COMPLEMENTOS DE FORMACIÓN EN TÉCNICAS DE MINERÍA DE DATOS

CAIO FERNANDES MORENO

UCM - Mineria de Datos

ANÁLISIS DE CLUSTERS

# 1. Los datos.

Los datos utilizados en este trabajo son datos datos de algunos paises.

Para importar los datos se ha utilizado el codigo sas abajo:

**PROC** **IMPORT** OUT= ucm.paises

DATAFILE= "C:\Users\win\Documents\GitHub\ucm\complementos\trabajocomplementos31enero16\DatosPaises-SAS.xls"

DBMS=EXCEL5 REPLACE;

GETNAMES=YES;

**RUN**;

|  |
| --- |
| The SAS System |

| **Obs** | **PAIS** | **POBL** | **NATALIDA** | **ESPERANZ** | **MORTALID** | **BALANZAC** | **PIB** | **PRODCERE** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | Afganistán | 27963 | 35.6 | 59.8 | 8.6 | -4766 | 566 | 157.13532 |
| **2** | Albania | 2902 | 13.1 | 77.5 | 7.2 | -2861 | 3786 | 577.685262 |
| **3** | Alemania | 80435 | 8.3 | 80.7 | 10.8 | 205408 | 41100 | 2659.28619 |
| **4** | Angola | 21220 | 46.2 | 51.7 | 14.2 | 29864 | 4221 | 19.5415273 |
| **5** | Arabia Saudita | 28091 | 20.8 | 74.1 | 3.4 | 144283 | 19327 | 10.288944 |
| **6** | Argelia | 36036 | 25.1 | 74.4 | 5.1 | 17558 | 4350 | 113.11257 |
| **7** | Argentina | 41223 | 17.8 | 76 | 7.6 | 12057 | 11508 | 310.531847 |
| **8** | Armenia | 2963 | 13.3 | 74.6 | 9 | -2771 | 3125 | 191.307784 |
| **9** | Australia | 22163 | 13.5 | 82.1 | 6.7 | 10724 | 57593 | 82.6334118 |
| **10** | Austria | 8392 | 9.5 | 81.1 | 9.4 | -5712 | 46377 | 1593.5443 |

# 2. Ejecución

En el enunciado del ejercicio se pide para cargar el archivo y despues trabajar con una muestra de 100 paises utilizando el procedimiento en SAS llamado **proc surveyselect**, el codigo abajo se ejecuta lo que se ha pedido:

/\*

El archivo \DatosPaises.xlsx" (que podeis descargar del campus) contiene informacion sobre 7 variables socioeconomicas de 133 paises. Seleccionar aleatoriamente una muestra de 100 paises con el procedimiento surveyselect de la siguiente forma:

\*/

libname ucm 'C:\Users\win\Documents\GitHub\ucm\complementos\trabajocomplementos31enero16\';

**data** paises;

set ucm.paises;

**run**;

**proc** **print** data=paises (obs=**100**);

**run**;

**proc** **contents** data=paises out=sa;

**data**;set sa;if \_n\_=**1** then put 'LISTA DE VARIABLES CONTINUAS';if type=**1** then put name @@;**run**;

**data**;set sa;if \_n\_=**1** then put 'LISTA DE VARIABLES CATEGÓRICAS';if type=**2** then put name @@;**run**;

**proc** **surveyselect** data=paises method=srs n=**100** out=sample\_paises;

**run**;

Tambien se pide para trabajar solo con las variables POBL, NATALIDA, ESPERANZ e MORTALID.

/\*

Para la muestra obtenida, realizar un Analisis Cluster incluyendo SOLO

las variables demograficas (Pobl Natalidad EsperanzaVida Mortalidad

), que debe incluir como minimo:

\*/

**proc** **print** data=sample\_paises (obs=**100**);

var POBL NATALIDA ESPERANZ MORTALID;

**run**;

Por las variables no teneren la misma unidad de medida es necesario normalizar los datos. Esto se hace en SAS utilizando el procedimiento **proc stdize** para normalizar las variables POBL, NATALIDA, ESPERANZ e MORTALID.

Codigo SAS:

**proc** **stdize** data=sample\_paises out=sample\_paisesnorm;

var POBL NATALIDA ESPERANZ MORTALID;

**run**;

**proc** **print** data=sample\_paisesnorm;

var POBL NATALIDA ESPERANZ MORTALID;

**run**;

**proc** **print** data=sample\_paisesnorm;

**run**;

Otra etapa que se ha hecho ha sido estudiar las correlaciones entre las 4 variables (POBL, NATALIDA, ESPERANZ e MORTALID), pero en esta etapa se ha utilizado la herramienta R y SAS.

En el resultado del estudio se ha encontrado una correlacion muy alta entre EsperanzaVida y Natalidad de **-0.87063840**, se percibe que cuanto mayor es la Esperanza de Vida menos Natalidad hay en un pais. En R he utilizado las 133 observaciones, pero con SAS solo estudiamos las correlaciones de una amuestra de 100 observaciones.

Abajo la tabla de correlacion hecha en R y una figura para representar la correlacion:

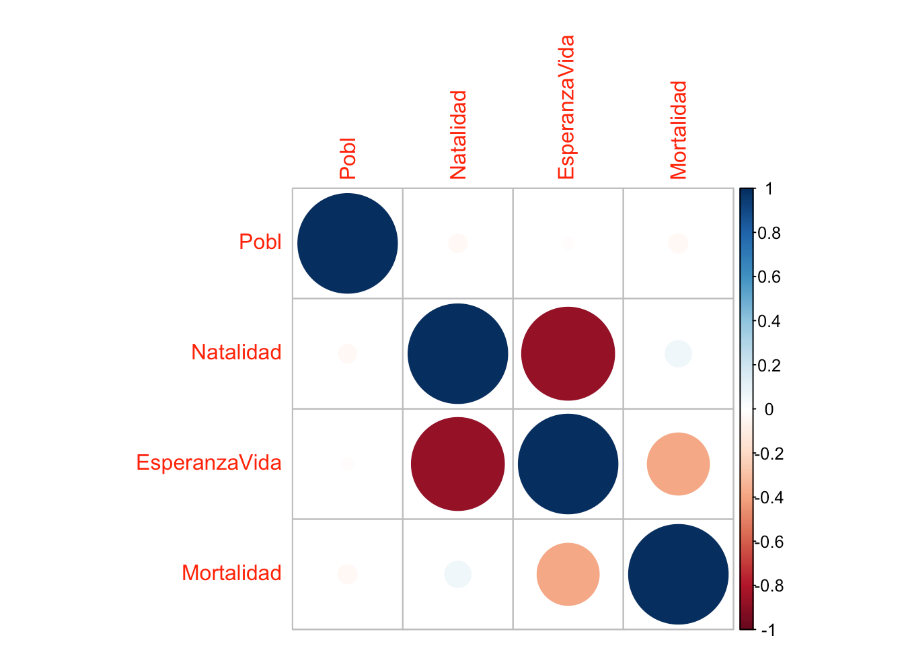
## Pobl Natalidad EsperanzaVida Mortalidad

## Pobl 1.00000000 -0.03243038 -0.01445153 -0.03519014

## Natalidad -0.03243038 1.00000000 -0.87063840 0.06833273

## EsperanzaVida -0.01445153 -0.87063840 1.00000000 -0.38664292

## Mortalidad -0.03519014 0.06833273 -0.38664292 1.00000000

****

En SAS para mirar la correlacion se hace con con el codigo abajo, con esto se analisa la correlacion y se puede ver tambien si hay datos atipicos o no.

**proc** **corr** data=sample\_paises outp=sample\_paisescorr;

var POBL NATALIDA ESPERANZ MORTALID;

**run**;

Resultados:

The CORR Procedure

|  |  |
| --- | --- |
| **4 Variables:** | POBL NATALIDA ESPERANZ MORTALID |

| **Simple Statistics** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **N** | **Mean** | **Std Dev** | **Sum** | **Minimum** | **Maximum** | **Label** |
| **POBL** | 100 | 27906 | 39987 | 2790573 | 622.00000 | 198615 | POBL |
| **NATALIDA** | 100 | 22.21200 | 11.28550 | 2221 | 8.30000 | 49.80000 | NATALIDAD |
| **ESPERANZ** | 100 | 70.68200 | 9.00345 | 7068 | 49.50000 | 83.30000 | ESPERANZAVIDA |
| **MORTALID** | 100 | 8.50600 | 2.84679 | 850.60000 | 2.70000 | 15.30000 | MORTALIDAD |

Se puede ver en las cuatro variables que hay datos atipicos en casos donde son muy altos y muy bajos.

Abajo se puede ver la correlacion hecha en SAS, el resultado es lo mismo con R con una pequena diferencia que con SAS tenemos una amuestra de 100 observaciones y con R he hecho con las 133 observaciones. Pero la correlacion es con Natalidad y Esperanza de Vida.

| **Pearson Correlation Coefficients, N = 100  Prob > |r| under H0: Rho=0** | | | | |
| --- | --- | --- | --- | --- |
|  | **POBL** | **NATALIDA** | **ESPERANZ** | **MORTALID** |
| |  | | --- | | **POBL** | | POBL | | |  | | --- | | 1.00000 | |  | | |  | | --- | | -0.04365 | | 0.6663 | | |  | | --- | | 0.00360 | | 0.9716 | | |  | | --- | | 0.00055 | | 0.9957 | |
| |  | | --- | | **NATALIDA** | | NATALIDAD | | |  | | --- | | -0.04365 | | 0.6663 | | |  | | --- | | 1.00000 | |  | | |  | | --- | | -0.88465 | | <.0001 | | |  | | --- | | 0.09054 | | 0.3703 | |
| |  | | --- | | **ESPERANZ** | | ESPERANZAVIDA | | |  | | --- | | 0.00360 | | 0.9716 | | |  | | --- | | -0.88465 | | <.0001 | | |  | | --- | | 1.00000 | |  | | |  | | --- | | -0.37909 | | 0.0001 | |
| |  | | --- | | **MORTALID** | | MORTALIDAD | | |  | | --- | | 0.00055 | | 0.9957 | | |  | | --- | | 0.09054 | | 0.3703 | | |  | | --- | | -0.37909 | | 0.0001 | | |  | | --- | | 1.00000 | |  | |

1. Analisis Cluster jerarquico con al menos dos metodos de agrupamiento. A partir de los procedimientos estudiados en clase, determinar el numero (o numeros) adecuado de grupos.

*“El análisis cluster (o de conglomerados) tiene como objetivo formar grupos de individuos con características similares con respecto a determinadas variables.”*

*“La idea básica es, a partir de un conjunto de individuos, crear grupos exluyentes 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).”*

Para hacer una analisis de cluster es necesario tres fases:

* Normalizar las variables;
* Ver la correlación entre las variables;
* Corregir el problema de los atípicos ya que distorsionarían la generación de cluster;

**Analisis de clusteres jerárquico**

Se ha utilizado los codigos SAS abajo para probar 2 metodos de agrupamiento (centroid y ward) que son los 2 mas utilizados.

El procedimento SAS para analise de cluster jerarquico es el proc cluster, se ha anadido los parametros pseudo, RSQUARE para que en los resultados se muestre la pseudo-T2 y la pseudo-F y el R2.

Metodo de centroide.

**proc** **cluster** data=sample\_paisesnorm method=centroid pseudo ccc RSQUARE

outtree=sample\_paisesnormC

print=**10** plots=den(VERTICAL);

var POBL NATALIDA ESPERANZ MORTALID;

**run**;

Resultados con el metodo centroide:

The CLUSTER Procedure

Centroid Hierarchical Cluster Analysis

| **Eigenvalues of the Covariance Matrix** | | | | |
| --- | --- | --- | --- | --- |
|  | **Eigenvalue** | **Difference** | **Proportion** | **Cumulative** |
| **1** | 1.96555872 | 0.94160487 | 0.4914 | 0.4914 |
| **2** | 1.02395385 | 0.08070507 | 0.2560 | 0.7474 |
| **3** | 0.94324878 | 0.87601013 | 0.2358 | 0.9832 |
| **4** | 0.06723865 |  | 0.0168 | 1.0000 |

|  |  |
| --- | --- |
| **Root-Mean-Square Total-Sample Standard Deviation** | 1 |

|  |  |
| --- | --- |
| **Root-Mean-Square Distance Between Observations** | 2.828427 |

| **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** | **Norm Centroid Distance** | **Tie** |
| **10** | **CL21** | **OB57** | 3 | 0.0036 | .824 | .821 | 0.26 | 46.8 | 2.4 | 0.515 |  |
| **9** | **CL10** | **CL54** | 6 | 0.0087 | .815 | .805 | 0.90 | 50.2 | 6.2 | 0.5347 |  |
| **8** | **CL16** | **CL11** | 53 | 0.0754 | .740 | .785 | -3.1 | 37.4 | 38.3 | 0.5386 |  |
| **7** | **CL13** | **OB12** | 3 | 0.0039 | .736 | .761 | -1.7 | 43.2 | 1.7 | 0.5409 |  |
| **6** | **CL8** | **CL9** | 59 | 0.0479 | .688 | .731 | -2.6 | 41.5 | 14.4 | 0.6635 |  |
| **5** | **CL6** | **CL15** | 66 | 0.0934 | .595 | .691 | -5.2 | 34.8 | 24.3 | 0.8594 |  |
| **4** | **CL12** | **CL14** | 30 | 0.0418 | .553 | .636 | -4.3 | 39.6 | 19.7 | 0.8751 |  |
| **3** | **CL4** | **CL5** | 96 | 0.3815 | .171 | .529 | -12 | 10.0 | 81.3 | 0.9568 |  |
| **2** | **CL7** | **OB29** | 4 | 0.0163 | .155 | .379 | -6.9 | 18.0 | 5.2 | 1.0385 |  |
| **1** | **CL3** | **CL2** | 100 | 0.1551 | .000 | .000 | 0.00 | . | 18.0 | 1.4141 |  |

Para determinar con cuantos clusters se debe quedar hay que mirar el máximo relativo de Pseudo F-Statistic, Pseudo t-Squared y el R-Squared mas grande que 70%.

Por mirar el Pseudo t-Squared se ve el valor 81,3 que es el máximo relativo, por esto me quedo con 4 clusters, pues el 81,3 es en el cluster 3 y es necesario añadir más 1, el resultado es 4.

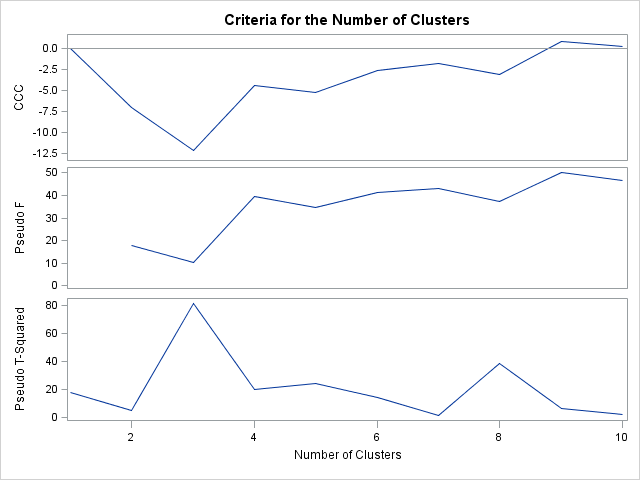
En el método Pseudo F-Statistic donde se elige el cluster por el máximo relativo entonces nos quedamos con 4.

Pero utilizando el R-Squared, dado que es mayor que 70% nos quedaríamos con 7 clusters.

Para el criterio ccc (**Cubic Clustering Criterion)** se puede ver valores negativos, esto es un indicador que hay dados atípicos. Valores de CCC más grandes de 2 a 3 indican buenos clusters.

He encontrado en el enlace abajo más detalles de como se puede interpretar el CCC.

<https://support.sas.com/documentation/onlinedoc/v82/techreport_a108.pdf>

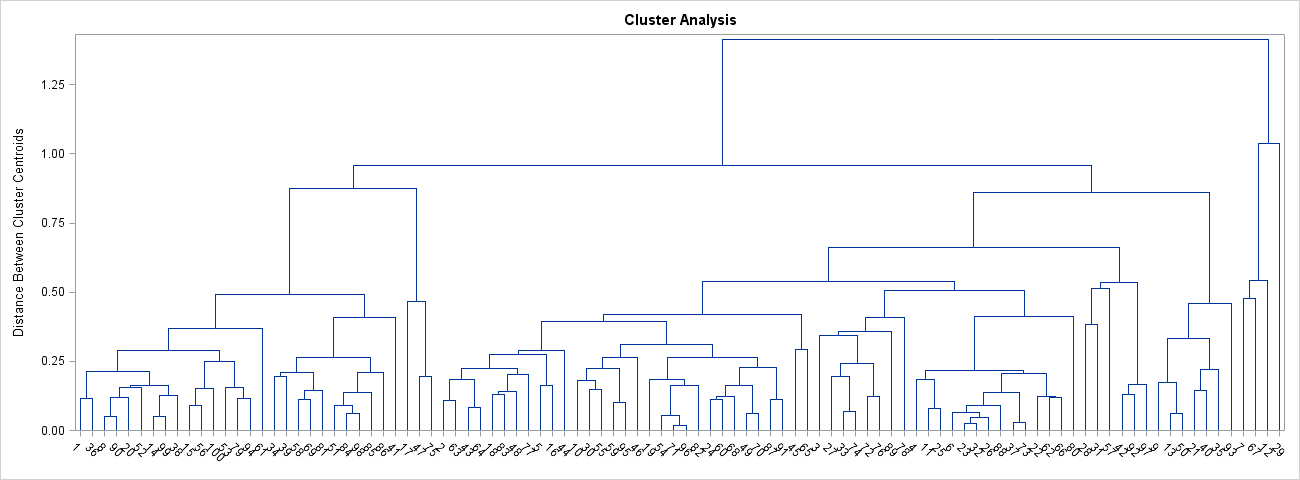


En los tres gráficos arriba se puede ver que lo mejor es quedar con 4 clusters.

Porque el Pseudo T-Squared tiene un pico en 3 clusters y se debe añadir más 1, entonces quedamos con 4.

Lo mismo con el Pseudo F que se puede concluir que se debe quedar con 4.

Dendograma del método de centroide.



Ahora probaremos con el Ward.

Metodo de ward.

**proc** **cluster** data=sample\_paisesnorm method=ward pseudo ccc RSQUARE

outtree=sample\_paisesnormW

print=**10** plots=den(VERTICAL);

var POBL NATALIDA ESPERANZ MORTALID;

**run**;

Los resultados son:

The CLUSTER Procedure

Ward's Minimum Variance Cluster Analysis

| **Eigenvalues of the Covariance Matrix** | | | | |
| --- | --- | --- | --- | --- |
|  | **Eigenvalue** | **Difference** | **Proportion** | **Cumulative** |
| **1** | 1.96555872 | 0.94160487 | 0.4914 | 0.4914 |
| **2** | 1.02395385 | 0.08070507 | 0.2560 | 0.7474 |
| **3** | 0.94324878 | 0.87601013 | 0.2358 | 0.9832 |
| **4** | 0.06723865 |  | 0.0168 | 1.0000 |

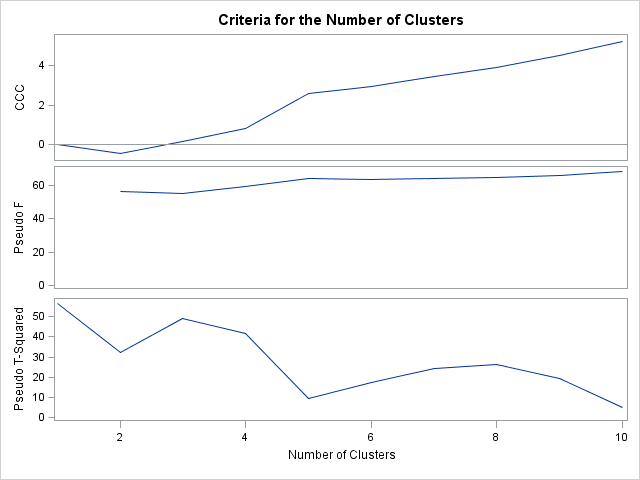
|  |  |
| --- | --- |
| **Root-Mean-Square Total-Sample Standard Deviation** | 1 |

|  |  |
| --- | --- |
| **Root-Mean-Square Distance Between Observations** | 2.828427 |

| **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** | **Tie** |
| **10** | **CL19** | **OB29** | 4 | 0.0163 | .872 | .821 | 5.22 | 68.2 | 5.2 |  |
| **9** | **CL15** | **CL13** | 28 | 0.0194 | .853 | .805 | 4.50 | 65.9 | 19.2 |  |
| **8** | **CL22** | **CL20** | 18 | 0.0220 | .831 | .785 | 3.92 | 64.5 | 26.1 |  |
| **7** | **CL14** | **CL24** | 19 | 0.0258 | .805 | .761 | 3.45 | 64.0 | 24.3 |  |
| **6** | **CL8** | **CL18** | 25 | 0.0334 | .772 | .731 | 2.93 | 63.5 | 17.6 |  |
| **5** | **CL10** | **CL12** | 11 | 0.0414 | .730 | .691 | 2.57 | 64.3 | 9.4 |  |
| **4** | **CL7** | **CL11** | 36 | 0.0802 | .650 | .636 | 0.81 | 59.4 | 41.4 |  |
| **3** | **CL9** | **CL6** | 53 | 0.1177 | .532 | .529 | 0.16 | 55.2 | 48.9 |  |
| **2** | **CL3** | **CL5** | 64 | 0.1664 | .366 | .379 | -.46 | 56.5 | 32.1 |  |
| **1** | **CL4** | **CL2** | 100 | 0.3659 | .000 | .000 | 0.00 | . | 56.5 |  |

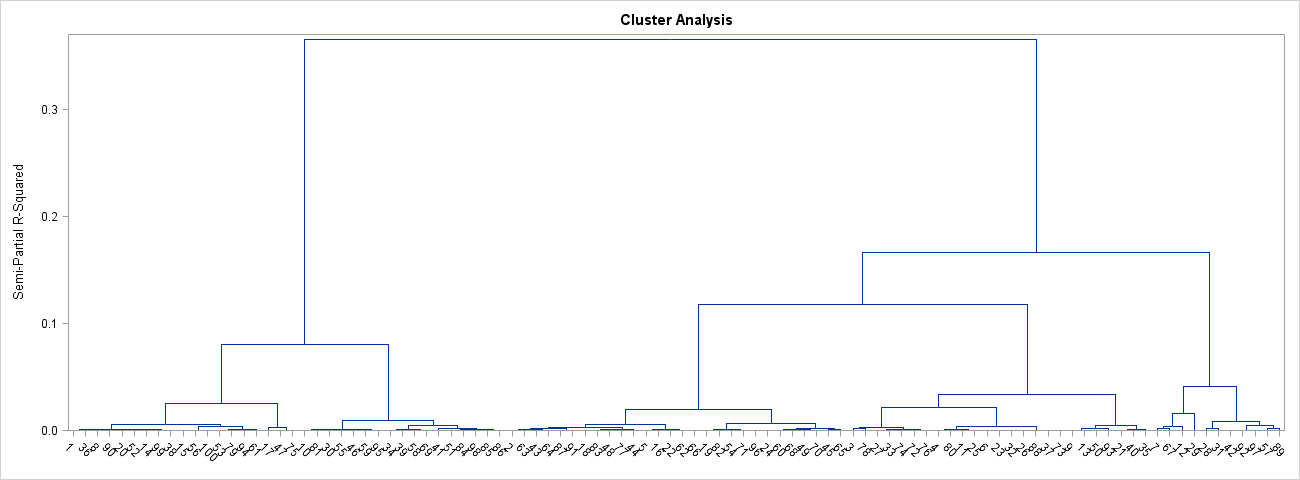
Por el Pseudo t-Squared se debe quedar con 4 clusters. Por el R-Square con 5.

Con el CCC quedo con 5 clusteres porque esta entre 2 y 3, el valor es 2.57.



Con el grafico arriba lo mismo 4 para el Pseudo T-Squared y 5 para el Pseudo F y CCC.

El dendograma con el metodo de Ward.



Dejo abajo el codigo para probar con el metodo average si se desea, pero no lo que utilizado.

**proc** **cluster** data=sample\_paisesnorm method=average pseudo RSQUARE ccc outtree=sample\_paisesnormA

print=**10** plots=den(VERTICAL);

var POBL NATALIDA ESPERANZ MORTALID;

**run**;

**Analisis de clusteres no jerárquico**

Abajo se puede ver los codigos SAS para analisis de clusters no jerárquico, donde se ha probado con grupos (clusters) de 4, 5 y 7, esto por que han sido las cantidades de grupos (clusters) que han salido mejor en la prueba de analisis de clusteres jerarquicos pero no se puede llegar a una conclusion exacta de la mejor cuantidad para se quedar.

Probando ambos jerárquico y no jerárquico la conclusión es que el número de grupos (clusters) que se debe quedar es 4.

Abajo se explicará como se ha llegado a esta conclusion, bien como los codigos y valores analisados.

**Prueba con grupos (clusteres) no jerárquico para 4 grupos.**

**PROC** **FASTCLUS** DATA=sample\_paisesnorm MAXCLUSTERS=**4** MEAN=MEDIAS2

DRIFT OUT=cluster4 maxiter=**30**;

var POBL NATALIDA ESPERANZ MORTALID;

**run**;

Resultados para la prueba con grupos (clusteres) no jerárquico para 4 grupos.

|  |  |
| --- | --- |
| **Pseudo F Statistic =** | 68.64 |

|  |  |
| --- | --- |
| **Approximate Expected Over-All R-Squared =** | 0.48331 |

|  |  |
| --- | --- |
| **Cubic Clustering Criterion =** | 14.195 |

Se ha puesto en amarillo los valores de Pseudo F Statistic y Cubic Clustering Criterion por que con 4 grupos (clusters) se ha obtenido los mejores valores comparados con 5 y 7 grupos.

**Prueba con grupos (clusteres) no jerárquico para 5 grupos.**

**PROC** **FASTCLUS** DATA=sample\_paisesnorm MAXCLUSTERS=**5** MEAN=MEDIAS2

DRIFT OUT=cluster5 maxiter=**30**;

var POBL NATALIDA ESPERANZ MORTALID;

**run**;

Prueba con grupos (clusteres) no jerárquico para 5 grupos.

|  |  |
| --- | --- |
| **Pseudo F Statistic =** | 58.69 |

|  |  |
| --- | --- |
| **Approximate Expected Over-All R-Squared =** | 0.58643 |

|  |  |
| --- | --- |
| **Cubic Clustering Criterion =** | 9.680 |

**Prueba con grupos (clusteres) no jerárquico para 7 grupos.**

**PROC** **FASTCLUS** DATA=sample\_paisesnorm MAXCLUSTERS=**7** MEAN=MEDIAS2

DRIFT OUT=cluster7 maxiter=**30**;

var POBL NATALIDA ESPERANZ MORTALID;

**run**;

Prueba con grupos (clusteres) no jerárquico para 7 grupos.

|  |  |
| --- | --- |
| **Pseudo F Statistic =** | 59.70 |

|  |  |
| --- | --- |
| **Approximate Expected Over-All R-Squared =** | 0.66546 |

|  |  |
| --- | --- |
| **Cubic Clustering Criterion =** | 11.147 |

Hasta el momento la conclusion es quedar con 4 grupos (clusters) porque para comparar los grupos (clusteres) hay que comparar los resultados de los valores Pseudo F Statistic y también el Cubic Clustering Criterion. Se sabe que cuanto más grande son el Pseudo F Statistic y el CCC (Cubic Clustering Criterion) mejor es el cluster.

Entonces como se puede ver en los resultados arriba los mas altos son del cluster 4.

**Ahora probaremos el teste de beale (contraste F de Beale):**

El codigo SAS abajo se puede probar el test de beale con grupos (clusters) de 4, 5 y 7.

**proc** **means** data=cluster4 ; var distance; output out=sumacuad4 uss=w4 ;

**run**;

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

**run**;

**proc** **means** data=cluster7 ; var distance; output out=sumacuad7 uss=w7 ;

**run**;

**data** beale;

merge sumacuad4 sumacuad5 sumacuad7;

k1=(\_freq\_-**4**)\*(**4**\*\*(-**2**/**8**));

k2=(\_freq\_-**5**)\*(**5**\*\*(-**2**/**8**));

k3=(\_freq\_-**7**)\*(**7**\*\*(-**2**/**8**));

fbeale1=(w4-w5)\*k2/(w5\*(k1-k2));

pvalor=**1**-probf(fbeale1,(k1-k2),k2);

fbeale2=(w4-w7)\*k3/(w7\*(k1-k3));

pvalor2=**1**-probf(fbeale2,(k1-k3),k3);

fbeale3=(w5-w7)\*k3/(w7\*(k2-k3));

pvalor3=**1**-probf(fbeale3,(k2-k3),k3);

**run**;

**proc** **print** data=beale;**run**;

Resultados del teste de Beale (contraste F de Beale):

| **Obs** | **\_TYPE\_** | **\_FREQ\_** | **w4** | **w5** | **w7** | **k1** | **k2** | **k3** | **fbeale1** | **pvalor** | **fbeale2** | **pvalor2** | **fbeale3** | **pvalor3** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | 0 | 100 | 125.909 | 114.090 | 81.6182 | 67.8823 | 63.5303 | 57.1753 | 1.51231 | 0.20544 | 2.89781 | .004545197 | 3.57938 | .003807254 |

Ahora haremos la interpretacion del resultado del teste de beale (contraste F de Beale).

**Comparativos de p-valores.**

fbeale1=(w4-w5)\*k2/(w5\*(k1-k2));

pvalor=**1**-probf(fbeale1,(k1-k2),k2);

fbeale2=(w4-w7)\*k3/(w7\*(k1-k3));

pvalor2=**1**-probf(fbeale2,(k1-k3),k3);

fbeale3=(w5-w7)\*k3/(w7\*(k2-k3));

pvalor3=**1**-probf(fbeale3,(k2-k3),k3);

**Valores de p-valores:**

Pvalor1 = 0.20544

Pvalor2 = .004545197

Pvalor3 = .003807254

El p-valor 1 (comparación de 4 clusters con 5) es 0.20544 es alto, entonces la comparación de 4 clusters con 5 clusters se dice que es muy difícil que 4 clusters sea peor que 5 clusters.

El p-valor2 es bajo siendo .004545197 donde se hace la comparación de 4 frente a 7, se entiende que si 4 clusters son mejores que 7.

El p-valor3 es bajo donde se compara 5 frente a 7, donde si 5 clusters son mejores que 7.

**Conclusión Final**

La conclusión final de la cantidad de grupos (clusters) baseado en el resultado de todos las pruebas es quedar con 4 grupos (clusters).

**Grupos (Clusters)**

El código SAS abajo se puede utilizar para estudiar los grupos.

**proc** **sort** data=cluster4 out=cluster4s;

by cluster;

**proc** **Freq** data=cluster4s;

by cluster; tables PAIS;

**run**;

**proc** **sort** data=cluster5 out=cluster5s;

by cluster;

**proc** **Freq** data=cluster5s;

by cluster; tables PAIS;

**run**;

**proc** **sort** data=cluster7 out=cluster7s;

by cluster;

**proc** **Freq** data=cluster7s;

by cluster; tables PAIS;

**run**;

**Pruebas con R**

He hecho algunas pruebas con R.

En el ejemplo abajo se utiliza el paquete NbClust para determinar el número correcto de grupos (clusters).

Enlace para el tutorial:

<http://www.inside-r.org/packages/cran/NbClust/docs/NbClust>

**Código R:**

setwd("/Users/caiomsouza/git/Bitbucket/ucm/COMPLEMENTOS\_DE\_FORMACION\_EN\_TECNICAS\_DE\_MINERIA\_DE\_DATOS/tareas-entregar/trabajo-31enero16")

paises <- read.csv(file="DatosPaises.csv",head=TRUE,sep=",")

#head(paises, 10)

#Dejar solo POBL NATALIDA ESPERANZ MORTALID

paises.valores <- paises

# Remove la columna Paises

paises.valores$Pais <- NULL

# Remove la columna BalanzaComercial

paises.valores$BalanzaComercial <- NULL

# Remove la columna PIB

paises.valores$PIB <- NULL

# Remove la columna ProdCereales

paises.valores$ProdCereales <- NULL

#head(paises.valores,10)

# Normaliza las variables

paises.valores.normalizar <- scale(paises.valores)

#head(paises.valores.normalizar, 10)

## Prueba el mejor cluster

data1 <- paises.valores

data2 <- paises.valores.normalizar

#data<-iris[,-c(5)]

#data <- data1

data <- data2

res<-NbClust(data, diss=NULL, distance = "euclidean", min.nc=2, max.nc=6,

method = "ward.D2", index = "kl")

res$All.index

res$Best.nc

res$Best.partition

res<-NbClust(data, diss=NULL, distance = "euclidean", min.nc=2, max.nc=6,

method = "kmeans", index = "hubert")

res$All.index

res<-NbClust(data, diss=NULL, distance = "manhattan", min.nc=2, max.nc=6,

method = "complete", index = "all")

res$All.index

res$Best.nc

res$All.CriticalValues

res$Best.partition

He probado los datos normalizados y no normalizados por curiosidad, pero si lo sé que por las unidades de medidas no tener las mismas es obligatorio normalizar los datos.

Con el método ward se recomienda 6 grupos/clusters, pero con kmeans y complete se recomienda 2 con los datos normalizados.

Cuando los datos no son normalizados se recomienda 5 grupos con ward y después con los otros métodos 2.

La conclusión es que el mejor número de grupos son 2.

Pero los valores son distintos cuando los valores están normalizados o no.

Con los datos normalizados tenemos las siguientes recomendaciones para el numero de clusters(grupos):

PseudoT2: 2

Test de Beale: 2

CCC: 2

Silhouette: 2

**Conclusiones con R:**

Se puede concluir que con R el número de grupos es de 2 a 6 grupos, pero se recomienda 2 grupos, lo que hemos visto con SAS que no es la mejor recomendación.

No se puede decir que R se equivoca porque no conozco este paquete **NbClust** o suficiente para quizás hacer un trabajo con la calidad exigida.

Quizás con más tiempo y conocimiento de R y del paquete **NbClust** se pudiera llegar a números y resultados más próximos de los conseguidos con SAS.

Pero si se puede decir que con R se recomienda de 2 a 6. Dejo esta análisis con R solo como un complemento al trabajo, pero tengo conocimiento que no ha sido pedido hacer nada con R, solo con SAS y con lo que hemos visto en clase.

**Explicando los grupos (Clusters)**

El código SAS abajo se imprime los grupos para que se pueda interpretar sus resultados y características de cada grupo.

Lo analizáremos solo los resultados con 4 grupos (clusters).

**proc** **sort** data=cluster4 out=cluster4s;

by cluster;

**proc** **Freq** data=cluster4s;

by cluster; tables PAIS;

**run**;

**proc** **print** data=cluster4;**run**;

Abajo los países que pertenecen a cada grupo (cluster).

**Grupo 1 (Cluster 1)**

Cluster=1

| **PAIS** | | | | |
| --- | --- | --- | --- | --- |
| **PAIS** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| **Afganistán** | 1 | 3.45 | 1 | 3.45 |
| **Benin** | 1 | 3.45 | 2 | 6.90 |
| **Burkina Faso** | 1 | 3.45 | 3 | 10.34 |
| **Camerún** | 1 | 3.45 | 4 | 13.79 |
| **Chad** | 1 | 3.45 | 5 | 17.24 |
| **Congo** | 1 | 3.45 | 6 | 20.69 |
| **Etiopía** | 1 | 3.45 | 7 | 24.14 |
| **Gabón** | 1 | 3.45 | 8 | 27.59 |
| **Ghana** | 1 | 3.45 | 9 | 31.03 |
| **Guinea** | 1 | 3.45 | 10 | 34.48 |
| **Haití** | 1 | 3.45 | 11 | 37.93 |
| **Lesoto** | 1 | 3.45 | 12 | 41.38 |
| **Madagascar** | 1 | 3.45 | 13 | 44.83 |
| **Malawi** | 1 | 3.45 | 14 | 48.28 |
| **Malí** | 1 | 3.45 | 15 | 51.72 |
| **Mozambique** | 1 | 3.45 | 16 | 55.17 |
| **Namibia** | 1 | 3.45 | 17 | 58.62 |
| **Níger** | 1 | 3.45 | 18 | 62.07 |
| **Papúa Nueva Guinea** | 1 | 3.45 | 19 | 65.52 |
| **Republica Checa** | 1 | 3.45 | 20 | 68.97 |
| **República Democrática Popular de Corea** | 1 | 3.45 | 21 | 72.41 |
| **Rumania** | 1 | 3.45 | 22 | 75.86 |
| **Senegal** | 1 | 3.45 | 23 | 79.31 |
| **Serbia** | 1 | 3.45 | 24 | 82.76 |
| **Togo** | 1 | 3.45 | 25 | 86.21 |
| **Uganda** | 1 | 3.45 | 26 | 89.66 |
| **Yemen** | 1 | 3.45 | 27 | 93.10 |
| **Zambia** | 1 | 3.45 | 28 | 96.55 |
| **Zimbabue** | 1 | 3.45 | 29 | 100.00 |

**Grupo 2 (Cluster 2)**

The FREQ Procedure

Cluster=2

| **PAIS** | | | | |
| --- | --- | --- | --- | --- |
| **PAIS** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| **Bangladesh** | 1 | 20.00 | 1 | 20.00 |
| **Brasil** | 1 | 20.00 | 2 | 40.00 |
| **Federación Rusa** | 1 | 20.00 | 3 | 60.00 |
| **Filipinas** | 1 | 20.00 | 4 | 80.00 |
| **Pakistán** | 1 | 20.00 | 5 | 100.00 |

**Grupo 3 (Cluster 3)**

The FREQ Procedure

Cluster=3

| **PAIS** | | | | |
| --- | --- | --- | --- | --- |
| **PAIS** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| **Albania** | 1 | 2.56 | 1 | 2.56 |
| **Australia** | 1 | 2.56 | 2 | 5.13 |
| **Bolivia (Estado Plurinacional de)** | 1 | 2.56 | 3 | 7.69 |
| **Canadá** | 1 | 2.56 | 4 | 10.26 |
| **Chile** | 1 | 2.56 | 5 | 12.82 |
| **Colombia** | 1 | 2.56 | 6 | 15.38 |
| **Cuba** | 1 | 2.56 | 7 | 17.95 |
| **Ecuador** | 1 | 2.56 | 8 | 20.51 |
| **Fiji** | 1 | 2.56 | 9 | 23.08 |
| **Irak** | 1 | 2.56 | 10 | 25.64 |
| **Irlanda** | 1 | 2.56 | 11 | 28.21 |
| **Irán (República islámica de)** | 1 | 2.56 | 12 | 30.77 |
| **Israel** | 1 | 2.56 | 13 | 33.33 |
| **Jordania** | 1 | 2.56 | 14 | 35.90 |
| **Kazajstán** | 1 | 2.56 | 15 | 38.46 |
| **Libia** | 1 | 2.56 | 16 | 41.03 |
| **Líbano** | 1 | 2.56 | 17 | 43.59 |
| **Marruecos** | 1 | 2.56 | 18 | 46.15 |
| **Mongolia** | 1 | 2.56 | 19 | 48.72 |
| **Myanmar** | 1 | 2.56 | 20 | 51.28 |
| **Nepal** | 1 | 2.56 | 21 | 53.85 |
| **Nicaragua** | 1 | 2.56 | 22 | 56.41 |
| **Nueva Caledonia** | 1 | 2.56 | 23 | 58.97 |
| **Nueva Zelanda** | 1 | 2.56 | 24 | 61.54 |
| **Omán** | 1 | 2.56 | 25 | 64.10 |
| **Panamá** | 1 | 2.56 | 26 | 66.67 |
| **Paraguay** | 1 | 2.56 | 27 | 69.23 |
| **Perú** | 1 | 2.56 | 28 | 71.79 |
| **República Dominicana** | 1 | 2.56 | 29 | 74.36 |
| **República Unida de Tanzanía** | 1 | 2.56 | 30 | 76.92 |
| **República de Moldova** | 1 | 2.56 | 31 | 79.49 |
| **Ruanda** | 1 | 2.56 | 32 | 82.05 |
| **Sri Lanka** | 1 | 2.56 | 33 | 84.62 |
| **Tailandia** | 1 | 2.56 | 34 | 87.18 |
| **Turquía** | 1 | 2.56 | 35 | 89.74 |
| **Túnez** | 1 | 2.56 | 36 | 92.31 |
| **Uzbekistán** | 1 | 2.56 | 37 | 94.87 |
| **Venezuela (República Bolivariana de)** | 1 | 2.56 | 38 | 97.44 |
| **Viet Nam** | 1 | 2.56 | 39 | 100.00 |

**Grupo 4 (Cluster 4)**

The FREQ Procedure

Cluster=4

| **PAIS** | | | | |
| --- | --- | --- | --- | --- |
| **PAIS** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| **Alemania** | 1 | 3.70 | 1 | 3.70 |
| **Armenia** | 1 | 3.70 | 2 | 7.41 |
| **Austria** | 1 | 3.70 | 3 | 11.11 |
| **Bielorrusia** | 1 | 3.70 | 4 | 14.81 |
| **Bosnia y Herzegovina** | 1 | 3.70 | 5 | 18.52 |
| **Bulgaria** | 1 | 3.70 | 6 | 22.22 |
| **Croacia** | 1 | 3.70 | 7 | 25.93 |
| **Dinamarca** | 1 | 3.70 | 8 | 29.63 |
| **Eslovaquia** | 1 | 3.70 | 9 | 33.33 |
| **Eslovenia** | 1 | 3.70 | 10 | 37.04 |
| **España** | 1 | 3.70 | 11 | 40.74 |
| **Finlandia** | 1 | 3.70 | 12 | 44.44 |
| **Francia** | 1 | 3.70 | 13 | 48.15 |
| **Georgia** | 1 | 3.70 | 14 | 51.85 |
| **Grecia** | 1 | 3.70 | 15 | 55.56 |
| **Hungría** | 1 | 3.70 | 16 | 59.26 |
| **Lituania** | 1 | 3.70 | 17 | 62.96 |
| **Noruega** | 1 | 3.70 | 18 | 66.67 |
| **Países Bajos** | 1 | 3.70 | 19 | 70.37 |
| **Polonia** | 1 | 3.70 | 20 | 74.07 |
| **Portugal** | 1 | 3.70 | 21 | 77.78 |
| **Reino Unido** | 1 | 3.70 | 22 | 81.48 |
| **República Democrática Popular Lao** | 1 | 3.70 | 23 | 85.19 |
| **República Democrática del Congo** | 1 | 3.70 | 24 | 88.89 |
| **República de Corea** | 1 | 3.70 | 25 | 92.59 |
| **Suecia** | 1 | 3.70 | 26 | 96.30 |
| **Ucrania** | 1 | 3.70 | 27 | 100.00 |

**Conclusiones de los 4 grupos.**

El código SAS abajo se imprime los grupos para que se pueda interpretar sus resultados y características de cada grupo.

Grupo 1: Paises con

Grupo 2: Paises

Grupo 3: Paises

Grupo 4: