Introducción al Análisis Cluster con SAS

Javier Portela
Facultad de Estadística
UCM

El Análisis de conglomerados

El análisis de conglomerados para casos consiste en agrupar a los elementos de la muestra en "Clusters" o conglomerados a partir de los valores que toman en diversas variables, de manera que éstos grupos sean lo más homogéneos posible en su interior (mínima varianza intra-grupos) y lo más alejados posible entre ellos (máxima varianza entre-grupos). Se exige que las variables sean continuas o todo lo más ordinales para el buen funcionamiento del análisis, aunque eventualmente puede realizarse el análisis con variables categóricas utilizando cierto tipo de distancias.

El análisis de conglomerados se utiliza para tratar problemas de **clasificación**:

- Cómo agrupar especies naturales (en Biología o incluso Geología o Química): éste es el problema de la Taxonomía, que originó los primeros trabajos sobre Clasificación jerárquica.
- Clasificar consumidores-tipo según sus características y preferencias, para estudiarlos mejor por separado. (Marketing).
- Clasificar seres vivos con los mismos síntomas y características patológicas. (Medicina, Genética).
- Encontrar grupos de los mismos comportamientos sociales o psicológicos mediante encuestas (Sociología, Psicología).

Elementos del problema

Variables y casos

Para enfocar un análisis de conglomerados con PROC CLUSTER han de tenerse en cuenta algunos detalles importantes:

- El **número de casos** debe ser limitado, debido a que la mayor parte de las veces el tiempo de CPU de los cálculos realizados por el PROC CLUSTER varían en proporción al cuadrado o cubo del número n de observaciones. Para grandes matrices de datos se debe utilizar el PROC FASTCLUS, cuyo tiempo de proceso es proporcional a n. En todo caso, en el *ejemplo 3* se trata con 500 casos.
- Los datos atípicos (outliers), aunque no sean errores de medición (solamente casos extremos), provocan distorsiones en la mayoría de los métodos. Aunque el PROC CLUSTER facilita su eliminación mediante la opción TRIM=*p*, es posible hacer un análisis previo de los outliers según las diferentes variables utilizadas.
- En general se deben utilizar **variables continuas**. En caso de existir también variables discretas se debe otorgar especial atención a que su variabilidad tenga la misma influencia en el análisis que la de otras variables. Para ello se suele **estandarizar** las variables con la opción STD del PROC CLUSTER. También se suele utilizar esta opción con variables continuas, pero en general si todas las variables tienen **la misma escala de medición** ésta se debe respetar para no adulterar los resultados.
- **Las variables** utilizadas **no** deben estar altamente **correlacionadas**: Si dos lo estuvieran, una de ellas estaría aportando información redundante ya dada por la otra. Los casos serían agrupados casi siempre en orden a los valores tomados doblemente en esas variables, despreciando a menudo la influencia de otras. Para evitar ello, se debe depurar la matriz de datos observando el resultado de PROC CORR y eliminar del análisis algunas de las variables significativamente correlacionadas con otras.
- A continuación del proceso de creación de conglomerados es necesario analizar por qué se han agrupado de determinada manera los casos. Si se ha utilizado un gran número de variables esta tarea es complicada y puede llevar a errores; no conviene, pues, utilizar más de 5 ó 6 variables.

El análisis jerárquico

De los diferentes métodos de este tipo de análisis sólo presentaremos los jerárquicos ascendentes: Los individuos se van uniendo uno a uno, creando grupos, hasta formar el grupo final, todos en uno. En el primer paso del proceso habrá **n** grupos (pues hay **n** casos), en el segundo n-1, hasta llegar al último paso, en el que se crea (mediante unión de los dos grupos que quedan) un único grupo. Esto se suele representar como un árbol donde cada raíz es un caso que se une a otro, formando así el siguiente nivel de unión. Este método de análisis de conglomerados es el más utilizado, y el único presentado en el PROC CLUSTER del SAS.

Asimismo, de los métodos presentados por el SAS se verán sólo aquellos que no toman en cuenta distribuciones de probabilidad sobre las variables, aunque sean estimaciones no paramétricas.

Las distancias

Distancia entre casos

Para determinar cómo unir los elementos de la muestra es necesario definir previamente una **distancia entre casos**. Así se unirían cada vez los dos casos o conglomerados más próximos. Por defecto, el PROC CLUSTER toma la distancia Euclídea, considerando cada caso de los n como un vector de p componentes, siendo p el número de variables utilizadas.

Se puede sin embargo aportar al PROC CLUSTER un archivo SAS que consista en una matriz de distancias declarando en la sentencia data (type=distance); el *ejemplo 4* ilustra esta posibilidad.

Distancia entre conglomerados

Cuando tras varios pasos del algoritmo todos los casos pertenecen a algún conglomerado se hace necesario establecer una **distancia entre conglomerados** para poder calcular cuáles son los más próximos y unirlos en el siguiente paso del proceso. Realmente esta distancia determina el método de creación de grupos, pues en clasificación jerárquica el proceso es similar para todos los métodos. Por ello la distancia entre conglomerados se define en el apartado METHOD=método del PROC CLUSTER. Las más importantes son:

AVERAGE -. El "Average linkage" o Promedio entre grupos, es la distancia entre grupos más utilizada: Corresponde a tomar el promedio de distancias entre **todos** los pares (A,B) siendo A un caso del primer grupo y B un caso del segundo. Hay que destacar que toma toda la información posible de las distancias entre casos, aunque consume mucho tiempo de cálculo.

CENTROID -. Calcula un centro de gravedad o **centroide** para cada grupo (media del grupo para todas las variables). A continuación define la distancia entre grupos como la distancia entre los centroides de los grupos.

SINGLE -. El "Single linkage" o método de las distancias mínimas considera la distancia entre dos grupos como la **mínima** distancia de entre todos los A y B tales que A es un caso del primer grupo y B un caso del segundo grupo.

COMPLETE -. El "Complete linkage" o método de las **distancias máximas** toma como distancia entre dos grupos la distancia entre los dos casos A y B más alejados.

WARD -. El método de Ward o método de la **mínima varianza intra-cluster** une aquellos grupos que dan lugar a un grupo-unión de mayor homogeneidad interna (mínima varianza) y máxima variabilidad con los demás. Aunque optimiza las exigencias del análisis de conglomerados tiende a crear grupos del mismo tamaño y es muy sensible a datos atípicos.

Es aconsejable **utilizar varios métodos** y analizar las diferencias que se obtienen, para no llegar a falsas conclusiones.

Los empates

Si algunas de las variables son **discretas**, puede darse en algún paso del proceso que la distancia entre dos grupos sea igual a la distancia entre otros dos. ¿Qué par de grupos se debe unir ?. El PROC CLUSTER une aquel par que contiene el caso con menor número de identificación (orden en la matriz de datos).

Si el empate se da en los primeros pasos del proceso no tiene mayor relevancia, pero si se da al final puede llevar a resultados muy diferentes el tomar un par u otro: Es conveniente en estos casos recurrir a **reordenar aleatoriamente los datos** mediante un programa SAS y ejecutar el PROC CLUSTER con diferentes reordenaciones. Como alternativa, en algunos casos se puede escoger el método de unión de conglomerados que da el menor número de empates en el transcurso del algoritmo.

¿ Cuántos grupos se deben tomar ?

Si tenemos predeterminado el número k de grupos que nos interesa, en el dendograma (dibujo en forma de árbol que representa el orden de aglomeración) se puede trazar manualmente una línea vertical cortando en el lugar en que hay k grupos formados.

Pero en general no sabemos cuántos grupos es razonable obtener, o incluso si tenemos ideas preconcebidas los mismos datos pueden desmentirlas. En primera instancia se suele recurrir al dendograma para observar en qué momentos del proceso hay **grandes "huecos"** entre varios grupos (siendo "varios" un número deseable de grupos) y éstos contienen ya un gran número de casos.

El PROC CLUSTER también aporta el estadístico "pseudo" F (pseudo pues no sigue una F de Snedecor más que con fuertes exigencias) y el "pseudo" t.

El estadístico pseudo-F representa la relación en cociente entre la variabilidad total de todos los casos respecto a su media (suma de cuadrados) respecto a la suma de variabilidades dentro de cada grupo. Es conveniente que la primera sea grande respecto a la segunda, por lo que como regla general, se tomará el **número de grupos k tal que corresponda al máximo valor de F**. Cuando las variables son discretas y con pocas categorías, F tiende sensiblemente a aumentar a medida que aumenta el número de grupos, por lo que se pondrá más atención a los **saltos en los valores de F** o a los **máximos locales**.

El estadístico pseudo-t representa el **descenso de la variabilidad intra-grupos** al unir dos grupos en cada paso del proceso. Por lo tanto el momento del proceso en que el descenso (deseable) de esta variabilidad sea más brusco nos indicará el punto en que el número k de grupos es el ideal.

Para observar la evolución de estos estadísticos a medida que aumenta el número de grupos es conveniente:

a) Grabar en un archivo SAS los estadísticos F y t y la variable N número de grupos (con la opción OUTTREE=)

b) Hacer un gráfico PLOT que cruce la variable F y N para ver cómo evoluciona F respecto al aumento del número de clusters. Realizar el mismo gráfico para t y N.

Para determinar el número de clusters también es conveniente realizar el mismo análisis con **varios métodos diferentes** (opción METHOD=) y tomar el número de clusters que sea mayoritariamente apropiado según los diferentes métodos.

¿Cómo analizar los grupos creados ?

Tras determinar los grupos creados es necesario estudiar:

a) Por qué son diferentes entre ellos (qué diferencia, en general, los casos de un grupo con los de otro).

b) Por qué son homogéneos (que tienen en común los casos de cada grupo).

Con pocos casos se puede recurrir a un simple **listado de los valores** que toman los individuos de cada grupo en las variables de interés. Si se tiene un gran número de casos se puede además realizar análisis básicos con el procedimiento MEANS, UNIVARIATE, FREQ o TABULATE para observar las diferencias entre los grupos según **estadísticos** (media, d. típica, etc.) de las diferentes variables utilizadas. A veces se debará recurrir a **gráficos** para hacer estas diferencias más evidentes. Posteriormente se suele utilizar el **análisis de la varianza** o incluso el análisis factorial dentro de cada cluster (para ver qué variables son más relevantes en ese grupo de individuos).

En todo caso, el análisis a posteriori de los clusters depende mucho de los intereses iniciales y del tipo de datos, con lo que es difícil dar una pauta general para su desarrollo.

PROC CLUSTER

proc cluster data=*archivo opciones*; **VAR** *variables*; **BY** *variables;*

Opciones del PROC CLUSTER

Lectura y escritura de archivos SAS

DATA=*archivo*

Nombra el archivo del que se van a leer los datos. Éste puede ser un archivo de datos SAS (variables*casos) o bien un archivo de distancias SAS (creado previamente con la opción *data (type=distance)*.

OUTTREE=*archivo*

Nombra un *archivo* de salida, destinado a ser utilizado con el PROC TREE, que lo utiliza para dibujar el diagrama de árbol (dendograma).

Además, el *archivo* de OUTTREE contiene información valiosa para determinar el número de clusters ideal: Se trata de las variables _NCL_ (número de clusters en proceso) , _PSF_ (estadístico F), _PST_ (estadístico t) y ocasionalmente _RMSSTD_ (d. típic del cluster en proceso). Estas variables nos permiten observar cómo varían los estadísticos según el número de clusters creados y decidir en consecuencia. (Sólo se crean con los métodos AVERAGE, CENTROID o WARD).

Este mismo *archivo* contiene información exhaustiva del proceso de formación de clusters. El PROC TREE, a su vez, lo utiliza como base para la creación de un nuevo archivo de la siguiente manera:

> PROC TREE DATA=*archivo del outtree* N=*número de clusters* OUT=*archivo*; COPY *variables*; RUN;

N es el número de grupos que el usuario considera ideal para subdividir la muestra, tras un análisis previo que contiene, entre otros, una ejecución anterior del PROC TREE para ver el diagrama de árbol. **Las variables del COPY** son las variables originales de la matriz de datos. **El archivo de salida** contendrá, para cada caso, el cluster al que pertenece, su identificación (si existía la variable ID en el archivo original) y los valores de las variables declaradas en el COPY.

Este segundo archivo creado será la fuente de información para determinar las características de los individuos de cada grupo, y, *por consiguiente*, uno de los primeros objetivos del análisis.

Opciones de formación de clusters

METHOD=*método*

Donde *método* puede ser, entre otros:

AVERAGE	Método del promedio entre grupos.
CENTROID	Método del centroide.
SINGLE	Método de la mínima distancia
COMPLETE	Método de la distancia máxima
WARD	Método de la mínima varianza

Opciones de tratamiento previo de casos

STD

Estandarización de variables. Útil cuando las variables no pertenecen a la misma escala de medición.

TRIM=*p*

Donde *p* es la proporción de casos a omitir. Es una opción de eliminación de outliers, muy influyentes en los análisis con los métodos WARD, COMPLETE e incluso SINGLE. Opción muy recomendada para los dos primeros. En caso de tomarse la opción TRIM=*p* debe también especificarse la opción

K=n

donde n se utiliza para eliminar outliers mediante el criterio de los "n vecinos más cercanos" (se eliminan los casos más alejados de grupos de tamaño n).

Opciones de presentación del OUTPUT

PRINT=n

Especifica el número de pasos del proceso de creación de clusters que se mostrarán. Por defecto se muetra información de todos los pasos.

PSEUDO

Presenta los estadísticos F y t en los métodos CENTROID, AVERAGE y WARD.

RMSSTD

Presenta la desviación típica dentro cada cluster creado.

Ejemplos

Ejemplo1

Características psicológicas

(Variables discretas con 9 valores numéricos)

Se trata de agrupar una muestra de 20 individuos de alumnos de institutos, por sus características psicológicas medidas en 8 variables. Aunque éstas son discretas, tienen un número suficiente de categorías para no tener empates en la formación de clusters.

Matriz de correlaciones

Es necesario hacer un estudio previo de las correlaciones enter las variables, pues si algunas están muy correladas y otras no, la información dada por las primeras tendrá más peso en la formación de clusters. El proc corr aporta el coeficiente de correlación y el valor del p-value del test asociado.

```
proc corr data=uno;var ansi obse depre desor bamot refis satra conce;
run;
```

CORRELATION ANALYSIS

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 20

	ANSI	OBSE	DEPRE	DESOR
ANSI	1.00000 0.0	0.12524 0.5988	0.66140 0.0015	0.18372 0.4381
OBSE	0.12524 0.5988	1.00000 0.0	0.05709 0.8111	-0.41866 0.0662
DEPRE	0.66140 0.0015	0.05709 0.8111	1.00000 0.0	0.35135 0.1288
DESOR	0.18372 0.4381	-0.41866 0.0662	0.35135 0.1288	1.00000 0.0
BAMOT	0.47056 0.0363	-0.31186 0.1807	0.49451 0.0267	0.74423 0.0002
REFIS	0.63164 0.0028	-0.21549 0.3616	0.52654 0.0171	0.57870 0.0075
SATRA	-0.16442 0.4885	0.05136 0.8297	-0.04839 0.8394	0.02869 0.9044
CONCE	0.01860 0.9380	0.46722 0.0378	0.33970 0.1428	0.12902 0.5877

	BAMOT	REFIS	SATRA	CONCE
BAMOT	1.00000 0.0	0.81404 0.0001	-0.14799 0.5335	0.14459 0.5431
REFIS	0.81404 0.0001	1.00000 0.0	-0.04529 0.8496	0.03594 0.8804
SATRA	-0.14799 0.5335	-0.04529 0.8496	1.00000 0.0	0.46518 0.0388
CONCE	0.14459 0.5431	0.03594 0.8804	0.46518 0.0388	1.00000 0.0

Escogemos para el análisis las variables ANSI, OBSE, DESOR, SATRA y CONCE.

Análisis de conglomerados

A continuación realizamos el análisis de conglomerados con varios métodos para decidir con más criterio cuál será el número ideal de grupos que se debe crear.

```
proc cluster data=uno method=average std pseudo outtree=dos;
var ansi obse desor satra conce;
run;
```

La opción STD estandariza las variables; en nuestro caso no todas las variables siguen la misma escala. La opción PSEUDO aporta los estadísticos F y t2 en la salida. El archivo destinado a realizar el árbol con el Procedimiento TREE es el archivo *dos* .

Average Linkage Cluster Analysis

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	1.71004	0.389074	0.342008	0.34201
2	1.32096	0.155493	0.264193	0.60620
3	1.16547	0.585590	0.233094	0.83929
4	0.57988	0.356235	0.115976	0.95527
5	0.22365	.	0.044729	1.00000

The data have been standardized to mean 0 and variance 1
Root-Mean-Square Total-Sample Standard Deviation = 1
Root-Mean-Square Distance Between Observations = 3.162278

Number of Clusters	Clusters Joined		Frequency of New Cluster	Pseudo F	Pseudo t**2	Normalized RMS Distance	Tie
19	OB7	OB16	2	24.59	.	0.206965	
18	OB2	OB3	2	21.70	.	0.244170	
17	CL18	OB6	3	15.75	2.03	0.325182	
16	CL19	OB8	3	10.53	5.73	0.441411	
15	OB5	OB18	2	9.00	.	0.505889	
14	OB19	OB20	2	8.41	.	0.513417	
13	OB1	CL16	4	7.59	2.55	0.541371	
12	OB9	OB10	2	7.64	.	0.543158	
11	OB4	OB12	2	7.74	.	0.572822	
10	CL17	CL15	5	6.70	4.97	0.629030	
9	CL13	CL10	9	5.20	4.90	0.695966	
8	CL11	OB13	3	5.42	1.79	0.722595	
7	OB15	OB17	2	5.82	.	0.770683	
6	CL9	OB14	10	6.11	2.11	0.789833	
5	OB11	CL7	3	6.87	1.24	0.837754	
4	CL12	CL14	4	7.47	4.32	0.861887	
3	CL6	CL8	13	5.50	8.27	1.064148	
2	CL3	CL4	17	4.09	5.80	1.091084	
1	CL2	CL5	20	.	4.09	1.152224	

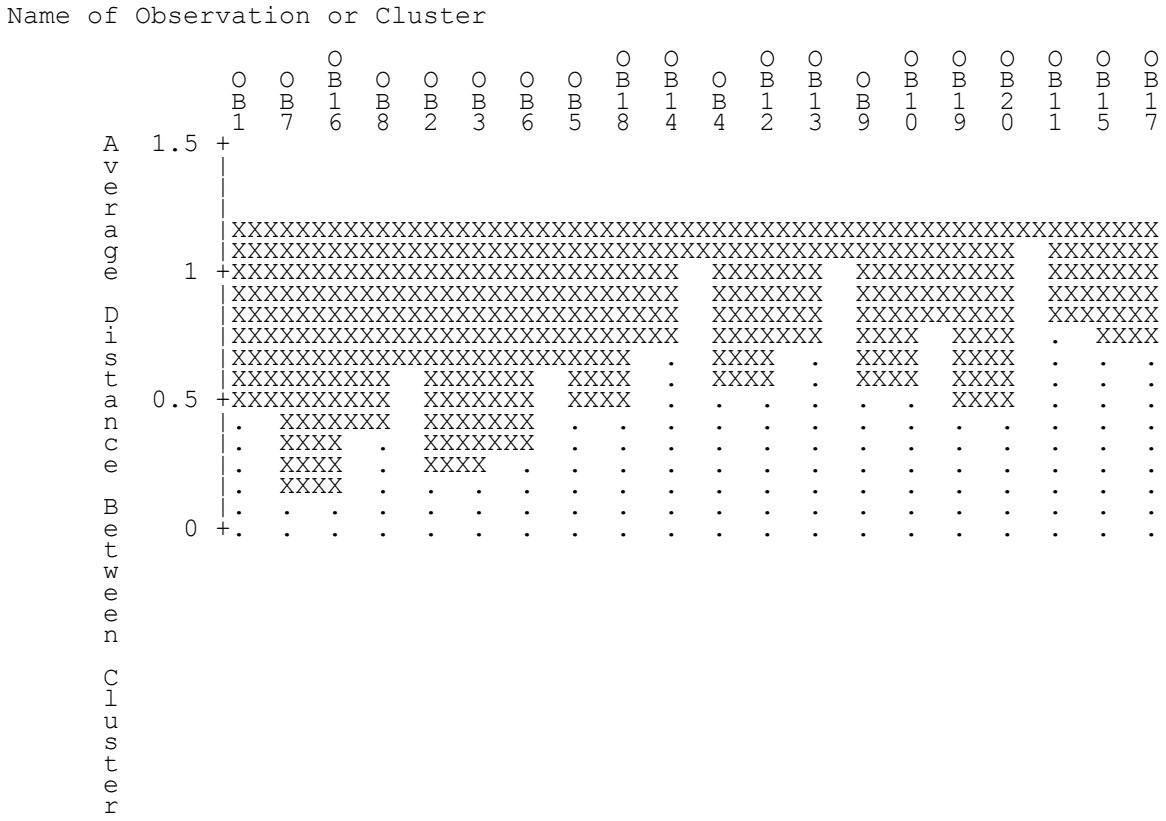
El procedimiento aporta datos sobre la diagonalización de la matriz de correlaciones (ver Análisis de Componentes Principales). Es útil para saber cómo está distribuida la información aportada por las variables. (Si la mayor parte de la proporción de la traza cae sobre un sólo autovalor es porque casi todas las variables del estudio "explican lo mismo" y se agrupan en un factor: esto sería consecuencia de tomar variables correladas). En nuestro caso la información está suficientemente repartida, con tres autovalores de valores parecidos.

Sigue un historial de la evolución del algoritmo, indicando en cada paso los clusters existentes, las observaciones o clusters que se unen, la distancia entre ellas (RMS en este método) y el número de casos que hay en el nuevo cluster. También se aportan los estadísticos F y t2. Nótese que un **máximo local de F** está en 4 clusters y **un mínimo local** de t2 también.

A continuación se presenta el árbol de formación de clusters, con el Procedimiento TREE:

```
proc tree data=dos pos=20;
run;
```

En este procedimiento se lee el archivo creado en OUTTREE y se denota pos=20 para que el árbol no ocupe demasiado en su salida.



La formación de clusters más sólida se da a partir de la altura de la D de "Distance" (4 clusters) al pasar previamente de 9 clusters (La "s")a 6 en la "i" y pasar a 4 de la D a la "g". Esto significa que la **distancia** entre los 4 clusters recién creados (columna izquierda) es grande y **ha dado un salto** desde la distancia entre clusters en la **anterior formación de grupos**.

Aquí se ha pasado desde una distancia media de entre los 6 clusters de 0.75 a la distancia entre los 4 clusters de 1.

Otros métodos

A continuación damos los resultados para los otros métodos cluster:

```
proc cluster data=uno method=centroid std pseudo outtree=dos;
var ansi obse desor satra conce;
run;

proc tree data=dos pos=20;
run;
```

Centroid Hierarchical Cluster Analysis

Eigenvalues of the Correlation Matrix

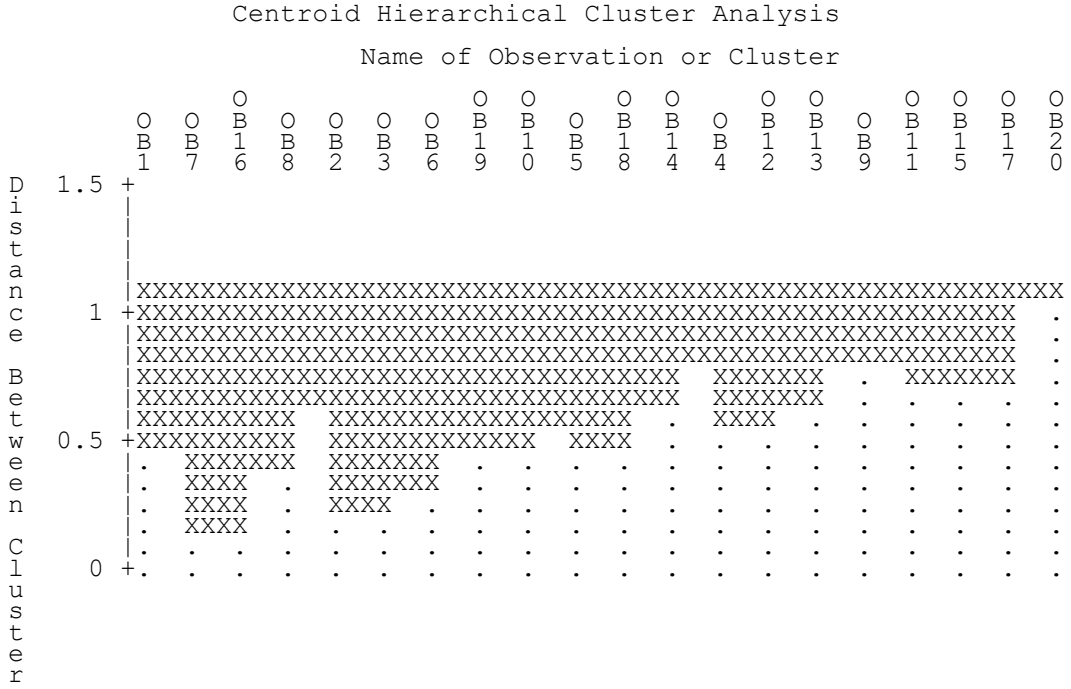
	Eigenvalue	Difference	Proportion	Cumulative
1	1.71004	0.389074	0.342008	0.34201
2	1.32096	0.155493	0.264193	0.60620
3	1.16547	0.585590	0.233094	0.83929
4	0.57988	0.356235	0.115976	0.95527
5	0.22365	.	0.044729	1.00000

The data have been standardized to mean 0 and variance 1
Root-Mean-Square Total-Sample Standard Deviation = 1
Root-Mean-Square Distance Between Observations = 3.162278

Centroid Hierarchical Cluster Analysis

Number of Clusters	Clusters Joined		Frequency of New Cluster	Pseudo F	Pseudo t**2	Normalized Centroid Distance	Tie
19	OB7	OB16	2	24.59	.	0.206965	
18	OB2	OB3	2	21.70	.	0.244170	
17	CL18	OB6	3	15.75	2.03	0.301395	
16	CL19	OB8	3	10.53	5.73	0.429109	
15	OB1	CL16	4	7.75	2.55	0.495000	
14	CL17	OB19	4	6.80	4.10	0.496835	
13	OB5	OB18	2	6.99	.	0.505889	
12	CL14	OB10	5	6.56	2.36	0.520950	
11	OB4	OB12	2	6.78	.	0.572822	
10	CL12	CL13	7	5.38	4.14	0.600076	
9	CL15	CL10	11	3.52	6.41	0.639970	
8	CL11	OB13	3	3.78	1.79	0.663410	
7	CL9	OB14	12	3.85	1.82	0.705473	
6	OB15	OB17	2	4.36	.	0.770683	
5	OB11	CL6	3	4.97	1.24	0.743870	
4	CL8	OB9	4	5.73	2.16	0.810709	
3	CL7	CL4	16	3.79	7.14	0.815484	
2	CL3	CL5	19	2.44	4.64	0.842438	
1	CL2	OB20	20	.	2.44	1.092594	

Centroid Hierarchical Cluster Analysis



```
proc cluster data=uno method=complete std pseudo outtree=dos;
var ansi obse desor satra conce;
run;
proc tree data=dos pos=20;
run;
```

Complete Linkage Cluster Analysis

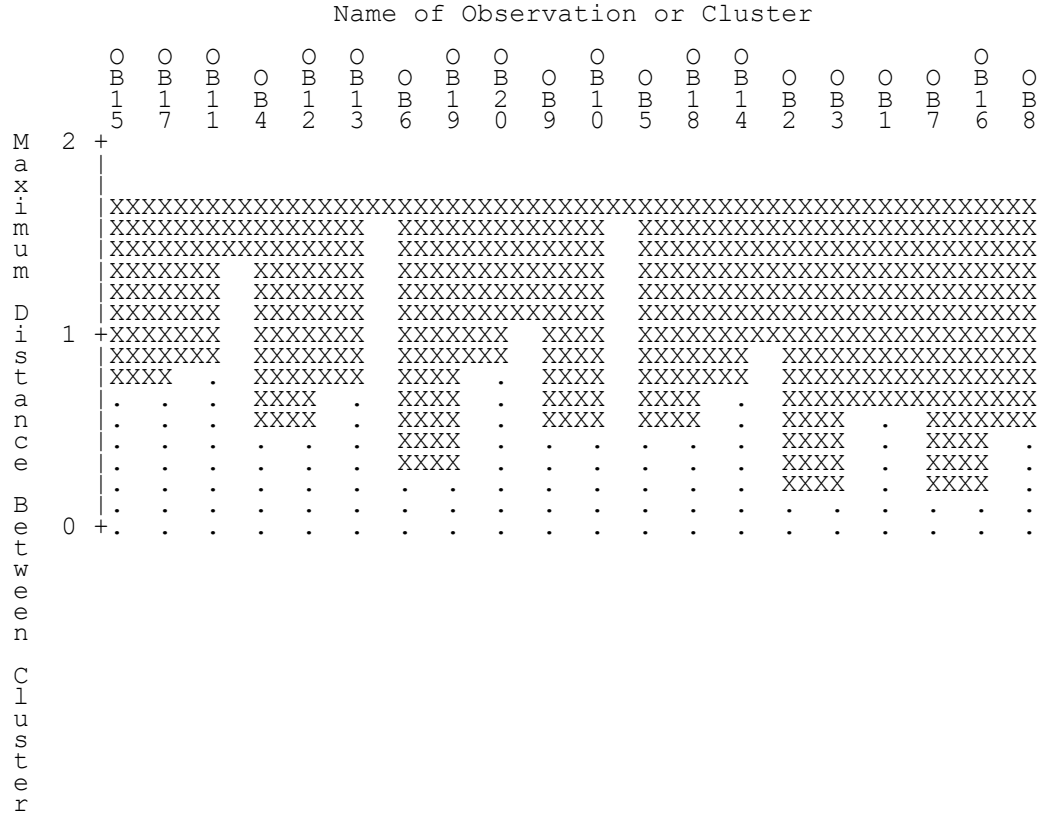
Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	1.71004	0.389074	0.342008	0.34201
2	1.32096	0.155493	0.264193	0.60620
3	1.16547	0.585590	0.233094	0.83929
4	0.57988	0.356235	0.115976	0.95527
5	0.22365	.	0.044729	1.00000

Complete Linkage Cluster Analysis

Number of Clusters	Clusters Joined		Frequency of New Cluster	Pseudo F	Pseudo t**2	Normalized Maximum Distance	Tie
19	OB7	OB16	2	24.59	.	0.217000	
18	OB2	OB3	2	21.70	.	0.256009	
17	OB6	OB19	2	15.64	.	0.367111	
16	CL19	OB8	3	10.50	5.73	0.509757	
15	OB5	OB18	2	8.98	.	0.530418	
14	OB9	OB10	2	8.12	.	0.569493	
13	OB4	OB12	2	7.63	.	0.600595	
12	OB1	CL18	3	6.96	7.51	0.646704	
11	CL12	CL16	6	6.59	2.44	0.705455	
10	CL15	OB14	3	6.29	2.22	0.785122	
9	OB15	OB17	2	6.21	.	0.808050	
8	CL13	OB13	3	6.36	1.79	0.811846	
7	CL17	OB20	3	6.71	4.92	0.879105	
6	OB11	CL9	3	7.10	1.24	0.883458	
5	CL11	CL10	9	7.08	4.00	1.024747	
4	CL7	CL14	5	7.67	3.57	1.085735	
3	CL8	CL6	6	6.17	5.73	1.457566	
2	CL3	CL4	11	5.90	3.86	1.647226	
1	CL5	CL2	20	.	5.90	1.698103	

Complete Linkage Cluster Analysis



El método del centroide no da resultados análogos al del promedio entre grupos, pero el de la máxima distancia sí: se ve en el gráfico cómo es conveniente tomar 4 clusters.

Los estadísticos F y t

A continuación se utilizan los resultados de este último método (la matriz del OUTTREE) para observar la evolución de los estadísticos F y t . Previamente hay que ordenar esos datos por la variable del eje horizontal (el número de clusters, _ncl_). Para el gráfico se ha utilizado el PROC PLOT (aunque también se puede utilizar el proc gplot, posiblemente con la opción symbol i=join).

```
proc sort data=dos;by _ncl_;

proc plot data=dos;plot (_psf_ _pst2_)*_ncl_;
run;
```

```
            Plot of _PSF_*_NCL_.  Legend: A = 1 obs, B = 2 obs, etc.

     P 30 +
     s    |                                                          A
     e    |                                                       A
     u    |
     d 20 +
     o    |                                                    A
     F    |
       10 +                                              A  A
     S    |           A  A  A  A  A     A  A  A  A  A
     t    |     A  A                 A
     a    |
     t  0 +
     i    ---+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--
     s       1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
     t
                                Number of Clusters

          Plot of _PST2_*_NCL_.  Legend: A = 1 obs, B = 2 obs, etc.

     P 10 +
     s    |
     e    |
     u    |                                      A
     d    |
     o    |  A     A                                        A
        5 +                     A
     T    |     A        A
     -    |              A
     S    |                              A  A
     q    |                        A
     u    |                   A
     a  0 +
     r    ---+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--
     e       1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
     d
```

Se observa en los gráficos lo mismo que se había advertido al ver los valores de F y t: Existe un máximo local para F en 4 clusters (gráfico superior) y un mínimo local en 4 clusters y también en 6 en el estadístico t (gráfico inferior). Pero 6 clusters parecen muchos para un conjunto de 20 observaciones, aunque todo depende del objetivo del análisis.

Los 4 clusters creados

Se realiza el proc tree solamente para crear una matriz donde cada caso lleve asociado su número de cluster, además de las variables originales (opción COPY). Es necesario designar el número de clusters elegido (opción n=).

```
proc tree data=dos noprint out=tres n=4;copy ansi obse desor satra conce;
run;
```

La matriz tres será analizada con el PROC TABULATE de manera a obtener la media de las variables de interés en cada cluster:

```
proc tabulate data=tres;
class cluster;
var ansi obse desor satra conce;
table cluster,(ansi obse desor satra conce)*mean;
run;
```

	ANSI	OBSE	DESOR	SATRA
	MEAN	MEAN	MEAN	MEAN
CLUSTER				
1	2.22	5.11	2.00	5.22
2	1.80	2.00	2.40	3.40
3	5.67	2.33	6.33	5.00
4	5.67	5.33	2.33	1.00

	CONCE
	MEAN
CLUSTER	
1	5.67
2	2.20
3	5.00
4	4.00

Se pueden observar las diferencias entre los clusters en términos de esas medias. Es posible aportar en el PROC TABULATE otros estadísticos como MIN, MAX, SUM, etc. En este caso (variables discretas de 9 categorías) puede que tenga más interés la **mediana** que la media; para ello se debería recurrir al PROC UNIVARIATE con la opción BY cluster; el PROC TABULATE no dispone de la mediana.

El grupo 1 "de los estudiosos" tiene una alta media en obsesión, satisfacción por el trabajo y nivel de concentración. El **grupo 2** denota un cierto relajo psicológico, combinado con una baja satisfacción por el trabajo. El **grupo 3** combina curiosamente una gran concentración a la vez que alta desorganización y alta satisfacción con el trabajo, designando aparentemente un grupo de "creativos". El **grupo 4** son individuos de máxima ansiedad y obsesión y mínima satisfacción con el trabajo.

Posibles análisis posteriores

La variable *cluster* ofrece amplias posibilidades estadísticas: interesa analizar con detalle si discrimina grupos de una manera clara. El análisis de la varianza es aconsejable. Se puede optar también a análisis no paramétricos como el de Kruskal Wallis, útil en este caso, donde todas las variables son discretas. Todo tipo de gráficos serían de gran ayuda.

Sería posible también observar como varían los valores medios de **otras variables** no utilizadas según los cluster. Otra posibilidad sería añadir algún caso más con características especiales (dirigidas, incluso artificialmente: valores extremos en las variables de interés) y observar cómo varía el proceso de creación de cluster.

Ejemplo2

Clasificación según pruebas deportivas

(Variables continuas correlacionadas)

Presentación del problema

Disponemos de una matriz de datos tomada de pruebas físicas realizadas a 30 chicos de 15 años en institutos de Madrid (1993):

ABD	Prueba de abdominales
GP11	Golpeo de placas (flexibilidad)
SLO11	Salto de longitud
DIM11	Dinamometría manual
CNA1	Carrera en espacios cortos (10x5)

Nos interesa obtener grupos de individuos de las mismas características deportivas, para desarrollar sus capacidades.

Todas las variables tienen alta correlación

```
proc corr data=uno; var abd gp11 slo11 dim11 cna1;
run;
```

CORRELATION ANALYSIS

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 30

	ABD	GP11	SLO11	DIM11	CNA1
ABD	1.00000 0.0	0.50027 0.0049	0.65838 0.0001	0.66335 0.0001	0.72693 0.0001
GP11	0.50027 0.0049	1.00000 0.0	0.28929 0.1210	0.48694 0.0064	0.78480 0.0001
SLO11	0.65838 0.0001	0.28929 0.1210	1.00000 0.0	0.61256 0.0003	0.48293 0.0069
DIM11	0.66335 0.0001	0.48694 0.0064	0.61256 0.0003	1.00000 0.0	0.55604 0.0014
CNA1	0.72693 0.0001	0.78480 0.0001	0.48293 0.0069	0.55604 0.0014	1.00000 0.0

Se observa que todas las variables están altamente correladas, como era de esperar; pero como todas lo están al mismo nivel, no habrá algunas facultades físicas que tengan preponderancia sobre otras a la hora de crear los clusters. Se puede continuar el proceso, teniendo en cuenta que cabe la posibilidad de eliminar alguna variable.

Conglomerados y outliers

En las opciones de PROC CLUSTER para este ejemplo se añade TRIM=10 (con k=3) para eliminar un 10% de los casos, considerados outliers, evitando así su intervención en el proceso, que en caso del método promedio entre grupos (METHOD=AVERAGE) puede llegar a ser fuerte, pues un outlier puede alterar mucho la distancia media entre su grupo y otro.

En el ejemplo anterior no se añadía esta opción: no tiene sentido hablar de outliers en variables de sólo 9 valores.

Esta opción tiende a "inflar" los valores de los estadísticos F y t.

Por supuesto, se añade la opción STD, pues las variables están tomadas en escalas muy diferentes.

```
proc cluster data=uno method=average pseudo trim=10 k=3 outtree=saluno
std;
var abd dim11 slo11 gp11 cna1;
run;
```

Average Linkage Cluster Analysis

3 observation(s) trimmed with estimated density 0.0005640409 or less.

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	2.56465	1.60499	0.512931	0.51293
2	0.95967	0.26795	0.191933	0.70486
3	0.69172	0.11537	0.138343	0.84321
4	0.57635	0.36874	0.115270	0.95848
5	0.20761	.	0.041522	1.00000

The data have been standardized to mean 0 and variance 1
Root-Mean-Square Total-Sample Standard Deviation = 1
Root-Mean-Square Distance Between Observations = 3.162278

Average Linkage Cluster Analysis

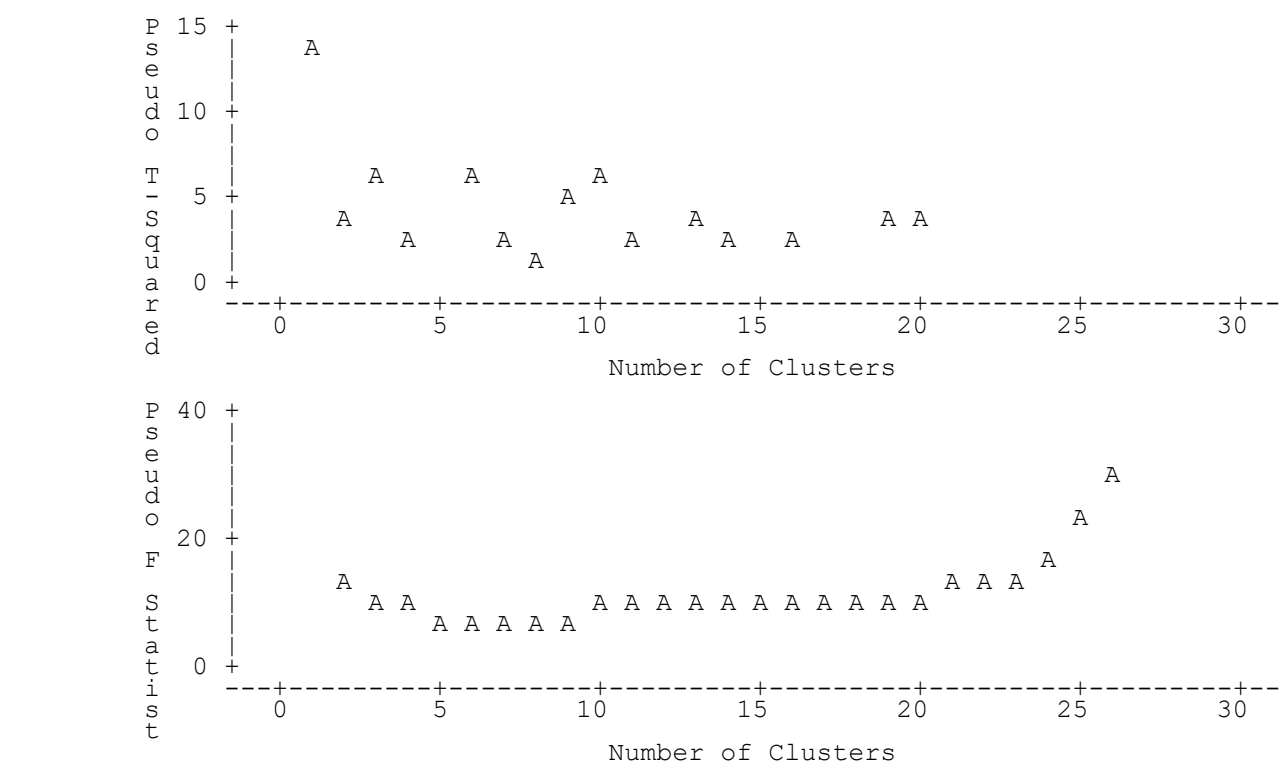
Number of Clusters	Clusters Joined		Frequency of New Cluster	Pseudo F	Pseudo t**2	Normalized RMS Distance	Tie
26	OB15	OB16	2	31.59	.	0.181341	
25	OB11	OB20	2	23.04	.	0.246604	
24	OB2	OB14	2	16.94	.	0.324055	
23	OB9	OB29	2	14.31	.	0.357171	
22	OB18	OB23	2	13.25	.	0.364236	
21	OB6	OB21	2	12.88	.	0.364781	
20	CL24	CL26	4	11.27	3.36	0.387494	
19	OB1	CL25	3	10.42	3.95	0.442219	
18	OB12	OB13	2	10.16	.	0.472617	
17	OB5	OB22	2	10.09	.	0.478351	
16	CL20	OB8	5	9.55	2.74	0.507063	
15	OB17	OB27	2	9.70	.	0.507616	
14	CL18	CL22	4	9.50	2.05	0.521422	
13	CL21	CL23	4	9.27	3.29	0.528909	
12	OB25	OB26	2	9.51	.	0.594380	
11	CL17	OB7	3	9.63	1.95	0.626043	
10	CL19	CL16	8	8.46	6.18	0.630568	
9	CL10	CL13	12	7.20	5.30	0.696969	
8	CL11	OB10	4	7.70	1.73	0.708625	
7	CL15	CL12	4	7.91	3.04	0.785760	
6	CL9	CL14	16	6.60	6.87	0.816007	
5	OB3	OB24	2	7.67	.	0.863201	
4	OB4	CL8	5	9.08	2.49	0.899765	
3	CL6	CL4	21	9.03	5.79	0.918590	
2	CL5	CL7	6	13.28	3.75	1.087370	
1	CL3	CL2	27	.	13.28	1.278456	

En los autovalores de la matriz de correlaciones se observa la influencia de la alta correlación entre variables: El primer autovalor agrupa el 50% de la traza. Habría que llegar al análisis de Componentes Principales para llegar a otras conclusiones más precisas.

Se ve que el estadístico F alcanza un máximo local en 4 clusters aunque también pueden tomarse 3, y el t un mínimo local en 4 y 8. Esto se observará en los siguientes gráficos:

```
proc sort data=saluno;by _ncl_;

proc plot data=saluno;plot (_pst2_ _psf_)*_ncl_ ;
run;
```



Dos agrupaciones distintas: la interpretación decide

El estadístico F no deja claro cuántos clusters tomar; probaremos con n=4 y n=3, observando en el PROC TABULATE sus resultados respectivos.

```
proc tree data=saluno noprint n=4 out=saldos;copy abd gp11 slo11 dim11
cna1;
run;
proc tabulate data=saldos;
class cluster;
var abd dim11 slo11 gp11 cna1;
table cluster,(abd dim11 slo11 gp11 cna1)*mean;
run;

proc tree data=saluno noprint n=3 out=saldos;copy abd gp11 slo11 dim11
cna1;
run;
proc tabulate data=saldos;
class cluster;
var abd dim11 slo11 gp11 cna1;
table cluster,(abd dim11 slo11 gp11 cna1)*mean;
run;
```

	ABD	DIM11	SLO11	GP11
	MEAN	MEAN	MEAN	MEAN
CLUSTER				
1	25.81	35.72	2.01	8.90
2	25.20	42.40	2.03	10.86
3	20.75	34.13	1.50	9.96
4	24.50	31.25	1.42	11.47

	CNA1
	MEAN
CLUSTER	
1	19.20
2	18.66
3	20.47
4	23.58

	ABD	DIM11	SLO11	GP11
	MEAN	MEAN	MEAN	MEAN
CLUSTER				
1	25.67	37.31	2.01	9.37
2	20.75	34.13	1.50	9.96
3	24.50	31.25	1.42	11.47

	CNA1
	MEAN
CLUSTER	
1	19.07
2	20.47
3	23.58

Parece que los grupos 1 2 y 3 son más fáciles de interpretar que los 4 formados en primer lugar (un criterio que se utiliza también en análisis factorial): El **grupo 1** contendría chicos con alta puntuación en casi todas las pruebas, menos en flexibilidad y carrera, presumiblemente chicos grandes y (perdón) toscos, pesados. El **grupo 2** lo formarían individuos de mediana capacidad física y quizás escasa talla, como indica su media en salto de longitud (a pies juntos). El **grupo 3** estaría constituido por chicos ligeros, ágiles (como indica su media en flexibilidad y carrera) pero de escasa fuerza (test de dinamometría).

Variables ilustrativas

Para corroborar estas hipótesis se procede a ampliar la última tabla con las variables TALLA y PESO, de las que también se dispone. Se une primero la matriz obtenida anteriormente con la matriz de datos original, donde están estas dos nuevas variables.

```
proc sort data=saldos;by cna1;
proc sort data=uno;by cna1;

data dos;
merge saldos uno;
by cna1;


proc tabulate data=dos;
class cluster;
var abd dim11 slo11 gp11 cna1 peso1 talla1;
table cluster,(abd dim11 slo11 gp11 cna1 peso1 talla1)*mean;
run;
```

	ABD	DIM11	SLO11	GP11
	MEAN	MEAN	MEAN	MEAN
CLUSTER				
1	25.67	37.31	2.01	9.37
2	20.75	34.13	1.50	9.96
3	24.50	31.25	1.42	11.47

	CNA1	PESO1	TALLA1
	MEAN	MEAN	MEAN
CLUSTER			
1	19.07	59.62	168.24
2	20.47	61.45	164.15
3	23.58	59.95	157.40

Los datos obtenidos confirman en parte las hipótesis, salvo en lo relativo al PESO: se observa que la talla es un factor fundamental en los resultados de estas pruebas deportivas.

Ejemplo 3

Facultades numéricas, facultades verbales y aptitud espacial (un ejemplo con 500 observaciones)

Presentación del problema

Pretendemos establecer una clasificación sobre 500 individuos de nuestra matriz de datos basada en las variables FNUM, FVERB y APTESP. El principal problema será el gran número de observaciones; mantenemos sin embargo a 10% la eliminación de outliers, pero eliminamos gran parte de la salida con las opciones NOPRINT.

Depuración de datos

En algunos casos, como éste, existen valores 0 que corresponden a valores missing; es necesario ponerlos a missing para que no sean tomados como valores reales de las variables.

```
data uno;
array x fnum fverb aptesp;
set pepe.matriz (obs=500);
do i=1 to 3;
   if x{i}=0 then delete;
end;
```

Análisis con la opción NOPRINT

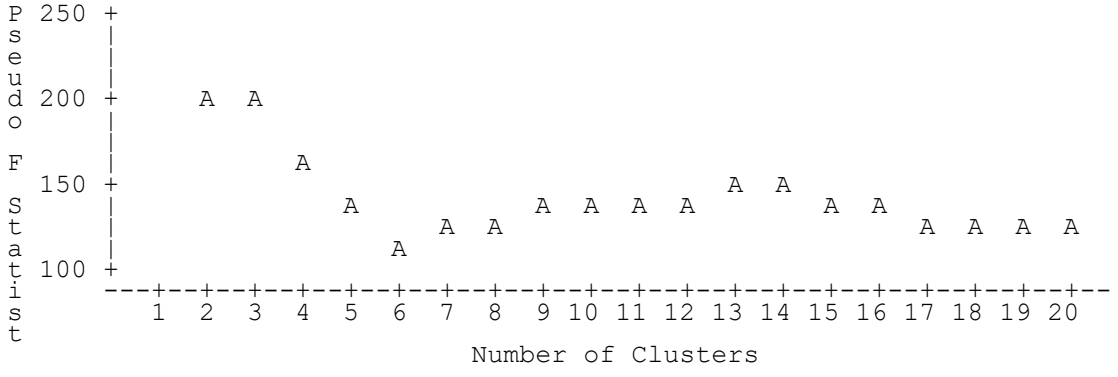
Dado que el proceso de creación de clusters es difícil de seguir (500 cluster), a menos que nos interesara hacer un seguimiento de algún individuo en particular , nos restringimos a extraer el gráfico de los estadísticos F y t, eliminando de esa matriz (para el gráfico) la evolución de más de 20 clusters en adelante.

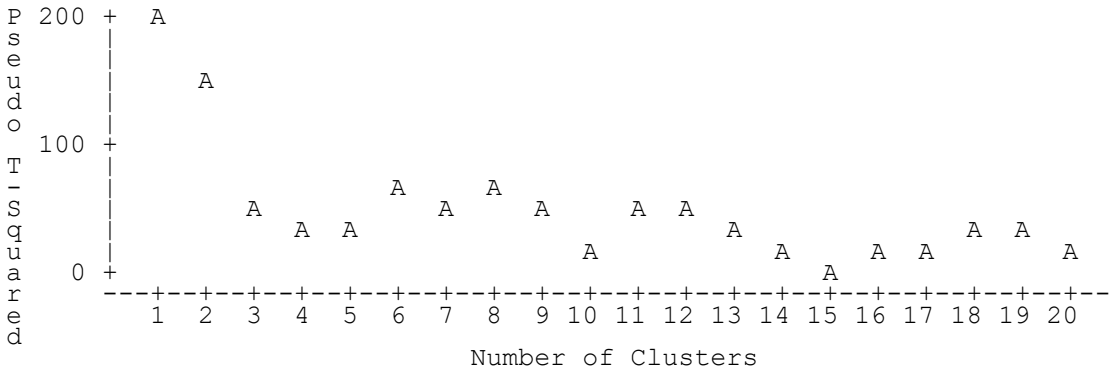
```
proc cluster data=uno method=average trim=10 k=3 noprint outtree=dos;
var fnum fverb aptesp;
run;

proc sort data=dos;by _ncl_;run;

data tres;
set dos;
if _ncl_ gt 20 then delete;
run;

proc plot data=tres;plot (_psf_ _pst2_)*_ncl_;
run;
```





Parece que se pueden tomar 4 o 3 clusters; tomamos esta última opción.

Contando los casos dentro de cada cluster

Existe la posibilidad de observar el número de casos que cae dentro de cada cluster; algunos métodos de formación de cluster tienden a formar grupos por un lado muy numerosos y por otro escasos. Añadir la opción n en el procedimiento tabulate es de mucho interés para análisis posteriores.

```
proc tree data=dos noprint n=3 out=sal;copy fnum fverb aptesp;
run;


proc tabulate data=sal;
class cluster;
var fnum fverb aptesp;
table cluster,(fnum fverb aptesp)*mean n;
run;
```

	FNUM	FVERB	APTESP	
	MEAN	MEAN	MEAN	N
CLUSTER				
1	37.73	33.62	35.16	116.00
2	31.80	29.64	14.74	200.00
3	45.34	37.17	55.23	63.00

El **grupo 3** (63 casos) agrupa claramente chicos con muy alta puntuación en todas las facultades. Estos cocos conviven con los del **grupo 2** (200), el caso opuesto, mientras que el **grupo 1** (116) parece tener inclinación a lo numérico.

Ejemplo 4

Clasificando ciudades por su distancia.

Presentación del problema

Queremos clasificar en grupos de proximidad las ciudades Albacete Alicante Almeria Avila Badajoz Barcelona, Bilbao y Burgos. Para ello introduciremos su matriz de distancias y realizaremos el análisis con los distintos métodos, comparando resultados.

Datos tipo distancia

Es una opción de la sentencia data: de la matriz simétrica de distancias sólo es necesaria la matriz triangular inferior: 366 es, por ejemplo, la distancia de Ávila (4ª fila-columna) a Albacete (1ª fila-columna). Es necesario añadir a los datos el nombre de cada caso (variable ciudad).

```
data dista (type=distance);
   input (Albacete Alicante ALmeria Avila Badajoz Barcelon Bilbao
   Burgos) (5.);
cards;
0
171  0
369  294  0
366  537  663  0
525  696  604  318  0
540  515  809  717  1022 0
646  817  958  401  694  620  0
488  659  800  243  536  583  158  0
;


data nombres;
input ciudad $10. @@;
cards;
Albacete  Alicante  Almeria   Avila     Badajoz    Barcelona  Bilbao    Burgos
;

data todo (type=distance);
merge dista nombres;
run;
```

Varios métodos

Aquí sí son interesantes los gráficos de árbol.

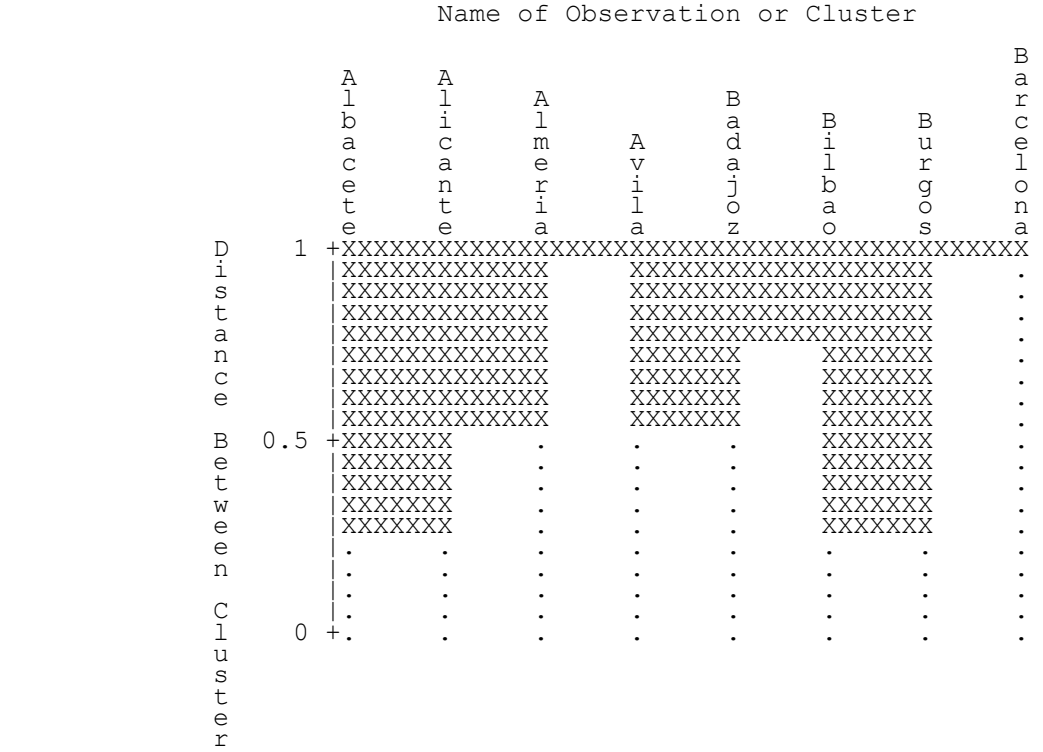
```
proc cluster data=todo method=centroid pseudo;
id ciudad;

proc tree spaces=5 pos=20;
run;
```

Centroid Hierarchical Cluster Analysis

Root-Mean-Square Distance Between Observations = 602.4539

Number of Clusters	Clusters Joined		Frequency of New Cluster	Pseudo F	Pseudo t**2	Normalized Centroid Distance	Tie
7	Bilbao	Burgos	2	16.80	.	0.262261	
6	Albacete	Alicante	2	18.35	.	0.283839	
5	Avila	Badajoz	2	11.52	.	0.527841	
4	CL6	Almeria	3	10.19	4.74	0.535264	
3	CL5	CL7	4	6.26	6.84	0.770858	
2	CL4	CL3	7	2.00	8.13	0.973757	
1	CL2	Barcelona	8	.	2.00	1.000173	

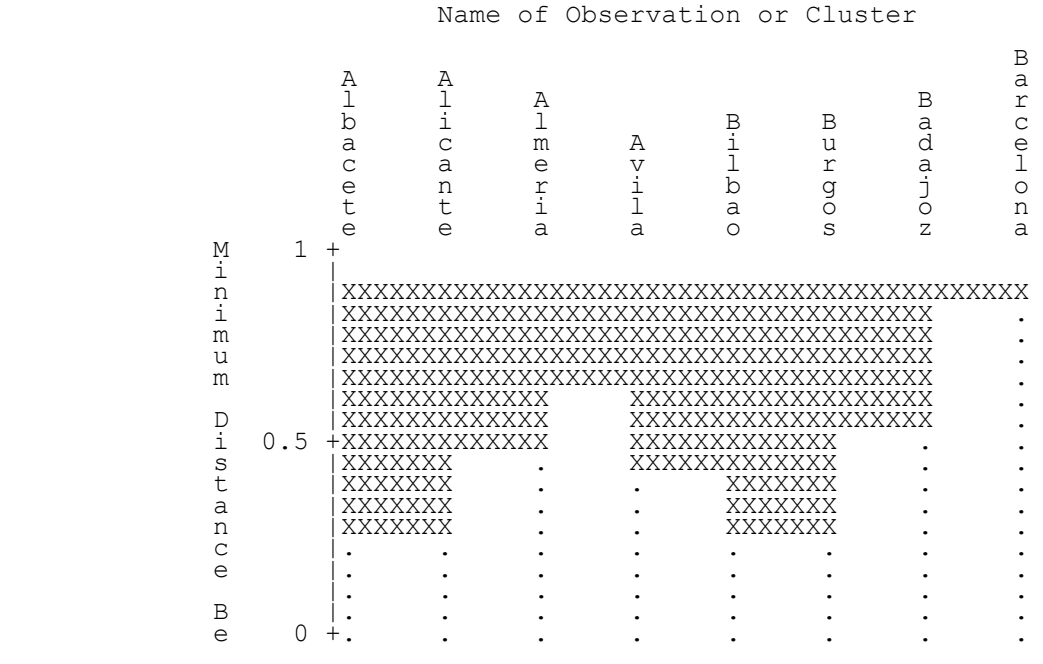



```
proc cluster data=todo method=single pseudo;
id ciudad;
proc tree spaces=5 pos=20;
run;
```

Single Linkage Cluster Analysis

Mean Distance Between Observations = 562.4643

Number of Clusters	Clusters Joined		Frequency of New Cluster	Normalized Minimum Distance	Tie
7	Bilbao	Burgos	2	0.280907	
6	Albacete	Alicante	2	0.304019	
5	Avila	CL7	3	0.432027	
4	CL6	Almeria	3	0.522700	
3	CL5	Badajoz	4	0.565369	
2	CL4	CL3	7	0.650708	
1	CL2	Barcelona	8	0.915614	

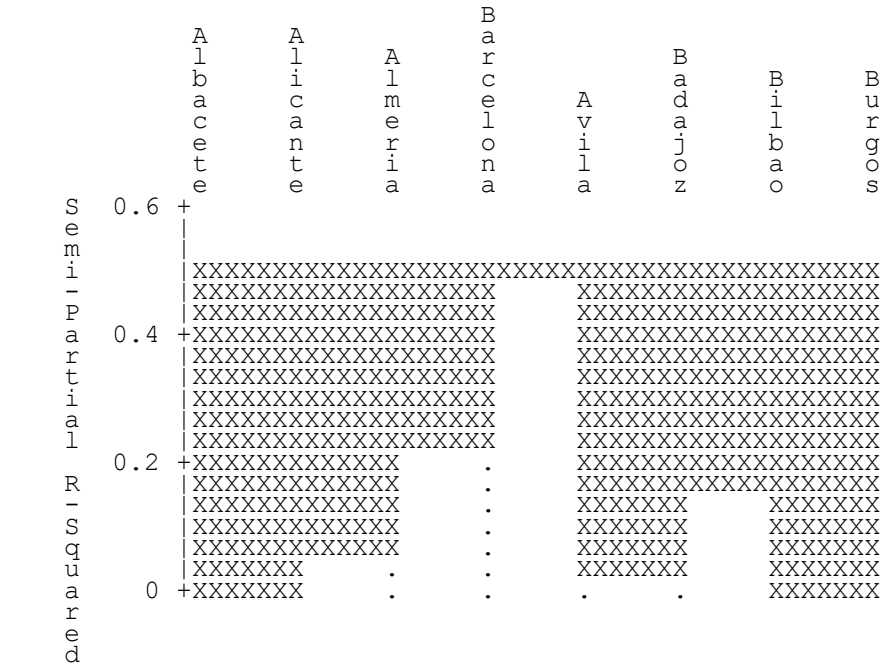


```
proc cluster data=todo method=ward pseudo;
id ciudad;
proc tree spaces=5 pos=20;
run;
```

Ward's Minimum Variance Cluster Analysis

Root-Mean-Square Distance Between Observations = 602.4539

NCL	Clusters Joined		FREQ	SPRSQ	RSQ	Pseudo F	Pseudo t**2	Tie
7	Bilbao	Burgos	2	0.009826	0.99017	16.8	.	
6	Albacete	Alicante	2	0.011509	0.97866	18.3	.	
5	Avila	Badajoz	2	0.039802	0.93886	11.5	.	
4	CL6	Almeria	3	0.054573	0.88429	10.2	4.7	
3	CL5	CL7	4	0.169778	0.71451	6.3	6.8	
2	CL4	Barcelona	4	0.221864	0.49265	5.8	6.7	
1	CL2	CL3	8	0.492648	0.00000	.	5.8	

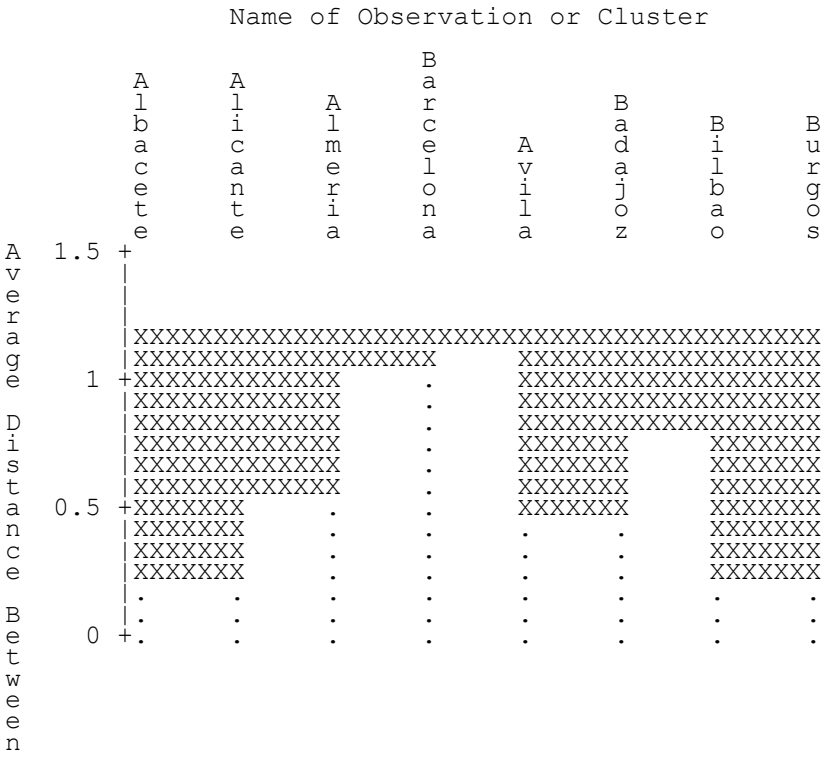



```
proc cluster data=todo method=average pseudo;
id ciudad;
proc tree spaces=5 pos=20;
run;
```

Average Linkage Cluster Analysis

Root-Mean-Square Distance Between Observations = 602.4539

Number of Clusters	Clusters Joined		Frequency of New Cluster	Pseudo F	Pseudo t**2	Normalized RMS Distance	Tie
7	Bilbao	Burgos	2	16.80	.	0.262261	
6	Albacete	Alicante	2	18.35	.	0.283839	
5	Avila	Badajoz	2	11.52	.	0.527841	
4	CL6	Almeria	3	10.19	4.74	0.553759	
3	CL5	CL7	4	6.26	6.84	0.825271	
2	CL4	Barcelona	4	5.83	6.71	1.054733	
1	CL2	CL3	8	.	5.83	1.142833	



¿Cómo agrupar las ciudades?

Casi todos los métodos coinciden en situar las divisiones ESTE (Almería, Albacete, Alicante) y NORTE-OESTE (Ávila, Badajoz, Bilbao, Burgos). Barcelona es difícil de clasificar, y forma un cluster aparte en la mayoría de los métodos. Ward es el único método que incluye Barcelona tempranamente en un cluster, y lo hace en del NOROESTE.

Tomamos como mejor opción el método promedio entre grupos (AVERAGE), con 3 clusters. Lo presentamos con el PROC PRINT, visto el corto número de casos.

```
proc tree noprint n=3 out=todo2;copy ciudad;
run;

proc sort data=todo2;by cluster;
proc print data=todo2;run;
```

OBS	_NAME_	CIUDAD	CLUSTER	CLUSNAME
1	Bilbao	Bilbao	1	CL3
2	Burgos	Burgos	1	CL3
3	Avila	Avila	1	CL3
4	Badajoz	Badajoz	1	CL3
5	Albacete	Albacete	2	CL4
6	Alicante	Alicante	2	CL4
7	Almeria	Almeria	2	CL4
8	Barcelona	Barcelona	3	Barcelona

Los tres grupos formados son los anteriormente mencionados. Es interesante saber que según este método (distancia promedio entre todas las ciudades de cada grupo) Barcelona se situaría más cerca del grupo del ESTE que del NOROESTE.

Ejemplo 5

Clasificación de países según su PIB y esperanza de vida. (Un ejemplo con dos únicas variables).

Presentación del problema

Disponemos de datos relativos al PIB y esperanza de vida de 19 países repartidos entre todo el mundo (datos de la ONU, 1995). Desearíamos establecer una clasificación objetiva de estos países según estas dos variables. Como se verá en el desarrollo del ejemplo, el hecho de que sólo se traten dos variables permite nuevos análisis gráficos en el plano (y más procedimientos estadísticos que no se mencionan).

```
data paises;
input pais $15. pib espe;
cards;
cuba            2000 75.6
sri-lanka       2650 71.2
uzbekistan      2790 69
china           2946 70.5
peru            3110 63.6
tayikistan      2180 70
jordania        2895 67.3
filipinas       2440 64.6
mongolia        2250 63
congo           2800 51.7
kenia           1350 58.6
pakistan        1970 58.3
surinam         3072 69.9
bolivia         2170 60.5
gabon           3498 52.9
vietnam         1250 63.4
zimbabwe        2160 56.1
argelia         2870 65.6
egipto          3600 60.9
;
```

Gráfico preliminar

Se traza en primer lugar el cruce entre las dos variables (por supuesto, base posible para otras técnicas estadísticas como modelos lineales).

```
proc sort data=paises;by espe;
proc plot data=paises;plot pib*espe;
run;
```

```
            Plot of PIB*ESPE.  Legend: A = 1 obs, B = 2 obs, etc.

4000 +
     |
     |        A                A
     |
3000 +                              A       A     AA
     |      A                           A       A    A
 PIB |                                     A
     |                A         A     A              A
2000 +                     A                                    A
     |
     |
     |                      A         A
1000 +
     ---+----------+----------+----------+----------+----------+----------+----------+--
       50         55         60         65         70         75         80

                                      ESPE
```

Análisis de conglomerados

Como en casos anteriores, se realizan varios análisis paralelos para determinar el número de clusters ideal.

```
proc cluster data=paises method=ward std pseudo outtree=pais2;
var pib espe;
id pais;
run;

proc tree pos=20;
run;
```

Ward's Minimum Variance Cluster Analysis

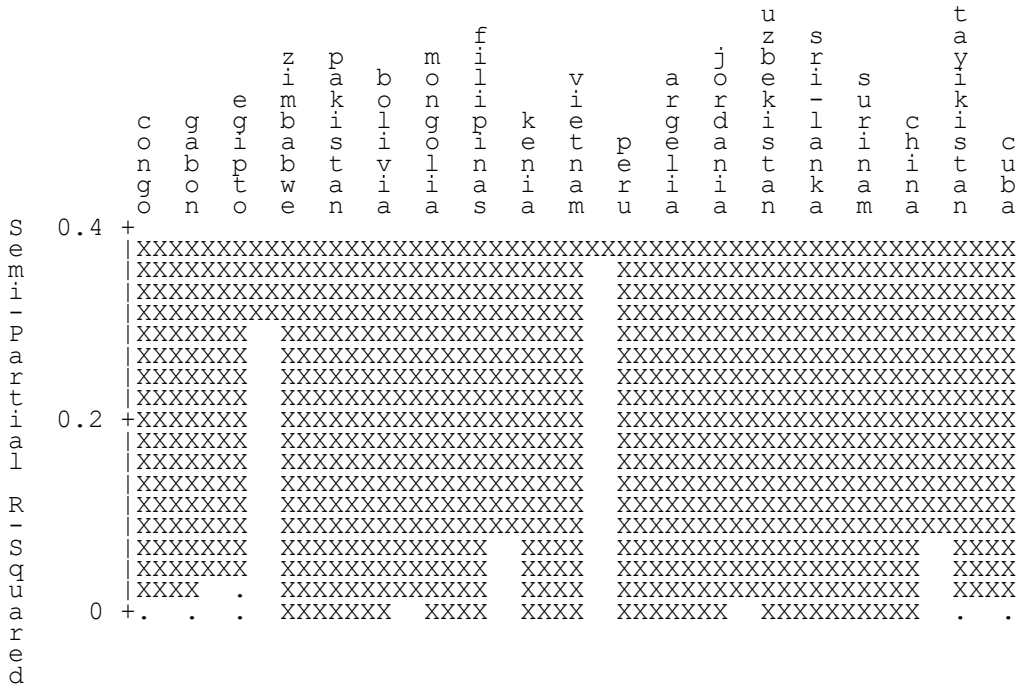
Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	1.01761	0.035223	0.508806	0.50881
2	0.98239	.	0.491194	1.00000

The data have been standardized to mean 0 and variance 1
Root-Mean-Square Total-Sample Standard Deviation = 1
Root-Mean-Square Distance Between Observations = 2

NCL	Clusters Joined		FREQ	SPRSQ	RSQ	PSF	PST2	Tie
18	surinam	china	2	0.000656	0.9993	89.7	.	
17	argelia	jordania	2	0.000972	0.9984	76.7	.	
16	mongolia	filipinas	2	0.002064	0.9963	54.0	.	
15	uzbekistan	sri-lanka	2	0.002256	0.9941	47.8	.	
14	zimbabwe	pakistan	2	0.002814	0.9912	43.5	.	
13	CL14	bolivia	3	0.005276	0.9860	35.1	1.9	
12	CL15	CL18	4	0.005658	0.9803	31.7	3.9	
11	peru	CL17	3	0.005899	0.9744	30.5	6.1	
10	kenia	vietnam	2	0.007921	0.9665	28.8	.	
9	tayikistan	cuba	2	0.011417	0.9551	26.6	.	
8	congo	gabon	2	0.016957	0.9381	23.8	.	
7	CL11	CL12	7	0.025419	0.9127	20.9	8.2	
6	CL13	CL16	5	0.028767	0.8839	19.8	8.5	
5	CL8	egipto	3	0.041629	0.8423	18.7	2.5	
4	CL6	CL10	7	0.078185	0.7641	16.2	8.3	
3	CL7	CL9	9	0.091935	0.6722	16.4	12.3	
2	CL5	CL4	10	0.303461	0.3687	9.9	13.2	
1	CL2	CL3	19	0.368713	0.0000	.	9.9	

Ward's Minimum Variance Cluster Analysis



```
proc cluster data=paises method=average std pseudo outtree=pais2;
var pib espe;
id pais;
run;

proc tree pos=20;
run;
```

Average Linkage Cluster Analysis

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	1.01761	0.035223	0.508806	0.50881
2	0.98239	.	0.491194	1.00000

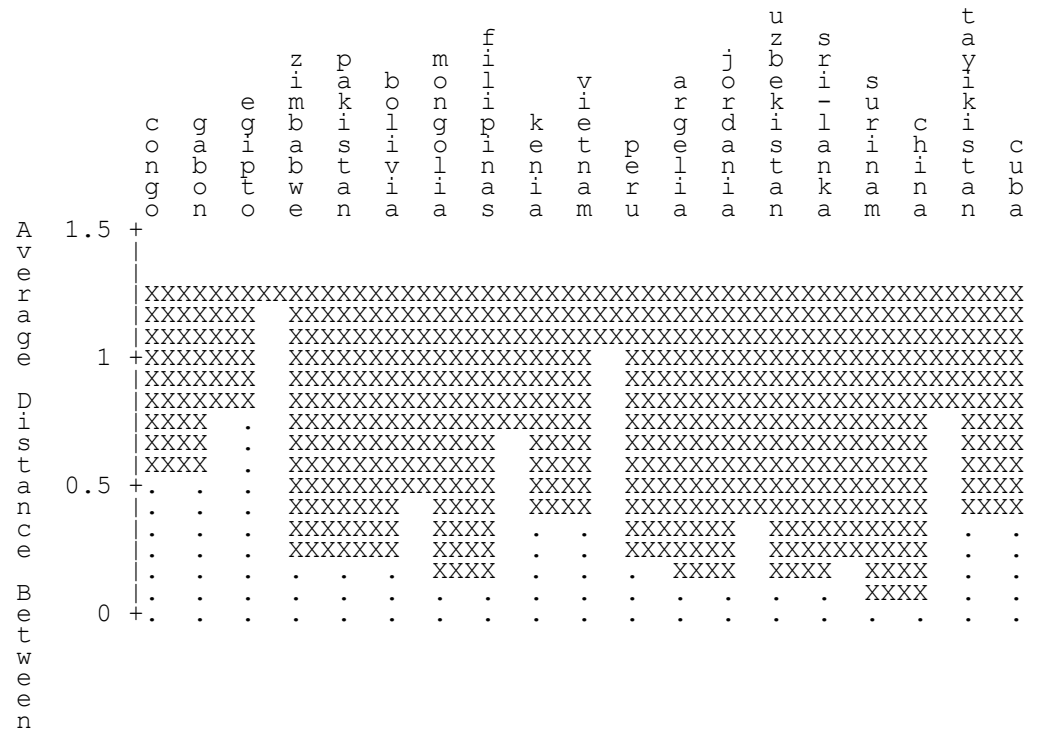
The data have been standardized to mean 0 and variance 1
Root-Mean-Square Total-Sample Standard Deviation = 1
Root-Mean-Square Distance Between Observations = 2

Average Linkage Cluster Analysis

NCL	Clusters Joined		FREQ	Pseudo F	Pseudo t**2	Norm RMS Dist	Tie
18	surinam	china	2	89.7	.	0.108631	
17	argelia	jordania	2	76.7	.	0.132290	
16	mongolia	filipinas	2	54.0	.	0.192741	
15	uzbekistan	sri-lanka	2	47.8	.	0.201513	
14	zimbabwe	pakistan	2	43.5	.	0.225068	
13	CL15	CL18	4	34.2	3.9	0.253029	
12	CL14	bolivia	3	31.7	1.9	0.289636	
11	peru	CL17	3	30.5	6.1	0.289845	
10	kenia	vietnam	2	28.8	.	0.377592	
9	CL11	CL13	7	20.0	8.2	0.416348	
8	tayikistan	cuba	2	20.8	.	0.453325	
7	CL12	CL16	5	18.2	8.5	0.499309	
6	congo	gabon	2	19.8	.	0.552474	
5	CL7	CL10	7	14.5	8.3	0.773476	
4	CL9	CL8	9	12.5	12.3	0.797384	
3	CL6	egipto	3	16.4	2.5	0.798939	
2	CL5	CL4	16	8.0	18.3	1.053680	
1	CL3	CL2	19	.	8.0	1.290309	

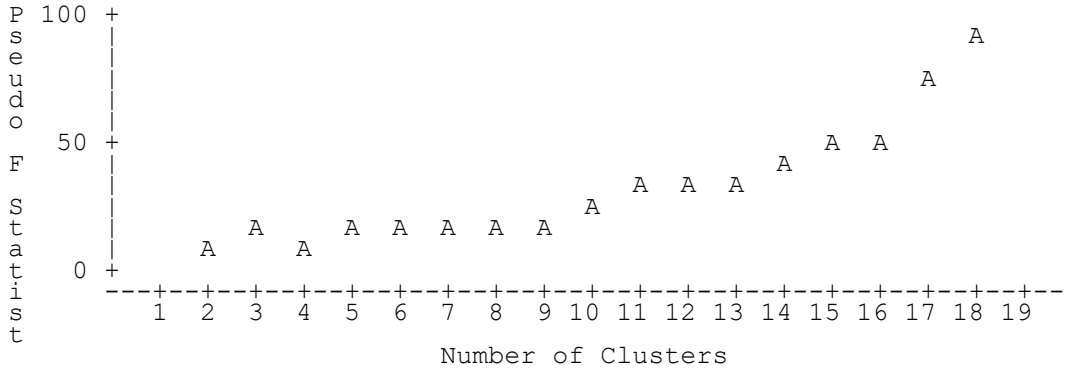
Average Linkage Cluster Analysis

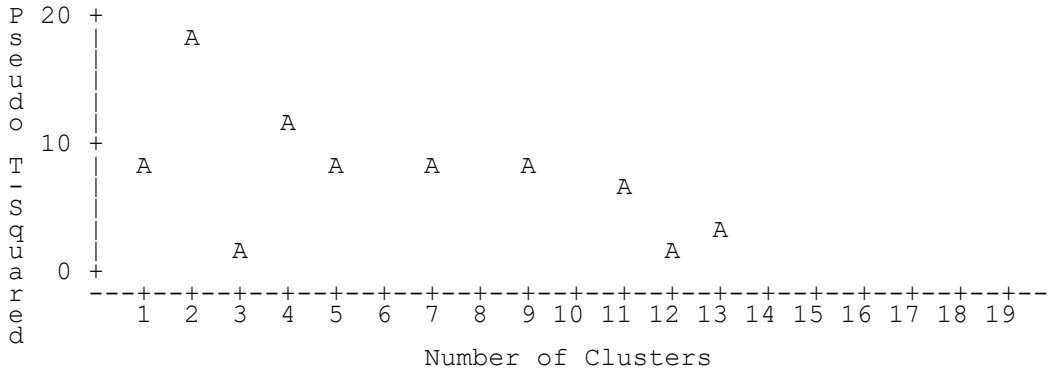
PAIS



```
proc sort data=pais2;by _ncl_;
run;

proc plot data=pais2;plot (_psf_ _pst2_)*_ncl_;
run;
```





El número de clusters es evidente

En los árboles derivados de los dos métodos se observa cómo la distancia que separa tres grupos es relativamente alta respecto a otras formaciones. En los gráficos de los estadísticos F y t también se observa un máximo local y mínimo local en 3, respectivamente. Se resume gráficamente y en forma de tablas la información final del proceso.

Esquemas finales sobre los grupos formados

En primer lugar se aporta un gráfico en el que se ve cómo están situados los países respecto a su PIB, Esperanza de vida y cluster al que pertenecen (opción plot pib*espe=cluster). Se ve cómo el **Cluster 1** presenta países con alto PIB y esperanza de vida, el **Cluster 2** con esperanza de vida media y bajo PIB, y el **Cluster 3**, al que pertenecen sólo los países Congo, Gabón y Egipto, con baja esperanza de vida pero relativamente alto PIB, debido principalmente al petróleo.

```
proc tree data=pais2 pos=20 out=pais3 n=3;copy pib espe pais;
id pais;


proc plot data=pais3;plot pib*espe=cluster;
run;
```

```
Plot of PIB*ESPE.  Symbol is value of CLUSTER.

4000 +
     |
     |         3                 3
     |
3000 +                                  1       1      11
     |      3                                1        1     1
 PIB |                                     2
     |                2          2      2                1
2000 +                     2                                        1
     |
     |
     |                       2          2
1000 +
     ---+----------+----------+----------+----------+----------+----------+----------+--
       50         55         60         65         70         75         80

                                      ESPE
```

Los datos precisos sobre la estructura de los clusters vienen dados a continuación, utilizando para ello los procedimientos PRINT y TABULATE.

```
proc print data=pais3;by cluster;var pais;
run;


proc tabulate data=pais3;class cluster;var pib espe;
table cluster, (pib espe)*mean n;
run;
```

```
--------------------------------- CLUSTER=1 ----------------------------------

                              OBS    PAIS

                                1    surinam
                                2    china
                                3    argelia
                                4    jordania
                                5    uzbekistan
                                6    sri-lanka
                                7    peru
                                8    tayikistan
                                9    cuba
--------------------------------- CLUSTER=2 ----------------------------------

                               OBS    PAIS

                                10    mongolia
                                11    filipinas
                                12    zimbabwe
                                13    pakistan
                                14    bolivia
                                15    kenia
                                16    vietnam
--------------------------------- CLUSTER=3 ----------------------------------

                               OBS     PAIS

                                17    congo
                                18    gabon
                                19    egipto
```

	PIB	ESPE	
	MEAN	MEAN	N
CLUSTER			
1	2723.67	69.19	9.00
2	1941.43	60.64	7.00
3	3299.33	55.17	3.00

BIBLIOGRAFÍA

Spath, Helmut	C*luster Analysis Algorithms*
Hair, J. y otros	*Multivariate Data Analysis with readings*
C. Chattfield, A. Collins	*Introduction to Multivariate Analysis*
Cuadras, C. M.	*Métodos de Análisis Multivariante*
Bisquerra, Rafael	*Introducción al Análisis Multivariante*

Manual del SAS/STAT: PROC CLUSTER, PROC TREE, PROC TABULATE.

Manual del SPSS: comando CLUSTER