otro tipo de distancias como la de Gower (no obstante, en este curso trataremos sólo con variables contínuas y nos centraremos en las distancias euclídeas por defecto).

Como ya se ha mencionado previamente, los métodos de clasificación jerárquica no producen una clasificación en un número determinado de clusters en un único paso, sino que configuran grupos con estructura arborescente de forma que clusters de niveles más bajos van siendo englobados en otros de niveles superiores.

Los pasos a seguir para realizar un Análisis Cluster Jerárquico son los siguientes:

i) Se parte de tantos grupos como observaciones.

ii) Se genera una matriz de dimensión $n \times n$ que indique las distancias entre todos los pares de observaciones (esta distancia debe haber sido definida con anterioridad).

iii) Se agrupan las dos observaciones (o clusters) más próximas. Con esto, el número de clusters existentes es uno menos que en el paso anterior.

iv) Se vuelve a obtener una matriz de distancias con los clusters formados en el paso anterior. Obsérvese que para obtener esta nueva matriz, es necesario elegir un método de cálculo de distancias entre clusters.

v) Repetir los pasos iii) y iv) hasta que todas las observaciones están agrupadas en un solo cluster.

Observemos que es necesario definir con claridad, tanto la distancia entre observaciones, como la distancia entre clusters o grupos de observaciones. Por lo tanto, bajo este esquema y con un mismo conjunto de datos, variando esas dos definiciones se podrán obtener múltiples clasificaciones diferenciadas.

3.1. Métodos para calcular la distancia entre dos clusters

Supongamos que tenemos una agrupación denominada C_k formada por las observaciones $x_{k,1}, \ldots, x_{k,n_k}$ y otra agrupación C_ℓ formada por las observaciones $x_{\ell,1}, \ldots, x_{\ell,n_\ell}$ los métodos más comunes para calcular las distancias entre clusters son:

- Enlace Simple o del vecino más cercano: la distancia entre dos clusters viene dada por la distancia mínima entre observaciones de distintos grupos, o en otras palabras, la distancia entre las observaciones más cercanas pertenecientes a distintos grupos:

$$D_{k,\ell} = \min_{\substack{i \in C_k \\ j \in C_\ell}} dist(x_i, x_j)$$

Tiende a crear grupos con muchas observaciones y alargados, que pueden incluir elementos muy distintos en los extremos.

- Enlace Completo o del vecino más alejado: la distancia entre dos clusters viene dada por la distancia máxima entre observaciones de distintos grupos, o en otras palabras, la distancia entre las observaciones más alejadas pertenecientes a distintos grupos:

$$D_{k,\ell} = \max_{\substack{i \in C_k \\ j \in C_\ell}} dist(x_i, x_j)$$

  Los grupos obtenidos con este método son más compactos que los obtenidos con el método del vecino más próximo.

- Enlace medio: la distancia entre dos clusters viene dada por la distancia media entre observaciones de distintos grupos:

$$D_{k,\ell} = \frac{1}{n_k n_\ell} \sum_{\substack{i \in C_k \\ j \in C_\ell}} dist(x_i, x_j)$$

  Los grupos así formados tienen varianza similar y pequeña.

- Distancia entre centroides: la distancia entre dos clusters viene dada por la distancia entre los centroides de cada grupo, que representarán el vector medio obtenido en las m variables para todos los individuos que formen parte del grupo:

$$D_{k,\ell} = dist(\bar{x_k}, \bar{x_\ell})$$

- Distancia de Ward o de la mínima varianza: en este caso, la distancia a minimizar viene dada por:

$$D_{k,\ell} = \frac{n_k n_\ell}{n_k + n_\ell} (\bar{x_k} - \bar{x_\ell})'(\bar{x_k} - \bar{x_\ell})$$

  Esta distancia permite seleccionar, entre todas las uniones de cluster posibles, aquella unión que minimiza la variabilidad interna de los cluster resultantes:

$$W = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (x_{k,i} - \bar{x_k})'(x_{k,i} - \bar{x_k})$$

  Los valores de W pueden ser de utilidad a la hora de determinar el número de clusters. Este método tiende a generar conglomerados pequeños y equilibrados en tamaño.

Ahora parece lógico preguntarse: ¿Cómo determinar en cada caso cuál es el método de agrupación más adecuado? No existe una respuesta exacta a esta pregunta, aunque los tres últimos son los más utilizados. Siempre es conveniente estudiar varios métodos y tomar una decisión en función de los resultados que se obtengan. Si varios métodos nos dan agrupaciones similares, se puede pensar que existe una forma natural de formarse grupos de observaciones.

Ejemplo 1 *En este ejemplo veremos como aplicar los tres primeros métodos en la realización de un análisis cluster. Supóngase que contamos con 4 datos A, B, C y D tales que*

su distancia (para aplicar los dos últimos métodos es necesario contar con el valor de las observaciones) está contenida en la siguiente matriz:

	A	B	C	D
A	0	1	4	2,5
B	1	0	2	3
C	4	2	0	4
D	2,5	3	4	0

- *Enlace Simple: El valor mínimo fuera de la diagonal es* 1 *por lo que deberemos unir* A *y* B*. A continuación, debemos recalcular las distancias:*

$$\left.\begin{array}{l} d(AB,C) = \text{mín}\{dist(A,C), dist(B,C)\} = \text{mín}\{4,2\} = 2 \\ d(AB,D) = \text{mín}\{dist(A,D), dist(B,D)\} = \text{mín}\{2{,}5,3\} = 2{,}5 \end{array}\right\}$$

	AB	C	D
AB	0	2	2,5
C	2	0	4
D	2,5	4	0

A continuación, la distancia mínima es 2*, por lo que debemos unir* AB *y* C*. La distancia de este cluster a la observación* D *es:*

$$d(ABC,D) = \text{mín}\{dist(AB,D), dist(C,D)\} = \text{mín}\{2{,}5,4\} = 2{,}5$$

Finalmente, agregamos D *al cluster* ABC*. Es importante conservar el orden y las distancias derivadas del análisis para obtener el dendrograma, que es un gráfico que explicaremos en la siguiente sección.*

- *Enlace Completo: El valor mínimo fuera de la diagonal es* 1 *por lo que de nuevo unimos* A *y* B*. A continuación, debemos recalcular las distancias:*

$$\left.\begin{array}{l} d(AB,C) = \text{máx}\{dist(A,C), dist(B,C)\} = \text{máx}\{4,2\} = 4 \\ d(AB,D) = \text{máx}\{dist(A,D), dist(B,D)\} = \text{máx}\{2{,}5,3\} = 3 \end{array}\right\}$$

	AB	C	D
AB	0	4	3
C	4	0	4
D	3	4	0

A continuación, la distancia mínima es 3*, por lo que debemos unir* AB *y* D*. La distancia de este cluster a la observación* C *es:*

$$d(ABD,C) = \text{máx}\{dist(AB,C), dist(D,C)\} = \text{máx}\{4,4\} = 4$$

Finalmente, agregamos C *al cluster* ABD*.*

- *Enlace medio: El valor mínimo fuera de la diagonal es* 1 *por lo que de nuevo unimos* A *y* B*. A continuación, debemos recalcular las distancias:*

$$\left.\begin{array}{l} d(AB,C) = \frac{1}{2}(dist(A,C) + dist(B,C)) = \frac{1}{2}(4+2) = 3 \\ d(AB,D) = \frac{1}{2}(dist(A,D) + dist(B,D)) = \frac{1}{2}(2{,}5+3) = 2{,}75 \end{array}\right\}$$

	AB	C	D
AB	0	3	2,75
C	3	0	4
D	2,75	4	0

A continuación, la distancia mínima es 2,75*, por lo que debemos unir* AB *y* D*. La distancia de este cluster a la observación* C *es:*

$$d(ABD,C) = \frac{1}{3}(dist(A,C) + dist(B,C) + dist(D,C)) = \frac{1}{3}(4+2+4) = \frac{10}{3}$$

Finalmente, agregamos C *al cluster* ABD*.*

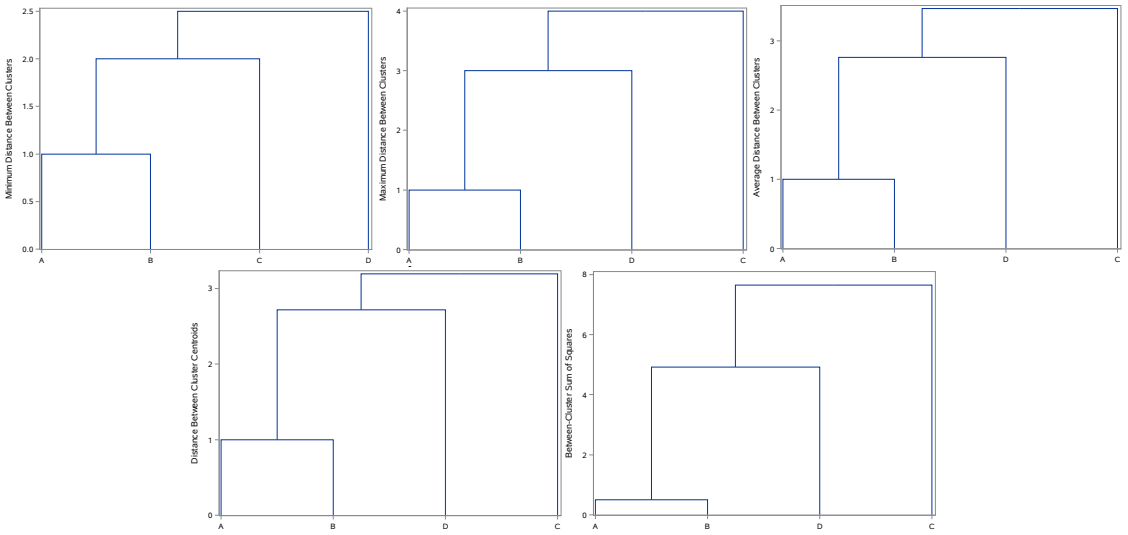

Figura 1: Dendrogramas del análisis cluster jerárquico sobre los datos del Ejemplo 1 para distintos tipos de distancias

3.2. El Dendrograma

El dendrograma es el gráfico más frecuente en análisis cluster jerárquico, pues permite plasmar el proceso de aglomeración y formación de grupos junto con la distancia entre cada dos grupos unidos en una gráfica.

El dendrograma se construye como sigue:

1. En la parte inferior del gráfico se disponen los n elementos iniciales.

2. Las uniones entre elementos se indican a partir de tres líneas rectas. Dos dirigidas a los elementos que se unen, y que son perpendiculares al eje de los elementos, y una paralela a este eje, que se sitúa al nivel que se unen (este nivel puede representar la distancia o algún otro estadístico que veremos más adelante).

3. El proceso se repite hasta que todos los elementos estén conectados por estas líneas rectas.

Este diagrama nos puede ayudar a determinar en qué momento del proceso de agrupación nos deberemos detener pues, si cortamos el dendrograma a un nivel dado, obtenemos una clasificación del número de grupos existentes a ese nivel y los elementos que los forman.

Ejemplo 2 *La Figura 1 muestra los dendrogramas obtenidos al aplicar análisis cluster jerárquico sobre los datos del Ejemplo 1 utilizando los 5 métodos explicados: enlace simple, enlace completo, enlace medio, distancia entre centroides y distancia de Ward. Nótese que el eje vertical no representa lo mismo en los cinco gráficos, sino que se representa la medida empleada para la realización de los clusters. Como ya se vió en el ejemplo 1, el orden no coincide para todos los métodos.*

3.3. Procedimientos para determinar el número adecuado de grupos

Después de llevar a cabo análisis Cluster, el paso siguiente es determinar el número real de grupos existentes en nuestros datos. Recordemos que nuestro objetivo es formar grupos lo más homogéneos "dentro de sí" y lo más diferentes "entre sí". Para medir esto,

la variabilidad total de los datos la dividiremos en dos partes: la que mide la variabilidad dentro de los grupos y la que mide la variabilidad entre los grupos:

$$\text{Variabilidad total: } T = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x_i})^2$$

$$\text{Variabilidad dentro del cluster k: } W_k = \sum_{i=1}^{m}\sum_{j\in C_k}(x_{ij} - \bar{x_i}^k)^2$$

$$\text{Variabilidad total intra-clusters: } W = \sum_{k=1}^{g} W_k$$

$$\text{Variabilidad total entre-clusters: } E = \sum_{i=1}^{m}\sum_{k=1}^{g}(\bar{x_i}^k - \bar{x_i})^2$$

donde x_{ij} es la j-ésima observación de la variable i-ésima y $\bar{x_i}$ y $\bar{x_i}^k$ son la media de la variable i-ésima para el conjunto total de individuos y para el cluster k-ésimo, respectivamente. Nótese que $T = W + E$.

Obviamente, el número de grupos no puede estimarse a partir de un criterio de minimización de la variabilidad interna, ya que la forma de alzanzar este objetivo consiste en hacer tantos grupos como observaciones, con lo que $W_k = W = 0\ \forall k$.

Algunos indicadores que nos ayudarán en la decisión son:

- R^2: se define como la proporción de variabilidad explicada por los cluster creados y viene dada por:
$$R^2 = \frac{E}{T} = 1 - \frac{W}{T}$$

- R^2-semiparcial: se define como la diferencia entre la proporción de variabilidad explicada con ℓ cluster o con k ($\ell > k$).
$$SPRSQ = R^2_\ell - R^2_k$$

- Pseudo Test de la T: se basa en la idea de que si las medias de dos agrupamientos diferentes no son significativamente diferentes, entonces esos dos agrupamientos podrían combinarse sin que la variabilidad dentro del cluster resultante fuese significativamente superior, pero si la diferencia entre esas medias es significativa, entonces los agrupamientos no deben combinarse.

  Así, si en una fase se agrupan los cluster k y ℓ para formar el cluster m, dentro de este último cluster la dispersión interna sería mayor que la suma de las dispersiones de los dos cluster k y ℓ de forma aislada.

$$Pseudo - T^2 = \frac{W_m - W_k - W_\ell}{\frac{W_k + W_\ell}{n_k + n_\ell - 2}}$$

  En la práctica, buscaremos cuando se produce un máximo relativo o un incremento excesivo (supongamos que para r cluster) y en ese caso se aconsejaría rechazar el agrupamiento en esa fase y recomendar la clasificación en $r + 1$ cluster.

- Pseudo F: este criterio compara la dispersión entre cluster con la dispersión dentro de los cluster. Lo que se pretende es que este cociente sea máximo. Por esta razón,

buscaremos máximos relativos o incrementos importantes del valor de este pseudo-estadístico. El cálculo del mismo viene dado por:

$$Pseudo - F = \frac{\frac{E}{g-1}}{\frac{W}{n-g}}$$

- Criterio Cúbico (CCC): este criterio, introducido por Searle, se aplica sólo si los datos se extraen de coordenadas y no es apropiado si el método es el enlace simple. Se puede representar gráficamente frente al número de agrupamientos para diversas selecciones de este último. Como con el pseudo-F, los picos o máximos relativos corresponden a números apropiados de agrupamientos, siempre que este índice sea superior a 2. Si toma valores negativos, existen indicios para pensar que hay observaciones atípicas. Y si los valores máximos se encuentran entre 0 y 2 las soluciones hay que tomarlas con cautela.

4. Análisis Cluster no jerárquico

Como ya se ha comentado, en el análisis cluster no jerárquico es necesario fijar de antemano el número de grupos en que se pretende dividir las observaciones. Las técnicas de agrupación siguen básicamente los siguientes pasos:

1. Seleccionar g observaciones como centroides iniciales de los clusters a construir, siendo g el número deseado de clusters.
2. Asignar cada una de las observaciones restantes al cluster más próximo.
3. Reasignar cada observación a uno de los g clusters de acuerdo con una regla de parada determinada previamente.
4. Parar si no se reasignan observaciones a un grupo distinto del de partida, o si la reasignación satisface la regla de parada. En caso contrario, volver a 2.

Los métodos para realizar agrupaciones se diferencian entre sí por el modo de escoger los centroides iniciales y por el criterio empleado para reasignar observaciones a los distintos grupos. Algunos de los métodos utilizados para obtener los centroides iniciales son:

- Seleccionar las g primeras observaciones con datos no-missing.
- Seleccionar la primera observación como primer centroide. El segundo centroide será aquella observación cuya distancia al primer centroide sea tan grande como una distancia seleccionada previamente. El tercer centroide, la observación cuya distancia a los dos primeros centroides sea mayor que la distancia seleccionada. Y así sucesivamente. Es el método empleado en SAS por defecto.
- Seleccionar aleatoriamente g observaciones con datos conocidos.
- Elegir centroides que estén entre sí lo más lejanos posible.
- Utilizar centroides que proporcione el investigador.

Una vez que se han identificado los centroides, se pueden formar los Clusters iniciales asignando cada una de las $n - g$ observaciones restantes al cluster correspondiente al centroide más próximo.

En cuanto a la forma de reasignar observaciones, algunas de las reglas más utilizadas son las siguientes:

- Calcular el centroide de cada cluster y reasignar sujetos a los clusters cuyo centroide es el más cercano. Los centroides no varían mientras se reasignan observaciones, sino que se recalculan después de hacer la nueva asignación de todas las observaciones. Si el cambio en el centroide es mayor que el valor determinado por un criterio de convergencia, se vuelve a hacer una reasignación de observaciones, y se vuelven a calcular los centroides. El proceso de reasignación continúa hasta que el cambio en los centroides es menor que el valor dado por el criterio de convergencia.

- Calcular el centroide de cada grupo y reasignar sujetos al cluster cuyo centroide es el más cercano. Para la asignación de cada observación se recalcula el centroide del cluster al que se asigna la observación y el centroide del cluster del que proviene la observación. La reasignación continúa hasta que la diferencia entre centroides es menor que el valor dado por el criterio de convergencia.

Por lo tanto, combinando los distintos métodos de selección de centroides iniciales y de reasignación de observaciones, se pueden desarrollar un gran número de algoritmos de clusters no jerárquicos.

4.1. Procedimientos para determinar el número adecuado de grupos

Algunos de los procedimientos vistos para análisis jerárquico, como el Pseudo-F o el CCC, son también válidos para análisis cluster no jerárquico. Otro procedimiento útil para este propósito es el test F de Beale que nos permitirá decidir entre dos agrupaciones con distinto número de grupos.

Si estamos decidiendo entre dos agrupaciones, una de ellas con c_1 grupos y otra con c_2 grupos ($c_1 > c_2$), sabemos que la ordenación con mas grupos será más homogénea (la suma de cuadrados de los errores dentro de los grupos será inferior). A cambio, el hecho de tener más grupos dificultará su interpretabilidad (principio de parsimonia). Por esta razón, es conveniente comprobar si el incremento de la suma de cuadrados de los errores dentro de los grupos no es muy grande, en cuyo caso compensará quedarnos con c_2 grupos.

Sean W_{c_1} y W_{c_2} las sumas de los cuadrados de las distancias entre cada observación y el centroide del grupo en el que ha sido asignada (tal y como se definieron para el análisis jerárquico). El estadístico F^* de Beale se calcula como:

$$F^* = \frac{W_{c_2} - W_{c_1}}{W_{c_1}} \frac{(n - c_1)c_1^{-\frac{2}{m}}}{(n - c_2)c_2^{-\frac{2}{m}} - (n - c_1)c_1^{-\frac{2}{m}}}$$

y se distribuye como una ley $F_{(n-c_2)c_2^{-\frac{2}{m}}-(n-c_1)c_1^{-\frac{2}{m}},(n-c_1)c_1^{-\frac{2}{m}}}$. Por lo que si F^* es mayor que el punto crítico de la F con esos grados de libertad, entonces se elegiría la agrupación con mayor número de grupos ya que las diferencias de la variabilidad interna serían significativas.

5. Caracterización de los clusters

Una vez que se ha decidido la partición de los clusters, el siguiente paso consistirá en caracterizarlos. Esta caracterización se debe realizar en dos sentidos:

- Por un lado, se debe realizar un análisis descriptivo sobre las variables activas utilizadas en el análisis, con lo que se determinarán las medias y varianzas de todas las variables. Ello nos permitirá una primera caracterización.
- Por otro lado, utilizaremos otras variables (suplementarias) que pueden ser categóricas o continuas y que nos permitirán explicar las variabilidades en cada grupo en base a otro tipo de criterios (sociodemográficos, físicos, etc.). En el caso de que sean categóricas evaluaremos la proporción de cada categoría presente en cada agrupación.

De forma conjunta se pueden examinar todas las características utilizando análisis discriminante, siendo el cluster al que pertenece la variable dependiente y utilizando el resto como explicativas o discriminantes.

También se puede aplicar análisis de correspondencias con respecto a las variables nominales (o a combinaciones de variables nominales) siendo las filas o las columnas los cluster generados.

Todo lo anterior nos dará una explicación verosímil del trabajo de clasificación realizado. Sin esta última parte el análisis quedaría muy incompleto.

6. Sistemática del Análisis Cluster

1. Análisis de los datos y detección de datos faltantes.
2. Elección de la medida de distancia entre observaciones a utilizar.
3. Decisión del tipo de análisis a realizar: Jerárquico o No Jerárquico.

En caso de Análisis Jerárquico:

4. Elección del método de agrupación de observaciones: Centroide, Ward, etc.
5. Decisión del número de grupos o Clusters a formar en función del valor de los estadísticos para los distintos números de grupos.
6. Examen de las características de los individuos de cada grupo: examen de las variables de manera individual, o de manera global a través de otros métodos multivariantes.
7. Caracterización de los distintos grupos formados.

En caso de Análisis No Jerárquico:

4. Decisión del número de grupos a formar.
5. Elección del algoritmo de agrupación.
6. Evaluación de la bondad del número de grupos seleccionado.
7. Examen de las características de los individuos de cada grupo: examen de las variables de manera individual, o de manera global a través de otros métodos multivariantes.
8. Caracterización de los distintos grupos formados.

7. Ejemplos resueltos con SAS

7.1. Ejemplo distancia entre ciudades

Se trata de un ejemplo sencillo que contiene la matriz de distancias en millas entre diferentes ciudades de Estados Unidos. En primer lugar, deberemos leer el conjunto de datos e identificarlo como distancias (opción *Type=distance* de la sentencia data):

```
Data america (type=distance);
input (atlanta chicago denver houston losangeles
       miami newyork sanfrancisco seattle washington) (5.) ciudad $15.;
 FORMAT CIUDAD $15.;
 cards;
0                                                    ATLANTA
587    0                                             CHICAGO
1212  920    0                                       DENVER
701   940  879    0                                  HOUSTON
1936 1745  831 1374    0                            LOSANGELES
604  1188 1726  968 2339    0                       MIAMI
748  713  1631 1420 2451 1092    0                  NEWYORK
2139 1858  949 1645  347 2594 2571    0            SANFRANCISCO
2182 1737 1021 1891  959 2734 2408  678    0        SEATTLE
543  597  1494 1220 2300  923  205 2442 2329    0   WASHINGTON
;
PROC PRINT;RUN;
```

En el programa anterior, llama la atención la sentencia *input(atlanta chicago ... washington)(5.)*. Esto nos permite poner el mismo formato de lectura a un conjunto de variables si tanto éstas como aquel van entre paréntesis y se acompañan de forma consecutiva. La sentencia *format* provoca no sólo que ciudad lea 15 caracteres (formato de lectura) sino que también los escriba (formato de escritura).

Dado que hemos indicado que el conjunto de datos es tipo distancia, SAS interpreta que los datos faltantes deben completarse con la información simétrica. A continuación procedemos a realizar análisis jerárquico con los métodos del centroide y de WARD. Las Figuras 2, 3 y 5 contienen las tablas de resultados y los dendrogramas para ambos métodos.

El procedimiento para realizar análisis cluster jerárquico en SAS es el *proc cluster*. Si los datos forman una matriz de distancias, debemos indicarlo. Debemos indicar también el método con *method=*. *nonorm* evita la normalización de las distancias, *pseudo* y *RSQUARE* indican que se muestre la pseudo-T^2 y la pseudo-F y el R^2, respectivamente y *plots=den(VERTICAL)* permite obtener un dendrograma vertical.

```
proc cluster data=america(type=distance) method=centroid nonorm pseudo
RSQUARE plots=den(VERTICAL) outtree=salida;
   id ciudad;
run;
```

Cluster History									
Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Pseudo F Statistic	Pseudo t-Squared	Centroid Distance	Tie
9	NEWYORK	WASHINGTON	2	0.0019	.998	66.7	.	205	
8	LOSANGELES	SANFRANCISCO	2	0.0054	.993	39.2	.	347	
7	ATLANTA	CHICAGO	2	0.0153	.977	21.7	.	587	
6	CL7	CL9	4	0.0296	.948	14.5	3.4	577.18	
5	CL8	SEATTLE	3	0.0391	.909	12.4	7.3	812.15	
4	DENVER	CL5	4	0.0475	.861	12.4	2.1	843.34	
3	CL6	MIAMI	5	0.0586	.803	14.2	3.8	907.49	
2	CL3	HOUSTON	6	0.0687	.734	22.1	2.6	962.45	
1	CL2	CL4	10	0.7339	.000	.	22.1	1853.7	

Figura 2: Tabla de resultados para el método del centroide

Cluster History									
Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Pseudo F Statistic	Pseudo t-Squared	Between Cluster Sum of Squares	Tie
9	NEWYORK	WASHINGTON	2	0.0019	.998	66.7	.	21013	
8	LOSANGELES	SANFRANCISCO	2	0.0054	.993	39.2	.	60205	
7	ATLANTA	CHICAGO	2	0.0153	.977	21.7	.	172285	
6	CL7	CL9	4	0.0296	.948	14.5	3.4	333134	
5	DENVER	HOUSTON	2	0.0344	.913	13.2	.	386321	
4	CL8	SEATTLE	3	0.0391	.874	13.9	7.3	439720	
3	CL6	MIAMI	5	0.0586	.816	15.5	3.8	658824	
2	CL3	CL5	7	0.1488	.667	16.0	5.3	1.67E6	
1	CL2	CL4	10	0.6669	.000	.	16.0	7.49E6	

Figura 3: Tabla de resultados para el método de Ward

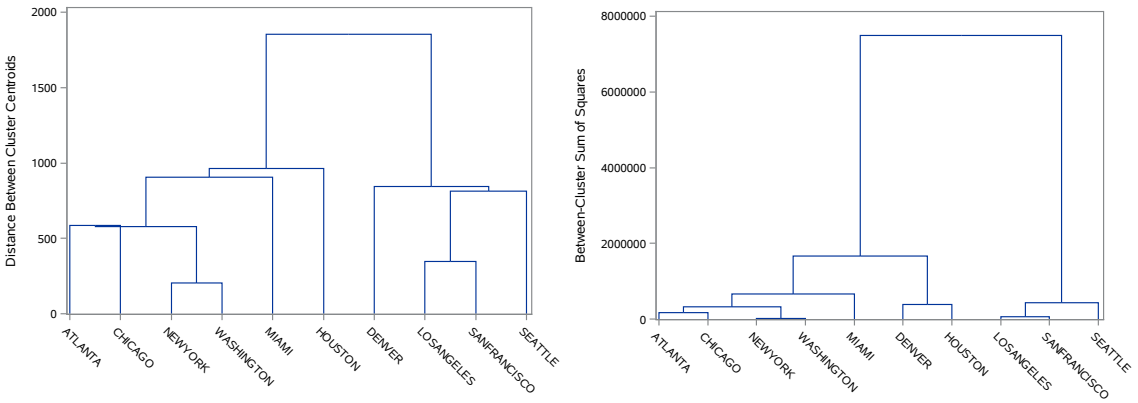

Figura 4: Dendrogramas para los métodos del centroide (izquierda) y de Ward (derecha)

Las Figuras 2 y 3 contienen los resultados del análisis. En particular, se muestra en qué orden se han realizado las agrupaciones y los estadísticos asociados a esas iteraciones del algoritmo. En el caso del método del centroide la secuencia de uniones es:

- En primer lugar se unen New York y Washington formando un cluster de frecuencia 2, y cuya distancia es 205. No se han producido empates (columna tie);
- en segundo lugar se unen Los Angeles y San Francisco;
- en tercer lugar se unen Atlanta y Chicago;
- en cuarto lugar se unen los cluster formados por (atlanta y chicago)=CL7 y (New York, Washington)=CL9. Siendo la distancia al centroide de estos dos cluster de 577,18; y así sucesivamente.

Podemos comprobar que la secuencia es similar (aunque no igual) con ambos métodos. Si nos fijamos en los valores de la pseudo-T^2 y la pseudo-F, en ambos casos concluimos que el número de grupos ha de ser 2. Además, en ambos casos la proporción de variabilidad explicada ronda el 70 %.

Los dendrogramas de la Figura 4 muestran de nuevo que la secuencia de uniones es similar. Además, la distancia entre las líneas horizontales también nos indica que el número de grupos es 2: coincidiendo con una división este/oeste de las mismas. No obstante, los métodos no coinciden en lo que a "Denver" se refiere. El método de los centroides se basa más en las distancias reales que el de Ward, por lo que asigna esta ciudad a la costa oeste (a la que geográficamente está más cercana). Sin embargo, el método de Ward busca minimizar la variabilidad dentro de los grupos y es por ello que el resultado puede resultar extraño desde el punto de vista geográfico. Sin embargo, nótese que a la vista del dendrograma, sería posible incluso crear 3 grupos: oeste, centro y este, lo que lograría una mayor homogeneidad de los grupos.

A la vista de los resultados, podemos comprobar que el análisis cluster carece de una solución única y que la misma depende del caso de aplicación. No obstante, suponiendo que damos por válida la solución ofrecida por el método del centroide, es posible obtener un conjunto de datos en el que a cada observación se le asigne su cluster en una nueva variable. Para ello, es necesario crear un conjunto de datos con los resultados en el *proc cluster* a partir de la sentencia *outtree=*. A continuación, utilizamos la siguiente sentencia para obtener el conjunto de datos con la variable "cluster":

```
proc tree data=salida n=2 out=CiudadesClasif;
id ciudad;
run;
proc print data=CiudadesClasif;
run;
```

8. Ejemplo Esperanza de vida

En este ejemplo se considera una muestra de los años de vida esperados de varios paises según la edad y el sexo (estos datos proceden de Keyfitz y Flieger (1971)). Se desea agrupar estos paises según la información de dichas variables.

```
Data esperanza;
input pais \$ m0 m25 m50 m75 w0 w25 w50 w75;
datalines;
Algeria 63.00 51.00 30.00 13.00 67.00 54.00 34.00 15.00
Cameroon 34.00 29.00 13.00 5.00 38.00 32.00 17.00 6.00
Madagascar 38.00 30.00 17.00 7.00 38.00 34.00 20.00 7.00
Mauritius 59.00 42.00 20.00 6.00 64.00 46.00 25.00 8.00
Reunion 56.00 38.00 18.00 7.00 62.00 46.00 25.00 10.00
Seychelles 62.00 44.00 24.00 7.00 69.00 50.00 28.00 14.00
South_Africa 65.00 44.00 22.00 7.00 72.00 50.00 27.00 9.00
Tunisia 56.00 46.00 24.00 11.00 63.00 54.00 33.00 19.00
Canada 69.00 47.00 24.00 8.00 75.00 53.00 29.00 10.00
Costa_Rica 65.00 48.00 26.00 9.00 68.00 50.00 27.00 10.00
Dominican_Rep 64.00 50.00 28.00 11.00 66.00 51.00 29.00 11.00
El_Salvador 56.00 44.00 25.00 10.00 61.00 48.00 27.00 12.00
Greenland 60.00 44.00 22.00 6.00 65.00 45.00 25.00 9.00
Grenada 61.00 45.00 22.00 8.00 65.00 49.00 27.00 10.00
Guatemala 49.00 40.00 22.00 9.00 51.00 41.00 23.00 8.00
Honduras 59.00 42.00 22.00 6.00 61.00 43.00 22.00 7.00
Jamaica 63.00 44.00 23.00 8.00 67.00 48.00 26.00 9.00
Mexico 59.00 44.00 24.00 8.00 63.00 46.00 25.00 8.00
Nicaragua 65.00 48.00 28.00 14.00 68.00 51.00 29.00 13.00
Panama 65.00 48.00 26.00 9.00 67.00 49.00 27.00 10.00
Trinidad 64.00 43.00 21.00 6.00 68.00 47.00 24.00 8.00
United_States 67.00 45.00 23.00 8.00 74.00 51.00 28.00 10.00
Argentina 65.00 46.00 24.00 9.00 71.00 51.00 28.00 10.00
Chile 59.00 43.00 23.00 10.00 66.00 49.00 27.00 12.00
Columbia 58.00 44.00 24.00 9.00 62.00 47.00 25.00 10.00
Ecuador 57.00 46.00 28.00 9.00 60.00 49.00 28.00 11.00
;
PROC PRINT;RUN;
```

En primer lugar realizaremos un análisis cluster jerárquico para intentar determinar el mejor número de grupos a realizar. A continuación, realizaremos un análisis cluster no jerárquico para verificar mediante otras medidas la adecuación de ese número de clusters y comprobar si los resultados utilizando ambas técnicas son similares. Nótese que no es necesario estandarizar las variables puesto que la unidad de todas ellas es el año.

```
proc cluster data=esperanza method=average nonorm pseudo RSQUARE ccc
print=15 plots=den(VERTICAL);
   id pais;
run;
```

En este caso se ha incluido la opción *ccc* puesto que los datos vienen dados en coordenadas. Además, se ha pedido que se muestren sólo las últimas 15 agrupaciones. Sólo se

Cluster History											
Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Approximate Expected R-Square	Cubic Clustering Criterion	Pseudo F Statistic	Pseudo t-Squared	RMS Distance	Tie
10	CL13	Trinidad	5	0.0058	.958	.	.	40.3	2.3	7.3485	
9	CL12	Reunion	6	0.0084	.949	.	.	39.8	3.5	8.438	
8	CL9	CL10	11	0.0225	.927	.	.	32.6	6.4	9.0885	
7	CL14	CL15	8	0.0234	.903	.	.	29.6	10.6	9.4274	
6	CL8	CL19	13	0.0198	.884	.	.	30.4	3.9	10.142	
5	Algeria	Tunisia	2	0.0124	.871	.891	-1.0	35.5	.	12.124	
4	CL6	CL7	21	0.0964	.775	.858	-2.8	25.2	16.3	13.287	
3	CL5	CL4	23	0.0532	.722	.804	-1.6	29.8	5.1	16.369	
2	CL3	Guatemal	24	0.0826	.639	.672	-.56	42.5	6.6	24.088	
1	CL2	CL11	26	0.6390	.000	.000	0.00	.	42.5	46.347	

Figura 5: Tabla de resultados para el método del enlace medio

Cluster History											
Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Approximate Expected R-Square	Cubic Clustering Criterion	Pseudo F Statistic	Pseudo t-Squared	Centroid Distance	Tie
10	CL12	Honduras	9	0.0087	.950	.	.	33.7	2.7	7.5921	
9	Seychell	CL13	5	0.0081	.942	.	.	34.4	3.7	7.7258	
8	CL9	CL14	9	0.0220	.920	.	.	29.5	7.2	7.6485	
7	CL10	CL19	11	0.0193	.901	.	.	28.7	5.3	8.3443	
6	CL7	Reunion	12	0.0119	.889	.	.	31.9	2.3	8.767	
5	CL6	CL8	21	0.1015	.787	.891	-3.9	19.4	18.0	10.803	
4	Algeria	Tunisia	2	0.0124	.775	.858	-2.8	25.2	.	12.124	
3	CL4	CL5	23	0.0532	.722	.804	-1.6	29.8	5.1	13.132	
2	CL3	Guatemal	24	0.0826	.639	.672	-.56	42.5	6.6	22.577	
1	CL2	CL11	26	0.6390	.000	.000	0.00	.	42.5	45.247	

Figura 6: Tabla de resultados para el método del centroide

muestra la sintaxis para el método del enlace medio pero también se ha obtenido el resultado para el método del centroide y de Ward. Las Figuras 5-7 muestran los resultados para los tres métodos.

Se puede comprobar que el método del enlace medio sugiere realizar 2 o 5 grupos y los métodos del centroide y de Ward, 2 o 6. Dado que no existen grandes diferencias entre los dendrogramas de los dos primeros métodos, sólo se ha representado para los métodos del enlace medio y de Ward (ver Figura 8). Se observa que el método de Ward sugiere la creación de dos únicos clusters (que además coinciden con los dos clusters creados por los otros dos métodos). No obstante, las agrupaciones en 5 o 6 no coinciden para dichos grupos. Por tanto, realizaremos el análisis cluster no jerárquico para dos, cinco y seis grupos y realiaremos el test F de Beale para determinar cuál es mejor. Como ya se ha comentado previamente, no es necesario estandarizar las variables pero, de ser así, sería necesario recurrir al *proc stdize* para obtener las variables estandarizadas y después realizar el análisis sobre éstas.

```
PROC FASTCLUS DATA=esperanza MAXCLUSTERS=2 MEAN=MEDIAS2
DRIFT OUT=cluster2 maxiter=30;
VAR m0 m25 m50 m75 w0 w25 w50 w75;
```

Cluster History											
Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Approximate Expected R-Square	Cubic Clustering Criterion	Pseudo F Statistic	Pseudo t-Squared	Between Cluster Sum of Squares	Tie
10	CL14	Trinidad	5	0.0058	.958	.	.	40.3	2.3	34.2	
9	CL12	Reunion	6	0.0084	.949	.	.	39.8	3.5	49.733	
8	Algeria	CL11	5	0.0121	.937	.	.	38.4	5.4	71.6	
7	Tunisia	CL19	3	0.0150	.922	.	.	37.6	9.3	88.5	
6	CL9	CL10	11	0.0225	.900	.	.	35.9	6.4	133.28	
5	CL8	CL15	9	0.0320	.868	.891	-1.2	34.4	8.8	189.26	
4	CL6	CL7	14	0.0388	.829	.858	-1.2	35.5	6.6	229.75	
3	CL4	Guatemal	15	0.0549	.774	.804	-.66	39.4	6.5	324.96	
2	CL5	CL3	24	0.1350	.639	.672	-.56	42.5	13.4	798.4	
1	CL2	CL13	26	0.6390	.000	.000	0.00	.	42.5	3779.7	

Figura 7: Tabla de resultados para el método de Ward

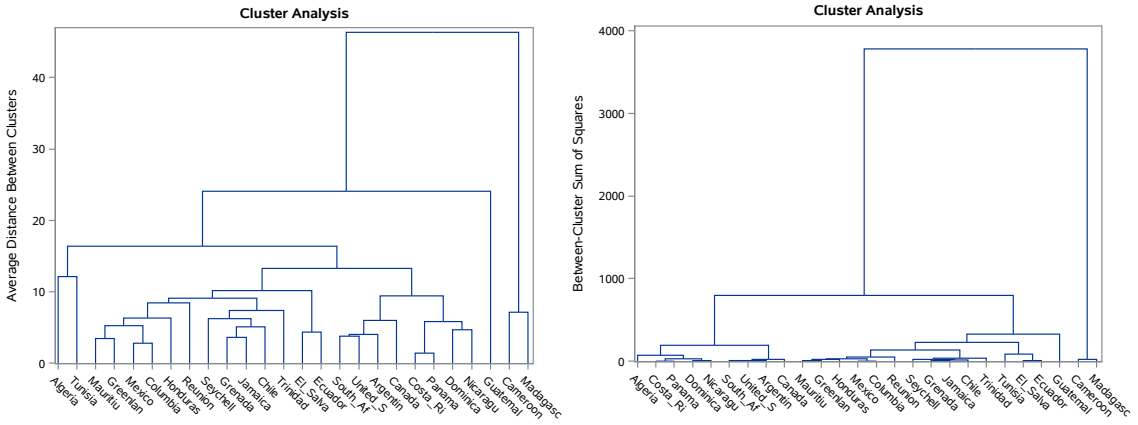

Figura 8: Dendrogramas para los métodos del enlace medio (izquierda) y de Ward (derecha)

```
RUN;
```

El procedimiento para análisis cluster no jerárquico es el *fastclus*. Debemos indicarle el número de clusters *MAXCLUSTERS* y el número máximo de iteraciones *maxiter*. Las sentencias *MEAN* y *OUT* permiten generar un conjunto de datos con las medias de los clusters y la asignación de las observaciones a los clusters, respectivamente. La sentencia *DRIFT* indica que debe recalcular el centroide cada vez que se produzca una reasignación. Como ya se ha indicado, SAS selecciona por defecto los centroides iniciales como aquellos separados a una distancia determinada. Este procedimiento también nos permite seleccionar aleatoriamente los centroides iniciales incluyendo *REPLACE=random*, así como seleccionar la semilla de la selección aleatoria a partir de la sentencia *random=* para que podamos comprobar si el resultado final cambia al seleccionar otros centroides iniciales.

Este procedimiento incluye una serie de salidas como son: la selección inicial de centroides, un resumen de la evolución de las iteraciones, si se ha satisfecho el criterio de convergencia o no, un resumen de los clusters creados, la proporción de variabilidad de las variables explicada por los clusters, el CCC, la PseudoF, el R^2 y las medias y desviaciones típicas de los clusters para las variables consideradas en el análisis.

Repetimos la sentencia anterior para 5 y 6 clusters, lo que nos permite realizar el test F de Beale a partir de las siguientes sentencias:

```
proc means data=cluster2 ; var distance; output out=sumacuad2 uss=w2 ;
run;

proc means data=cluster5 ; var distance; output out=sumacuad5 uss=w5 ;
run;

proc means data=cluster6 ; var distance; output out=sumacuad6 uss=w6 ;
run;

data beale;
    merge sumacuad2 sumacuad5 sumacuad6;
k1=(_freq_-2)*(2**(-2/8));
k2=(_freq_-5)*(5**(-2/8));
k3=(_freq_-6)*(6**(-2/8));
fbeale1=(w2-w5)*k2/(w5*(k1-k2));
pvalor=1-probf(fbeale1,(k1-k2),k2);
fbeale2=(w2-w6)*k3/(w6*(k1-k3));
pvalor2=1-probf(fbeale2,(k1-k3),k3);
fbeale3=(w5-w6)*k3/(w6*(k2-k3));
pvalor3=1-probf(fbeale3,(k2-k3),k3);
run;
 proc print data=beale;run;
```

La Figura 9 muestra los tres contrastes realizados: 2 vs. 5; 2 vs. 6; and 5 vs. 6. Como puede verse, este test sugiere clasificar los paises en 5 grupos. Por ello, analizamos a

w2	w5	w6	k1	k2	k3	fbeale1	pvalor	fbeale2	pvalor2	fbeale3	pvalor3
2135.54	732	596.310	20.1815	14.0435	12.7789	4.38699	0.010324	4.45592	.009694728	2.29926	0.15115

Figura 9: Contraste F de Beale

Cluster Means								
Cluster	m0	m25	m50	m75	w0	w25	w50	w75
1	49.00000000	40.00000000	22.00000000	9.00000000	51.00000000	41.00000000	23.00000000	8.00000000
2	36.00000000	29.50000000	15.00000000	6.00000000	38.00000000	33.00000000	18.50000000	6.50000000
3	59.50000000	48.50000000	27.00000000	12.00000000	65.00000000	54.00000000	33.50000000	17.00000000
4	65.22222222	46.66666667	25.00000000	9.11111111	70.00000000	50.66666667	28.00000000	10.77777778
5	59.25000000	43.25000000	22.66666667	7.75000000	63.66666667	46.91666667	25.50000000	9.50000000

Cluster Standard Deviations								
Cluster	m0	m25	m50	m75	w0	w25	w50	w75
1	.	.	.	.	.	.	.	.
2	2.828427125	0.707106781	2.828427125	1.414213562	0.000000000	1.414213562	2.121320344	0.707106781
3	4.949747468	3.535533906	4.242640687	1.414213562	2.828427125	0.000000000	0.707106781	2.828427125
4	1.922093766	2.061552813	2.121320344	2.204792759	3.162277660	1.118033989	0.866025404	1.641476300
5	2.490892502	2.005673770	2.534608929	1.544785952	2.570225789	1.831955405	1.623688282	1.623688282

Figura 10: Medias y desviaciones típicas para los 5 grupos y las 8 variables incluidas en el análisis

continuación las medias de las variables para intentar caracterizar dichos grupos. Sería interesante contar con otras variables que no se hayan incluido en el análisis para caracterizar dichos grupos y comprender mejor la misma.

La Figura 10 muestra las medias y las desviaciones típicas para los 5 grupos y las 8 variables incluidas en el análisis. El grupo 2 lo componen aquellos paises con menor esperanza de vida, mientras que el grupo 3 lo componen los paises con mayor. El grupo 4 se caracteriza por una mayor esperanza de vida al nacer y una relativamente alta esperanza de vida para las distintas edades. Por último, los grupos 1 y 5 son grupos con esperanzas de vida intermedias pero mayores en 5 que en 1.

En cuanto a la varibilidad de los grupos, se observa que el primer grupo no tiene valores asociados. Esto se debe al hecho de que sólo haya una observación en el grupo. En general, las varibilidades son similares, aunque destacan algunos valores pequeños que nos indican que los valores que toman dichas variables en esos grupos son muy similares. En ocasiones este hecho puede sernos de ayuda a la hora de caracterizar los datos.

Por último, veamos qué paises componen cada uno de los grupos. Para ellos recurrimos a la siguiente sintaxis:

```
proc sort data=cluster5 out=cluster5s;
by cluster;

proc Freq data=cluster5s;
  by cluster; tables pais;run;
```

La salida de la sentencia anterior nos indica que el grupo 1 lo compone Guatemala; el

Cluster History											
Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Approximate Expected R-Square	Cubic Clustering Criterion	Pseudo F Statistic	Pseudo t-Squared	Centroid Distance	Tie
10	CL47	OB144	3	0.0007	.951	.942	2.22	299	6.0	8.7321	
9	CL10	CL11	12	0.0025	.948	.936	2.71	322	5.8	8.7362	
8	CL20	CL33	49	0.0074	.941	.930	2.26	322	27.6	9.1445	
7	CL14	CL17	59	0.0198	.921	.921	-.10	278	47.1	9.8882	
6	CL7	OB113	60	0.0020	.919	.911	1.35	327	2.6	11.726	
5	CL12	CL9	36	0.0173	.902	.895	0.88	332	39.3	12.149	
4	CL8	OB44	50	0.0023	.899	.872	2.42	435	5.6	12.735	
3	CL6	CL29	64	0.0159	.883	.827	4.31	557	21.3	16.986	
2	CL3	CL5	100	0.1108	.773	.697	3.83	503	115	18.102	
1	CL4	CL2	150	0.7726	.000	.000	0.00	.	503	39.74	

Figura 11: Tabla de resultados para el método del centroide

Cluster History											
Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Approximate Expected R-Square	Cubic Clustering Criterion	Pseudo F Statistic	Pseudo t-Squared	Between Cluster Sum of Squares	Tie
10	CL26	CL20	22	0.0027	.959	.942	4.81	368	12.9	183.57	
9	CL27	CL17	31	0.0031	.956	.936	5.02	387	17.8	208.5	
8	CL35	CL15	23	0.0031	.953	.930	5.44	414	13.8	210.87	
7	CL10	CL47	26	0.0058	.947	.921	5.43	430	19.1	395.91	
6	CL8	CL13	38	0.0060	.941	.911	5.81	463	16.3	411.68	
5	CL9	CL19	50	0.0105	.931	.895	5.82	488	43.2	717.61	
4	CL12	CL11	36	0.0172	.914	.872	3.99	515	41.0	1175	
3	CL6	CL7	64	0.0301	.884	.827	4.33	558	57.2	2047.6	
2	CL3	CL4	100	0.1110	.773	.697	3.83	503	116	7565	
1	CL5	CL2	150	0.7726	.000	.000	0.00	.	503	52642	

Figura 12: Tabla de resultados para el método de Ward

2, Camerún y Madagascar; el 3, Algeria y Tunez; el 5 lo componen paises del centro y el sur de américa y el 4, los restantes. Si comparamos este resultado con los dendrogramas de la Figura 8, podemos ver que coincide con los del enlace medio.

9. Ejemplo Iris

En este caso vamos a retomar el conjunto de datos Iris del Análisis Discriminante. Contiene cuatro variables acerca del ancho y el largo del pétalo y el sépalo de tres clases de planta (dado que están todas en mm, no será necesario estandarizar) así como la clase de la misma. Queremos clasificar las hojas a partir de esas cuatro medidas para después comparar dichos grupos con las clases originales. Utilizando una sintaxis similar a la de los ejercicios anteriores, obtenemos las Figuras 11 y 12 referidas al análisis cluster jerárquico, donde puede observarse que el número de clusters a realizar ha de ser dos o tres.

A continuación realizamos análisis cluster no jerárquico para dos y tres clusters y realizamos el contraste F de Beale, obteniendo el resultado que se muestra en la Figura 13. El p-valor es menor que 0,05 por lo que la reducción de la variabilidad interna es significativa y debemos hacer 3 grupos. Por último, comprobamos el parecido entre los clusters obtenidos y los grupos originales.

Obs	_TYPE_	_FREQ_	w2	w3	k1	k2	fbeale	pvalor
1	0	150	15249.28	7885.14	104.652	84.8705	4.00695	.000003768

Figura 13: Contraste F de Beale

Table of CLUSTER by Species				
	Species(Iris Species)			
CLUSTER(Cluster)	Setosa	Versicolor	Virginica	Total
1	0 0.00 0.00 0.00	2 1.33 5.26 4.00	36 24.00 94.74 72.00	38 25.33
2	50 33.33 100.00 100.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	50 33.33
3	0 0.00 0.00 0.00	48 32.00 77.42 96.00	14 9.33 22.58 28.00	62 41.33
Total	50 33.33	50 33.33	50 33.33	150 100.00

Figura 14: Comparacion clusters vs. tipo de planta

```
proc freq data=cluster3;
   tables cluster*species;
run;
```

Comprobamos que el cluster 2 equivale totalmente a la clase “setosa”, el cluster 3 se corresponde principalmente con “versicolor”, aunque 14 observaciones han sido asignadas a “virginica”, y el cluster 1 representa a “virginica”, excepto por 2 observaciones que pertenecen a “versicolor”.