# PRÁCTICA A TÉCNICAS PARA REDUCIR LA DIMENSIÓN Y CLUSTERING 2024

Los ficheros asignados en la tabla a cada alumno, contienen información socio- económica, de salud, educación, etc de países de la Unión europea.

Alumno: Alexandro Bazán

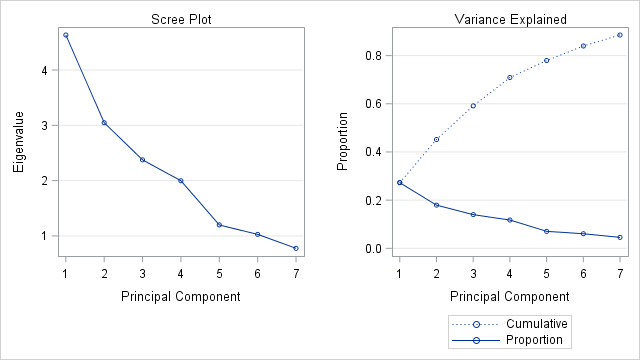
Análisis De Componentes Principales y Factorial:0

Para reducir el número de variables e intentar encontrar relaciones, tanto entre variables como entre países, realizar los siguientes apartados.

1. Realizar un análisis de componentes principales sobre la matriz de correlaciones. ¿Con cuantas componentes nos quedaríamos? (seguir el criterio de autovalores estrictamente mayor que 1)

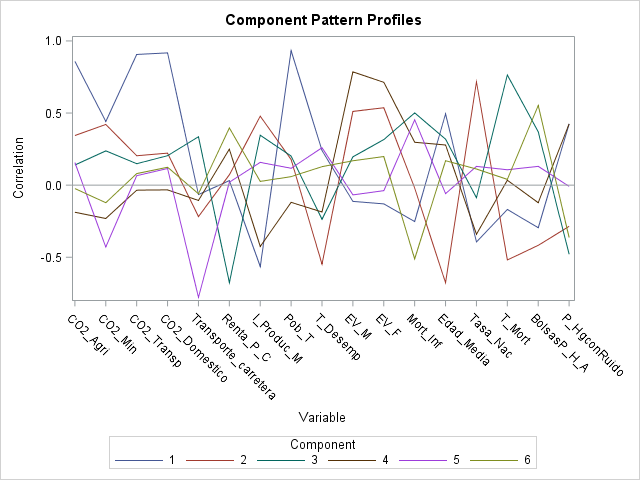
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Eigenvalues of the Correlation Matrix** | | | | |
|  | **Eigenvalue** | **Difference** | **Proportion** | **Cumulative** |
| **1** | 4.63332464 | 1.58664275 | 0.2725 | 0.2725 |
| **2** | 3.04668189 | 0.67149277 | 0.1792 | 0.4518 |
| **3** | 2.37518913 | 0.37721826 | 0.1397 | 0.5915 |
| **4** | 1.99797086 | 0.79923779 | 0.1175 | 0.7090 |
| **5** | 1.19873307 | 0.17053673 | 0.0705 | 0.7795 |
| **6** | 1.02819634 | 0.25210667 | 0.0605 | 0.8400 |
| **7** | 0.77608967 |  | 0.0457 | 0.8857 |

El número adecuado de componentes principales para este análisis es 6, ya que cumple con el criterio de Kaiser, estas tres componentes tienen valores propios mayores a 1 (4.63 , 3.046 , 2.375 , 1.997 , 1.198 y 1.028 ); segundo y explican en conjunto un alto porcentaje de la varianza total (84.00%), con contribuciones significativas de cada componente.



1. Hacer de nuevo el análisis, pero ahora indicando el número de componentes principales que hemos decidido retener. Sobre este análisis contestar los siguientes apartados.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Eigenvectors** | | | | | | | |
|  | | **Prin1** | **Prin2** | **Prin3** | **Prin4** | **Prin5** | **Prin6** |
| **CO2\_Agri** | CO2\_Agri | 0.398592 | 0.197322 | 0.090565 | -.133154 | 0.142754 | -.023644 |
| **CO2\_Min** | CO2\_Min | 0.205050 | 0.241062 | 0.154242 | -.163484 | -.392111 | -.120071 |
| **CO2\_Transp** | CO2\_Transp | 0.420910 | 0.116937 | 0.096422 | -.024760 | 0.059953 | 0.079065 |
| **CO2\_Domestico** | CO2\_Domestico | 0.426032 | 0.126992 | 0.132577 | -.023307 | 0.105118 | 0.122473 |
| **Transporte\_carretera** | Transporte carretera | -.030302 | -.125113 | 0.217701 | -.075153 | -.710842 | -.058810 |
| **Renta\_P\_C** | Renta\_P\_C | 0.014534 | 0.041033 | -.438834 | 0.176449 | 0.018827 | 0.391331 |
| **I\_Produc\_M** | I\_Produc\_M | -.262409 | 0.273901 | 0.224654 | -.301021 | 0.144457 | 0.025450 |
| **Pob\_T** | Pob\_T | 0.433014 | 0.101763 | 0.130973 | -.084343 | 0.106656 | 0.058580 |
| **T\_Desemp** | T\_Desemp | 0.111450 | -.313638 | -.153601 | -.131891 | 0.236987 | 0.126350 |
| **EV\_M** | EV\_M | -.052410 | 0.292935 | 0.127371 | 0.555591 | -.061393 | 0.166644 |
| **EV\_F** | EV\_F | -.060244 | 0.307671 | 0.205016 | 0.504569 | -.034922 | 0.195496 |
| **Mort\_Inf** | Mort\_Inf | -.117318 | -.011818 | 0.324930 | 0.210695 | 0.413920 | -.504750 |
| **Edad\_Media** | Edad Media | 0.229858 | -.387911 | 0.206985 | 0.196875 | -.054066 | 0.167126 |
| **Tasa\_Nac** | Tasa\_Nac | -.182996 | 0.410612 | -.056803 | -.240844 | 0.119218 | 0.110202 |
| **T\_Mort** | T\_Mort | -.078415 | -.297256 | 0.495612 | 0.023894 | 0.097233 | 0.039362 |
| **BolsasP\_H\_A** | BolsasP\_H\_A | -.137230 | -.239197 | 0.240272 | -.085977 | 0.119149 | 0.544395 |
| **P\_HgconRuido** | P\_HgconRuido | 0.197630 | -.163717 | -.310938 | 0.300150 | -.007899 | -.358301 |



* 1. ¿Cómo se calcula la Componente 2?

CP2=eT2 X\* = (0.197322CO2\_Agri\* ​)+(0.241062CO2\_Min\* ​)+(0.116937CO2\_Transp\* ​)+(0.126992CO2\_Domestico\* ​)+(-0.125113Transporte\_carretera\* )+( 0.041033Renta\_P\_C\*)+(0.273901I\_Produc\_M\* )+( 0.101763Pob\_T\* )+(−0.313638 T\_Desemp \* ​)+(0.292935 EV\_M \* ​)+(0.307671 EV\_F \* ​)+(-0.11818⋅ Mort\_Inf \*​) +(−0.387911 Edad\_Media\* )+(0.410612 Tasa\_Nac\*)+(-297256 T\_Mort\*

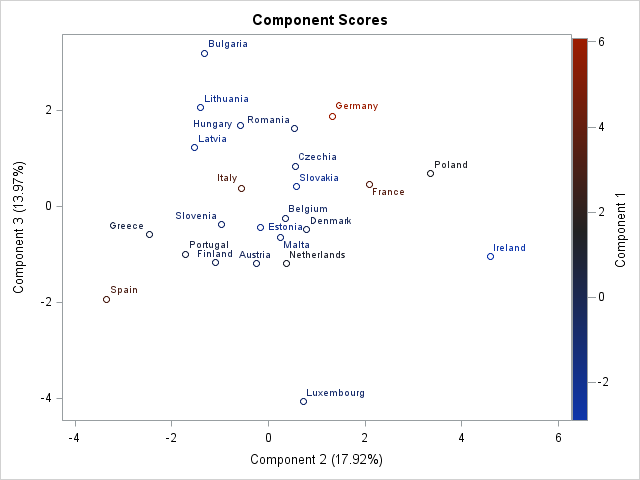
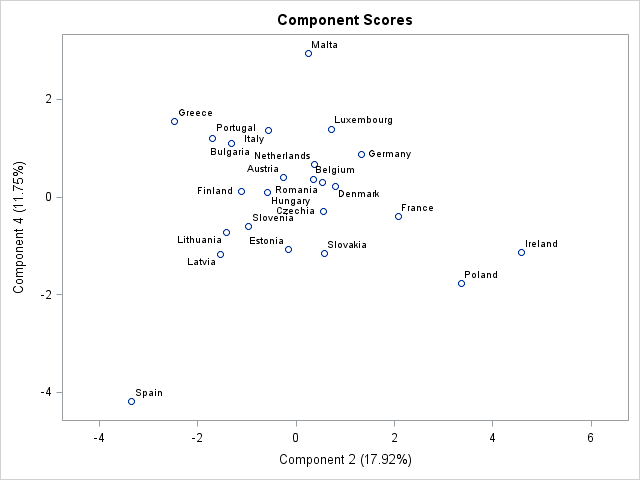
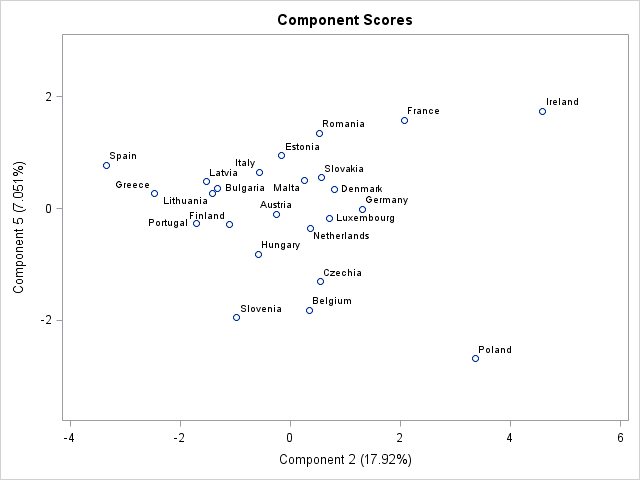
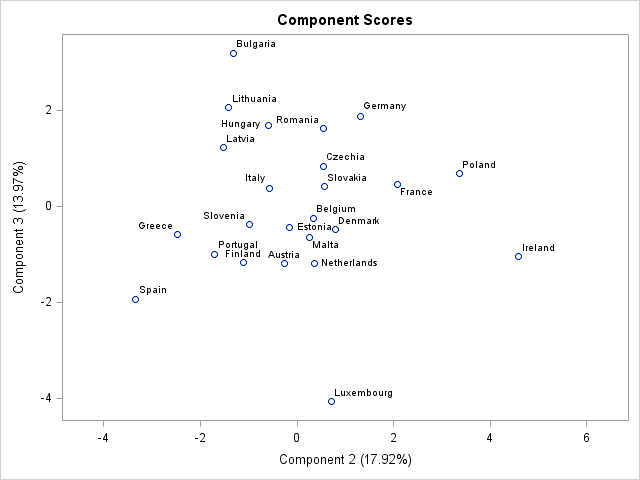
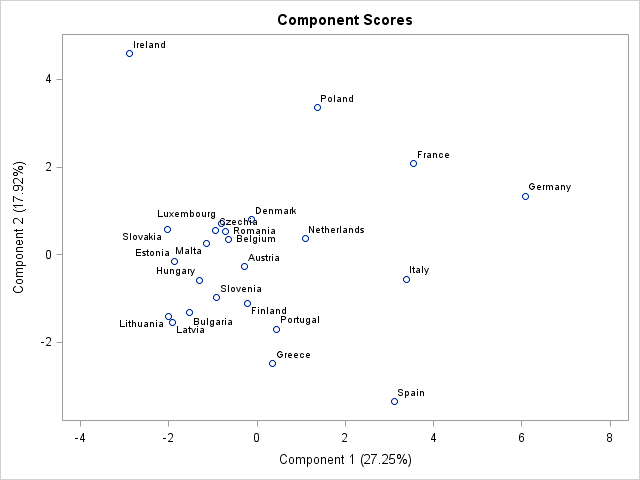
)+(-239197 BolsasP\_H\_A\*

)+(−0.163717P\_HgconRuido\*)

* 1. ¿Con que variables está más correlada?

Las variables positivas más altas en Prin2 son EV\_F (0.307671) y Tasa\_Nac (0.410612). Las variables con las cargas negativas más altas son T\_Desemp (-313638) y Edad\_Media (-387911). Así que estas serían las variables más fuertemente correlacionadas, ya sea positiva o negativamente, con la Componente 2.

* 1. ¿Qué país tiene mayor valor en dicha componente?



El país que tiene mayor valor en dicha componente es Irlanda.

1. Realizar un análisis Factorial.
   1. ¿Cuánto vale el MSA? ¿Qué variables presentan un peor valor? Si es necesario eliminar alguna variable para mejorar el MSA (hasta conseguir al menos 0.5).

Al principio , me da con las variables , un MSA de 0.34575900 , por lo que tengo que eliminar variables para conseguir al menos u MSA de 0.5

|  |  |  |
| --- | --- | --- |
| **Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.34575900** | **CO2\_Agri** | 0.5555579 |
| **CO2\_Min** | 0.1844543 |
| **CO2\_Transp** | 0.4232006 |
| **CO2\_Domestico** | 0.3880831 |
| **Transporte\_carretera** | 0.0989543 |
| **Renta\_P\_C** | 0.2500932 |
| **I\_Produc\_M** | 0.3391899 |
| **Pob\_T** | 0.4248354 |
| **T\_Desemp** | 0.4938527 |
| **EV\_M** | 0.2672229 |
| **EV\_F** | 0.2370938 |
| **Mort\_Inf** | 0.1512349 |
| **Edad\_Media** | 0.5791182 |
| **Tasa\_Nac** | 0.5442219 |
| **T\_Mort** | 0.4377959 |
| **BolsasP\_H\_A** | 0.3124401 |
| **P\_HgconRuido** | 0.2182464 |

Elimino Transporte\_carretera , Mort\_Inf y me da un MSA = **0.50514199** que es mayor a 0.5

|  |  |  |
| --- | --- | --- |
| **Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.50514199** | **CO2\_Agri** | 0.5762866 |
| **CO2\_Min** | 0.3991453 |
| **CO2\_Transp** | 0.6928153 |
| **CO2\_Domestico** | 0.5667236 |
| **Renta\_P\_C** | 0.2399347 |
| **I\_Produc\_M** | 0.5844185 |
| **Pob\_T** | 0.6205102 |
| **T\_Desemp** | 0.4483102 |
| **EV\_M** | 0.3383612 |
| **EV\_F** | 0.326246 |
| **Edad\_Media** | 0.5239881 |
| **Tasa\_Nac** | 0.461463 |
| **T\_Mort** | 0.4469588 |
| **BolsasP\_H\_A** | 0.4341915 |
| **P\_HgconRuido** | 0.5192563 |

1. Con el conjunto de variables que hemos decidido mantener decidir el número de factores adecuado.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Eigenvalues of the Correlation Matrix: Total = 14 Average = 1** | | | | |
|  | **Eigenvalue** | **Difference** | **Proportion** | **Cumulative** |
| **1** | 4.43338558 | 1.58784512 | 0.3167 | 0.3167 |
| **2** | 2.84554046 | 0.75820904 | 0.2033 | 0.5199 |
| **3** | 2.08733141 | 0.19245265 | 0.1491 | 0.6690 |
| **4** | 1.89487876 | 1.04133541 | 0.1353 | 0.8044 |
| **5** | 0.85354335 | 0.16781770 | 0.0610 | 0.8653 |
| **6** | 0.68572565 | 0.07639934 | 0.0490 | 0.9143 |
| **7** | 0.60932631 | 0.36360047 | 0.0435 | 0.9578 |
| **8** | 0.24572584 | 0.12068724 | 0.0176 | 0.9754 |
| **9** | 0.12503860 | 0.00682263 | 0.0089 | 0.9843 |
| **10** | 0.11821597 | 0.04966173 | 0.0084 | 0.9928 |
| **11** | 0.06855425 | 0.05021291 | 0.0049 | 0.9977 |
| **12** | 0.01834133 | 0.01018585 | 0.0013 | 0.9990 |
| **13** | 0.00815549 | 0.00191849 | 0.0006 | 0.9996 |
| **14** | 0.00623700 |  | 0.0004 | 1.0000 |

* 1. Realizar una rotación VARIMAX o QUARTIMAX (la que de mejor resultado). Comparar para la rotación escogida como han cambiado las cargas antes y después de la rotación.

Cargas antes de la rotación:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Factor Pattern** | | | | | |
|  | | **Factor1** | **Factor2** | **Factor3** | **Factor4** |
| **CO2\_Agri** | **CO2\_Agri** | 0.85921 | 0.35621 | 0.14752 | -0.17199 |
| **CO2\_Min** | **CO2\_Min** | 0.44183 | 0.45838 | 0.23075 | -0.12002 |
| **CO2\_Transp** | **CO2\_Transp** | 0.90732 | 0.21526 | 0.12301 | 0.00230 |
| **CO2\_Domestico** | **CO2\_Domestico** | 0.92210 | 0.23809 | 0.16534 | 0.01874 |
| **Renta\_P\_C** | **Renta\_P\_C** | -0.00352 | 0.01593 | -0.67816 | 0.01058 |
| **I\_Produc\_M** | **I\_Produc\_M** | -0.55831 | 0.49191 | 0.49056 | -0.25645 |
| **Pob\_T** | **Pob\_T** | 0.93783 | 0.19421 | 0.19247 | -0.07262 |
| **T\_Desemp** | **T\_Desemp** | 0.23426 | -0.56461 | -0.11442 | -0.28398 |
| **EV\_M** | **EV\_M** | -0.09835 | 0.51714 | -0.14943 | 0.81290 |
| **EV\_F** | **EV\_F** | -0.10925 | 0.55209 | -0.01806 | 0.78588 |
| **Edad\_Media** | **Edad Media** | 0.51292 | -0.65179 | 0.24092 | 0.42025 |
| **Tasa\_Nac** | **Tasa\_Nac** | -0.41158 | 0.70616 | 0.03380 | -0.34063 |
| **T\_Mort** | **T\_Mort** | -0.12352 | -0.46742 | 0.72440 | 0.32630 |
| **BolsasP\_H\_A** | **BolsasP\_H\_A** | -0.28711 | -0.40054 | 0.50461 | 0.12635 |
| **P\_HgconRuido** | **P\_HgconRuido** | 0.42018 | -0.32182 | -0.63533 | 0.14174 |

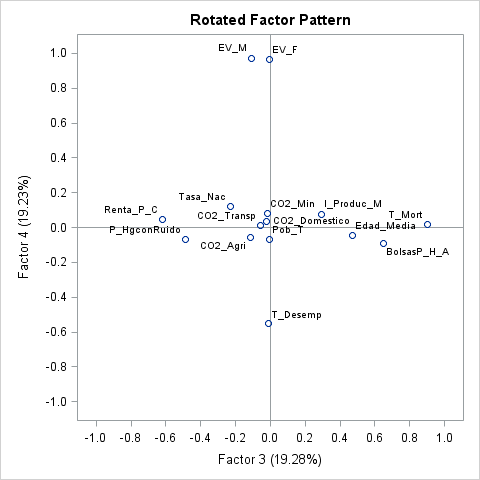
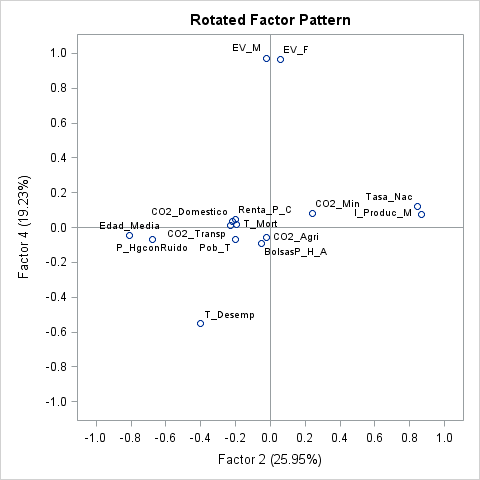
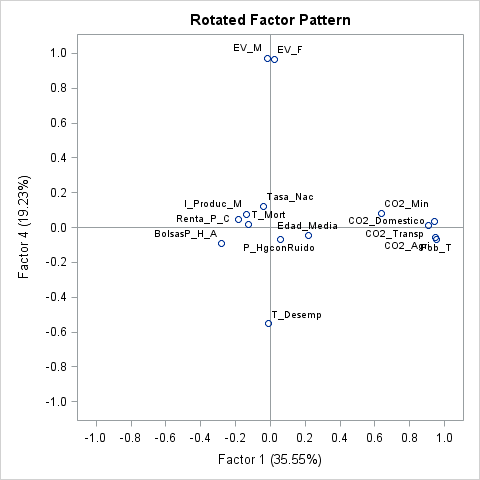
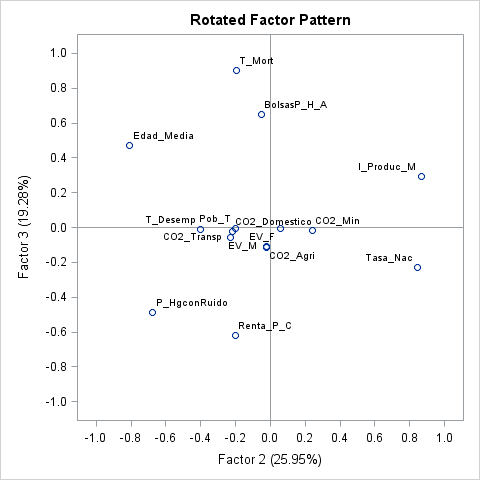
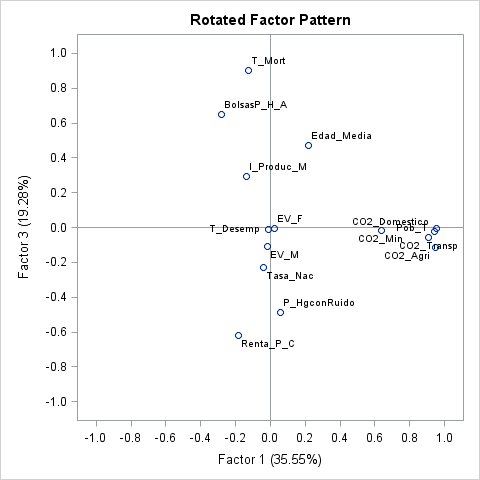
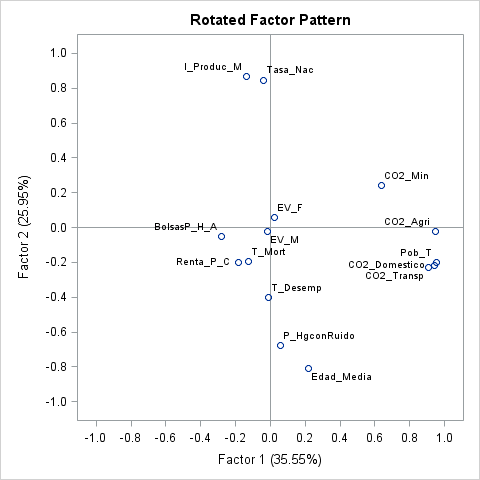
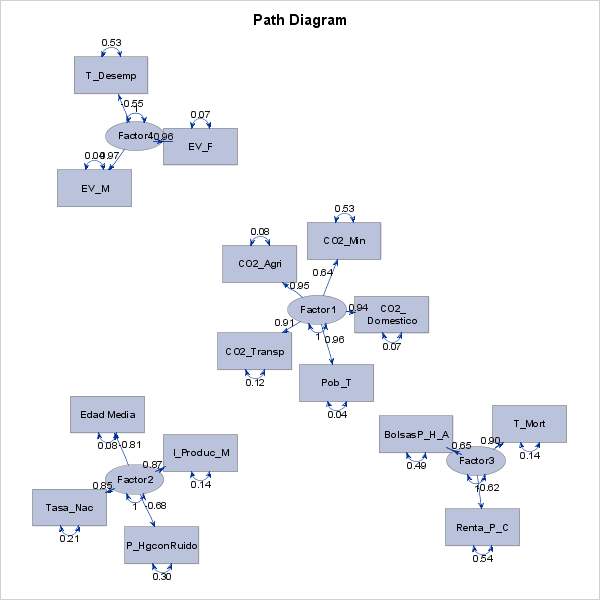
Con modelo : Varimax

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Rotated Factor Pattern** | | | | | |
|  | | **Factor1** | **Factor2** | **Factor3** | **Factor4** |
| **CO2\_Agri** | **CO2\_Agri** | 0.94837 | -0.02135 | -0.11500 | -0.05823 |
| **CO2\_Min** | **CO2\_Min** | 0.63872 | 0.24173 | -0.01446 | 0.07987 |
| **CO2\_Transp** | **CO2\_Transp** | 0.91023 | -0.22968 | -0.05725 | 0.01222 |
| **CO2\_Domestico** | **CO2\_Domestico** | 0.94154 | -0.21531 | -0.02397 | 0.03474 |
| **Renta\_P\_C** | **Renta\_P\_C** | -0.18164 | -0.20072 | -0.62050 | 0.04449 |
| **I\_Produc\_M** | **I\_Produc\_M** | -0.13873 | 0.86620 | 0.29069 | 0.07769 |
| **Pob\_T** | **Pob\_T** | 0.95607 | -0.20192 | -0.00751 | -0.06818 |
| **T\_Desemp** | **T\_Desemp** | -0.01054 | -0.40052 | -0.01146 | -0.55384 |
| **EV\_M** | **EV\_M** | -0.01938 | -0.02229 | -0.10832 | 0.97347 |
| **EV\_F** | **EV\_F** | 0.02305 | 0.05985 | -0.00487 | 0.96464 |
| **Edad\_Media** | **Edad Media** | 0.21716 | -0.80761 | 0.47018 | -0.04597 |
| **Tasa\_Nac** | **Tasa\_Nac** | -0.04216 | 0.84613 | -0.23048 | 0.11998 |
| **T\_Mort** | **T\_Mort** | -0.12814 | -0.19634 | 0.89980 | 0.01853 |
| **BolsasP\_H\_A** | **BolsasP\_H\_A** | -0.28277 | -0.04949 | 0.65032 | -0.09012 |
| **P\_HgconRuido** | **P\_HgconRuido** | 0.05672 | -0.67632 | -0.48872 | -0.06626 |

Escogemos Varimax debido a que las variables están más relacionadas con un único factor. Las variables CO2\_Agri y Pob\_T están claramente asignadas al Factor 1 , además si vemos las variables Tasa\_Nac y Edad\_Media están bien definidas en el Factor 2. Tras la rotación Varimax, las cargas factoriales se han redistribuido, logrando una mayor claridad en la asociación de las variables con cada factor. El Factor 1 se centra en el impacto ambiental (CO₂ y población total), el Factor 2 refleja industria y demografía (producción industrial y tasas de natalidad), el Factor 3 está relacionado con salud pública y problemas sociales (mortalidad y contaminación acústica), mientras que el Factor 4 se enfoca en la esperanza de vida y el desempleo

# Con el modelo factorial rotado:

1. Representar el Pathdiagram y los gráficos de las variables en los planos factoriales. ¿Qué representa cada factor?



 Factor 1: Impacto Ambiental.

 Factor 2: Demografía .

 Factor 3: Salud.

 Factor 4: Socioeconómica.

1. Sobre la tabla que contiene las cargas de los factores en las variables marcar para cada factor sobre las variables que más carga (>0.6) y comentar su signo. Escribir como sería la primera ecuación del modelo factorial.

Se indica cuales son mayores a 0.6

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Rotated Factor Pattern** | | | | | |
|  | | **Factor1** | **Factor2** | **Factor3** | **Factor4** |
| **CO2\_Agri** | **CO2\_Agri** | 0.94837 | -0.02135 | -0.11500 | -0.05823 |
| **CO2\_Min** | **CO2\_Min** | 0.63872 | 0.24173 | -0.01446 | 0.07987 |
| **CO2\_Transp** | **CO2\_Transp** | 0.91023 | -0.22968 | -0.05725 | 0.01222 |
| **CO2\_Domestico** | **CO2\_Domestico** | 0.94154 | -0.21531 | -0.02397 | 0.03474 |
| **Renta\_P\_C** | **Renta\_P\_C** | -0.18164 | -0.20072 | -0.62050 | 0.04449 |
| **I\_Produc\_M** | **I\_Produc\_M** | -0.13873 | 0.86620 | 0.29069 | 0.07769 |
| **Pob\_T** | **Pob\_T** | 0.95607 | -0.20192 | -0.00751 | -0.06818 |
| **T\_Desemp** | **T\_Desemp** | -0.01054 | -0.40052 | -0.01146 | -0.55384 |
| **EV\_M** | **EV\_M** | -0.01938 | -0.02229 | -0.10832 | 0.97347 |
| **EV\_F** | **EV\_F** | 0.02305 | 0.05985 | -0.00487 | 0.96464 |
| **Edad\_Media** | **Edad Media** | 0.21716 | -0.80761 | 0.47018 | -0.04597 |
| **Tasa\_Nac** | **Tasa\_Nac** | -0.04216 | 0.84613 | -0.23048 | 0.11998 |
| **T\_Mort** | **T\_Mort** | -0.12814 | -0.19634 | 0.89980 | 0.01853 |
| **BolsasP\_H\_A** | **BolsasP\_H\_A** | -0.28277 | -0.04949 | 0.65032 | -0.09012 |
| **P\_HgconRuido** | **P\_HgconRuido** | 0.05672 | -0.67632 | -0.48872 | -0.06626 |

CO2\_Agri\* = 0.94837 F1 -0.02135F2 -0.11500F3 -0.05823 F4

Cada factor tiene una relación clara con un conjunto específico de variables, y todas las asociaciones en este caso son positivas, según las cargas resaltadas en amarillo (>0.6).

1. ¿Qué coeficientes se utilizan para calcular el valor de los países en cada factor utilizando sus valores en las variables iniciales? Escribir la expresión resumida para calcular el valor del factor 1.

Factor1=(0.94837CO2\_Agri\*)+(0.63872 CO2\_Min\*)+(0.91023\* CO2\_Transp\*)+(0.94154CO2\_Domestico\*)+(-0.18164Renta\_P\_C\*)+(-0.13873I\_Produc\_M\*)+(0.95607Pob\_T)+(-0.01054\* T\_Desemp\*)+( -0.01938EV\_M\*)+(0.02305EV\_F)+( 0.21716Edad Media)+( -0.04216 Tasa\_Nac)+( -0.12814 T\_Mort)+( -0.28277 BolsasP\_H\_A) )+( 0.05672P\_HgconRuido)

1. A partir del fichero que contiene los valores que tienen los países en los nuevos factores, obtener una tabla en donde aparezcan los países:

España, Alemania y Grecia, y sus valores en los factores (solo estos).

Comentar que significado tienen estos valores

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Obs** | **Pais** | **Factor1** | **Factor2** | **Factor3** | **Factor4** |
| **1** | Germany | 2.86700 | 0.84852 | 0.78268 | 1.00252 |
| **2** | Greece | 0.18583 | -1.42901 | -0.77090 | 0.81424 |
| **3** | Spain | 1.39567 | -1.98422 | -0.02375 | -3.45855 |

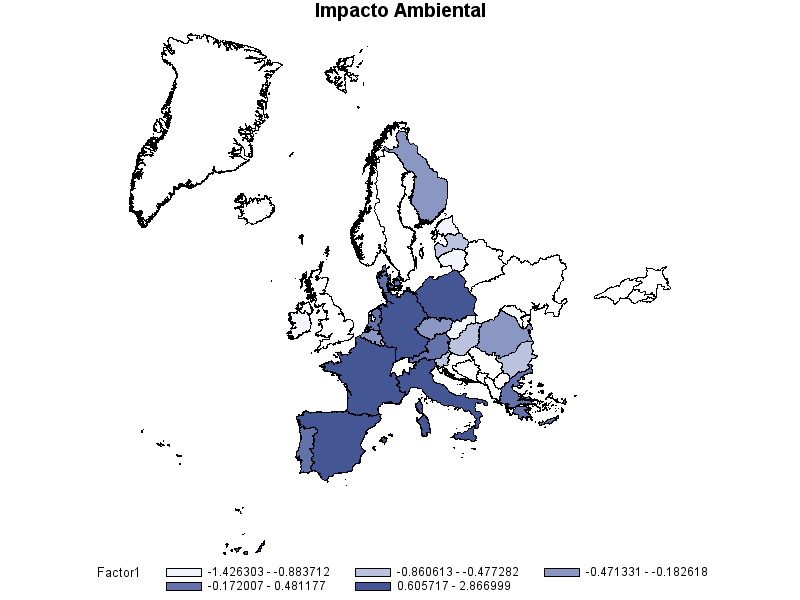
El Factor 1, denominado "Impacto Ambiental", refleja la relación entre las emisiones de CO2 provenientes de diversas fuentes (agricultura, minería, transporte y doméstico) y la población total. En este sentido, Alemania se destaca significativamente, con un valor muy por encima de la media, lo que sugiere un impacto ambiental considerable, probablemente debido a su alta industrialización y políticas medioambientales. Grecia, por otro lado, se sitúa ligeramente por encima de la media, lo que indica un desempeño moderado en términos de impacto ambiental. Finalmente, España presenta un desempeño positivo, con un valor intermedio entre Alemania y Grecia, lo que refleja una gestión ambiental relativamente buena, aunque sin llegar al nivel destacado de Alemania.

El Factor 2 , denominado “Demografia , refleja que Alemania (2.86700) presenta la mejor situación demográfica, posiblemente tasas de natalidad más equilibradas, mientras España (1.39567) muestra desafíos significativos por envejecimiento poblacional, y Grecia (0.18583) se mantiene en una posición intermedia con tendencias demográficas moderadas.

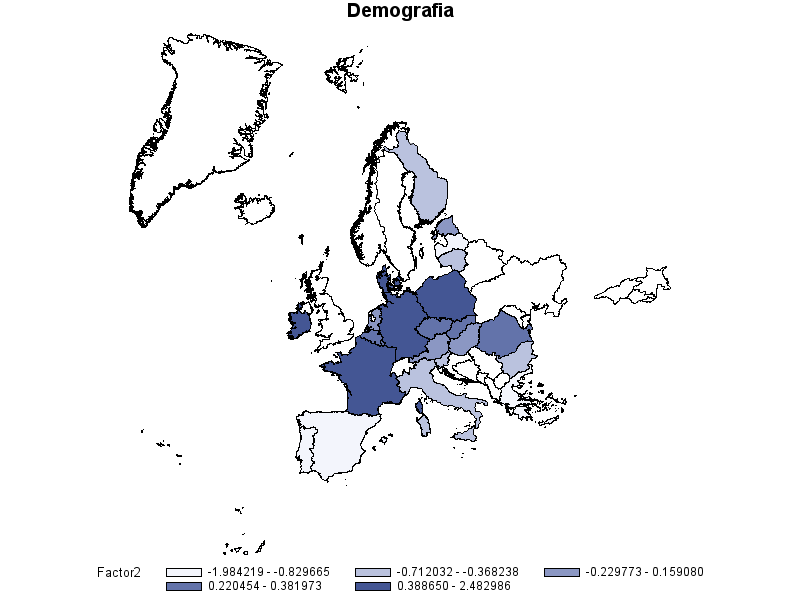
El Factor 3 , denominado “Salud” , Alemania (0.78268) lidera en condiciones de salud con un sistema médico robusto y alta inversión, España (-0.02375) mantiene un nivel promedio con capacidades sanitarias estables, y Grecia (-0.77090) evidencia mayores problemas estructurales en su sistema de salud

El Factor 4 , denominado “Socioeconómica” , Alemania (1.00252) muestra una economía sólida con indicadores positivos, Grecia (0.81424) presenta una situación económica intermedia, mientras España (-3.45855) enfrenta severos desafíos socioeconómicos, posiblemente derivados de una recuperación lenta tras crisis económicas recientes.

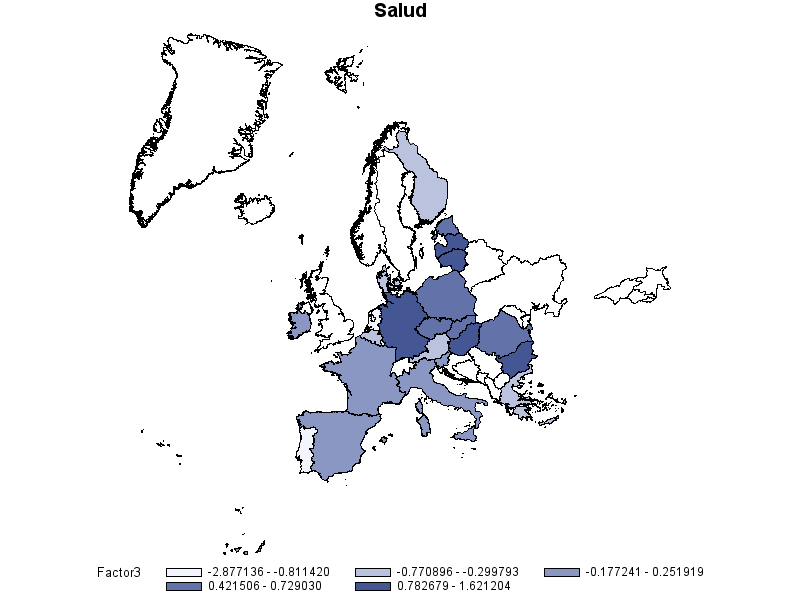
1. Representar el mapa de Europa por países coloreado según el valor del Factor 1, Factor2, Factor 3 y Factor4 (solo para estos independientemente del número de Factores). Comentar cada uno de los gráficos.



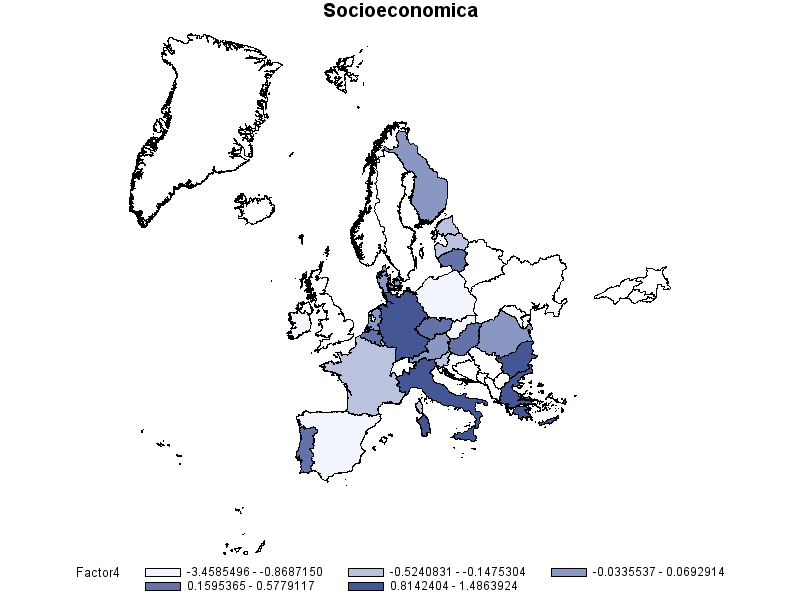
El gráfico del impacto ambiental muestra que los países del sur y oeste de Europa, representados por los tonos más oscuros como España , Francia Alemania Polonia e Italia . Esto podría estar relacionado con su alta población, un mayor nivel de consumo energético y actividades económicas intensivas en recursos. Por otro lado, los países del norte de Europa, como los escandinavos, presentan un menor impacto ambiental, probablemente debido a políticas más sostenibles, un uso eficiente de recursos y un enfoque en energías renovables.



El mapa muestra los valores del factor demográfico en los países europeos. Los tonos azules más oscuros en Europa Occidental y Central indican valores más altos, posiblemente debido a tasas de natalidad más elevadas han contribuido a un crecimiento demográfico más rápido en esas regiones. Por el contrario, los tonos azules más claros en el Sur, Este y Norte de Europa sugieren valores más bajos, relacionados quizás con tendencias de envejecimiento poblacional y bajas tasas de natalidad. El gradiente de color refleja la variación gradual de las características demográficas a través de las diferentes subregiones europeas.



Los países: Alemania , Hunguia , Letonia , Lituania y Bulgaria (tonos oscuros), indicando mejores condiciones de salud gracias una mayor renta y menores tasas de mortalidad. Irlanda y los países del sur de Europa, como España, Francia , Italia y Grecia, presentan valores intermedios, que reflejan un equilibrio relativo en estos indicadores. Por otro lado, los países del norte y oeste de Europa, como Noruega, Suecia y Reino Unido, muestran valores más bajos (tonos claros), posiblemente asociados a mayores tasas de mortalidad por envejecimiento poblacional.



Los países con tonos más oscuros, como Italia, Alemania, Italia , Grecia y Bulgaria tienen altas expectativas de vida y bajas tasas de desempleo, reflejando economías sólidas y altos estándares de calidad de vida. En contraste, los países con tonos claros, como Irlanda, Países Bajos , España , Polonia y Bielorrusia, enfrentan mayores tasas de desempleo y menores expectativas de vida, evidenciando retos estructurales en su desarrollo socioeconómico. Regiones como Francia, Finlandia , Lituania , Letonia , ,Hungría y Rumania presentan valores intermedios, reflejando un equilibrio moderado entre estas variables.

1. ¿Cuánto vale la raíz de la media de los cuadrados de los residuales RMSR?

¿Qué nos dice este valor? ¿Qué variable tiene mayor suma de los residuos de sus correlaciones? ¿Qué significa esto?

|  |  |  |
| --- | --- | --- |
| **Root Mean Square Off-Diagonal Residuals: Overall = 0.06389325** | **CO2\_Agri** | 0.0571044 |
| **CO2\_Min** | 0.0814352 |
| **CO2\_Transp** | 0.0362135 |
| **CO2\_Domestico** | 0.0514924 |
| **Renta\_P\_C** | 0.0898894 |
| **I\_Produc\_M** | 0.0502644 |
| **Pob\_T** | 0.0349104 |
| **T\_Desemp** | 0.0927131 |
| **EV\_M** | 0.0415733 |
| **EV\_F** | 0.05457 |
| **Edad\_Media** | 0.0325795 |
| **Tasa\_Nac** | 0.0785636 |
| **T\_Mort** | 0.0413805 |
| **BolsasP\_H\_A** | 0.0808522 |
| **P\_HgconRuido** | 0.0822762 |

La raíz de la media de los cuadrados de los residuales RMSR equivale a 0.06389325. Un valor de RMSR cercano a cero sugiere un buen ajuste, mientras que valores mayores indican que el modelo no representa adecuadamente las correlaciones observadas. En este caso , el valor está en buen margen. La variable con la mayor suma de los residuos de sus correlaciones es T\_Desemp con un residuo de 0.0927131, esto sugiere que los factores incluidos en el modelo no logran capturar adecuadamente la relación de esta variable con las demás

# Análisis Cluster:

1. Realizar un análisis clúster jerárquico del conjunto de datos.
   1. Utilizar el método de Ward. **Incluir la opción Standard** en la sentencia inicial del procedimiento para trabajar con las variables estandarizadas.

**proc** **cluster** data=Tarea2.Euro method=ward STANDARD RSQUARE

PSEUDO PRINT=**15** SIMPLE outtree=salida plots=all;

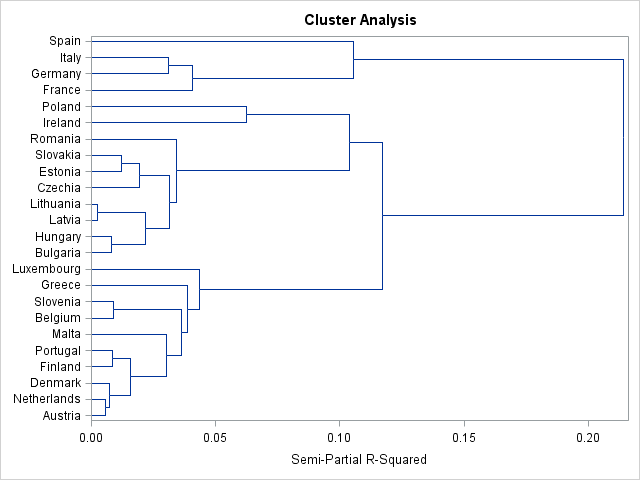
var CO2\_Agri--P\_HgconRuido;

ID Pais;

Copy Pais;

**run**;

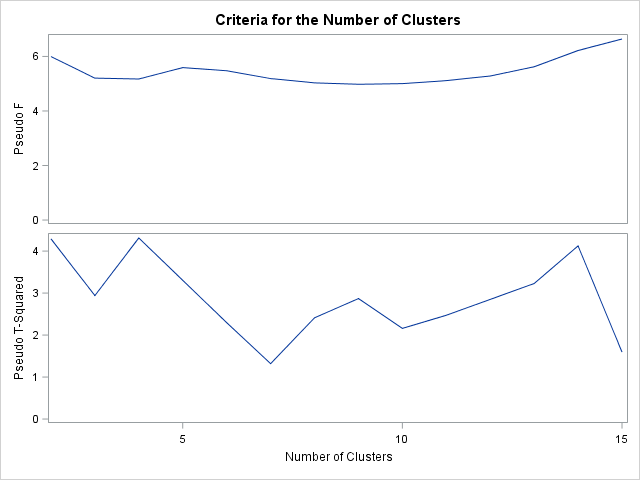
* 1. Representar el dendrograma. ¿Qué número de clústeres se intuye?



Se intuye un valor de 5 clusters al ver la imagen

* 1. Observando y la gráfica de los estadísticos pseudo F y pseudo T ¿Qué número de clústeres nos recomiendan estos criterios? (Rcuadrado como mínimo de 0.5)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Cluster History** | | | | | | | | |
| **Number of Clusters** | **Clusters Joined** | | **Freq** | **Semipartial R-Square** | **R-Square** | **Pseudo F Statistic** | **Pseudo t-Squared** | **Tie** |
| **15** | **Czechia** | **CL17** | 3 | 0.0194 | .912 | 6.6 | 1.6 |  |
| **14** | **CL20** | **CL23** | 4 | 0.0218 | .890 | 6.2 | 4.1 |  |
| **13** | **CL16** | **Malta** | 6 | 0.0301 | .860 | 5.6 | 3.2 |  |
| **12** | **Germany** | **Italy** | 2 | 0.0309 | .829 | 5.3 | . |  |
| **11** | **CL14** | **CL15** | 7 | 0.0316 | .797 | 5.1 | 2.5 |  |
| **10** | **CL11** | **Romania** | 8 | 0.0344 | .763 | 5.0 | 2.2 |  |
| **9** | **CL13** | **CL18** | 8 | 0.0365 | .726 | 5.0 | 2.9 |  |
| **8** | **CL9** | **Greece** | 9 | 0.0388 | .688 | 5.0 | 2.4 |  |
| **7** | **France** | **CL12** | 3 | 0.0408 | .647 | 5.2 | 1.3 |  |
| **6** | **CL8** | **Luxembourg** | 10 | 0.0435 | .603 | 5.5 | 2.3 |  |
| **5** | **Ireland** | **Poland** | 2 | 0.0626 | .541 | 5.6 | . |  |
| **4** | **CL10** | **CL5** | 10 | 0.1038 | .437 | 5.2 | 4.3 |  |
| **3** | **CL7** | **Spain** | 4 | 0.1055 | .331 | 5.2 | 2.9 |  |
| **2** | **CL6** | **CL4** | 20 | 0.1172 | .214 | 6.0 | 4.3 |  |
| **1** | **CL2** | **CL3** | 24 | 0.2142 | .000 | . | 6.0 |  |



Tras analizar los gráficos y la tabla, se concluye que el número óptimo de clusters es 7. Esto es porque al ver la gráfica , se observa un valor mínimo de 7 y lo mas importante es que al ver la tabla , el Pseudo  
t-Squared da 1.3

* 1. Realizar un proc tree sobre la salida del proc cluster para agrupar los individuos en el número de clústeres elegido teniendo en cuenta los dos apartados anteriores. Mostrar una tabla con los países para cada clúster.

CLUSTER=1

|  |  |
| --- | --- |
| **Obs** | **Pais** |
| **1** | Latvia |
| **2** | Lithuania |
| **3** | Bulgaria |
| **4** | Hungary |
| **5** | Estonia |
| **6** | Slovakia |
| **7** | Czechia |
| **8** | Romania |

CLUSTER=2

|  |  |
| --- | --- |
| **Obs** | **Pais** |
| **9** | Austria |
| **10** | Netherlands |
| **11** | Denmark |
| **12** | Finland |
| **13** | Portugal |
| **14** | Belgium |
| **15** | Slovenia |
| **16** | Malta |
| **17** | Greece |

CLUSTER=3

|  |  |
| --- | --- |
| **Obs** | **Pais** |
| **18** | Germany |
| **19** | Italy |
| **20** | France |

CLUSTER=4

|  |  |
| --- | --- |
| **Obs** | **Pais** |
| **21** | Luxembourg |

CLUSTER=5

|  |  |
| --- | --- |
| **Obs** | **Pais** |
| **22** | Ireland |

CLUSTER=6

|  |  |
| --- | --- |
| **Obs** | **Pais** |
| **23** | Poland |

CLUSTER=7

|  |  |
| --- | --- |
| **Obs** | **Pais** |
| **24** | Spain |

1. Realizar un análisis clúster no jerárquico utilizando el número de clústeres elegido sobre **los datos estandarizados**
   1. ¿Qué países forman cada uno de los clústeres? Mostrar una tabla con los países para cada clúster. Comparar con los obtenidos en el jerárquico.

Cluster=1

|  |  |
| --- | --- |
| **Obs** | **Pais** |
| **1** | Spain |

Cluster=2

|  |  |
| --- | --- |
| **Obs** | **Pais** |
| **2** | Austria |
| **3** | Denmark |
| **4** | Luxembourg |
| **5** | Malta |
| **6** | Netherlands |

Cluster=3

|  |  |
| --- | --- |
| **Obs** | **Pais** |
| **7** | Czechia |
| **8** | Poland |

Cluster=4

|  |  |
| --- | --- |
| **Obs** | **Pais** |
| **9** | Ireland |

Cluster=5

|  |  |
| --- | --- |
| **Obs** | **Pais** |
| **10** | France |
| **11** | Germany |
| **12** | Italy |

Cluster=6

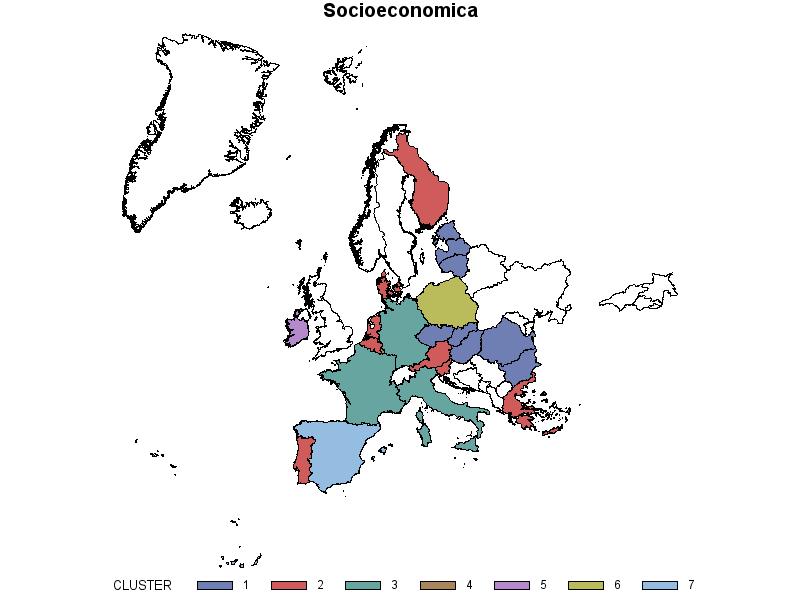
|  |  |
| --- | --- |
| **Obs** | **Pais** |
| **13** | Bulgaria |
| **14** | Estonia |
| **15** | Hungary |
| **16** | Latvia |
| **17** | Lithuania |
| **18** | Romania |
| **19** | Slovakia |

Cluster=7

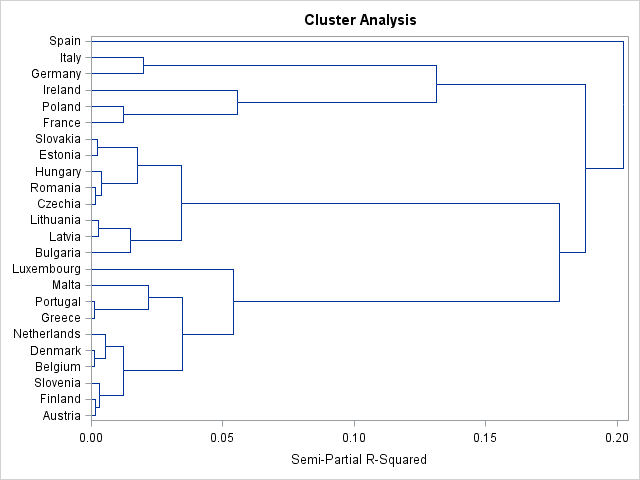
|  |  |
| --- | --- |
| **Obs** | **Pais** |
| **20** | Belgium |
| **21** | Finland |
| **22** | Greece |
| **23** | Portugal |
| **24** | Slovenia |

El método jerárquico agrupa mas países en los clusters , como se ve en los tres primeros , por lo que optaría por este método.

1. Elegir la agrupación más adecuada (jerárquica o no jerárquica) y representar el mapa de Europa con los países coloreados según el clúster al que pertenecen.



1. Utilizando como variables los factores rotados obtenidos en el apartado 4 de la primera parte:
   1. Realizar un análisis clúster jerárquico del conjunto de datos. Utilizar el método de Ward. **Incluir la opción Standard** en la sentencia inicial del procedimiento para trabajar con las variables estandarizadas. Representar el dendrograma. Observando solo la gráfica ¿Qué número de clústeres recomendarias?



Al observar , seria recomendable 5 clústeres , por la distancia que se observa

* 1. Realizar un análisis clúster no jerárquico utilizando el número de clústeres elegido antes. Puesto que las variables son factores no es necesario estandarizar.

**proc** **fastclus** data=PRO1F maxc=**5** out=solo10 OUTSTAT=ESTADISTICOS distance;

var Factor1 Factor2 Factor3 Factor4;

**run**;

**proc** **sort** data=solo10;

by cluster;

**run**;

**proc** **print** data=solo10;

var Pais;

by cluster;

**run**;

* 1. ¿Qué países forman cada uno de los clústeres? Mostrar una tabla con los países para cada clúster.

Cluster=1

|  |  |
| --- | --- |
| **Obs** | **Pais** |
| **1** | Spain |

Cluster=2

|  |  |
| --- | --- |
| **Obs** | **Pais** |
| **2** | Austria |
| **3** | Belgium |
| **4** | Denmark |
| **5** | Finland |
| **6** | Greece |
| **7** | Luxembourg |
| **8** | Malta |
| **9** | Netherlands |
| **10** | Portugal |

Cluster=3

|  |  |
| --- | --- |
| **Obs** | **Pais** |
| **11** | Bulgaria |
| **12** | Czechia |
| **13** | Estonia |
| **14** | Hungary |
| **15** | Latvia |
| **16** | Lithuania |
| **17** | Romania |
| **18** | Slovakia |
| **19** | Slovenia |

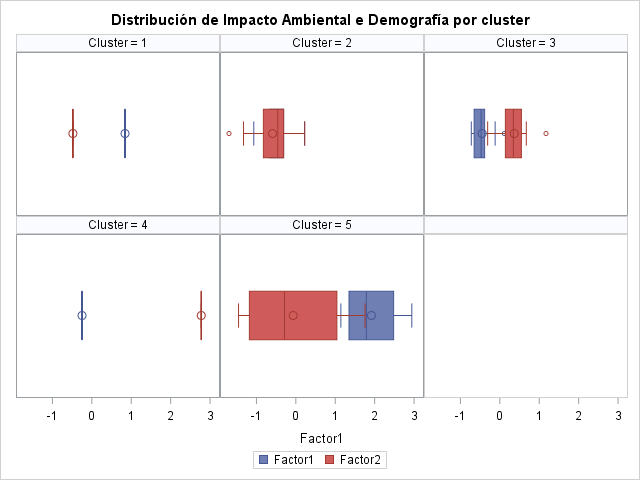
Cluster=4

|  |  |
| --- | --- |
| **Obs** | **Pais** |
| **20** | Ireland |

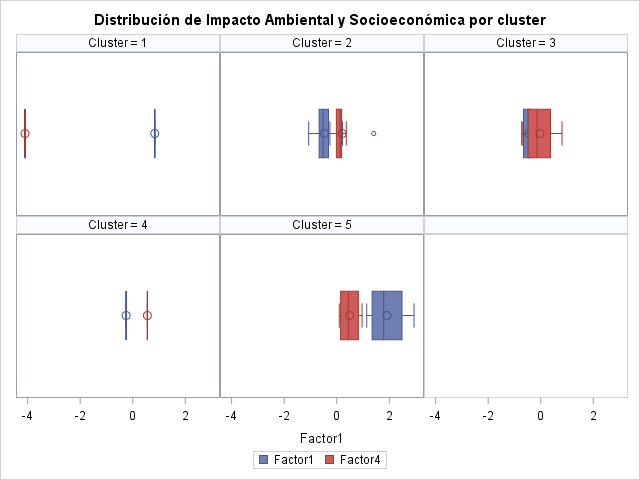
Cluster=5

|  |  |
| --- | --- |
| **Obs** | **Pais** |
| **21** | France |
| **22** | Germany |
| **23** | Italy |
| **24** | Poland |

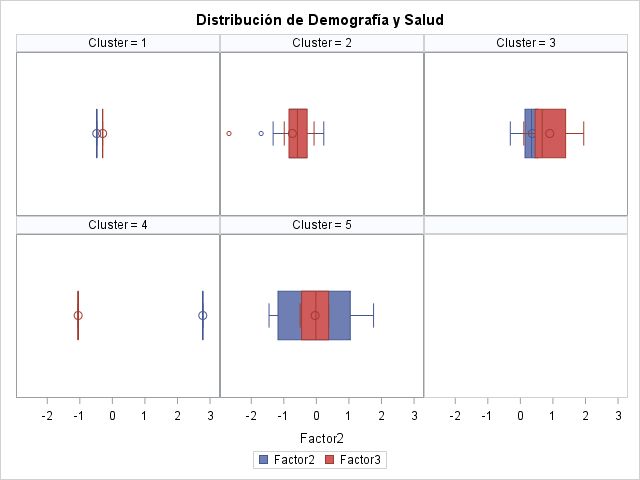
* 1. Representar los gráficos caja o los histogramas (lo que resulte más interpretable en cada caso) de los factores para cada uno de los clústeres.



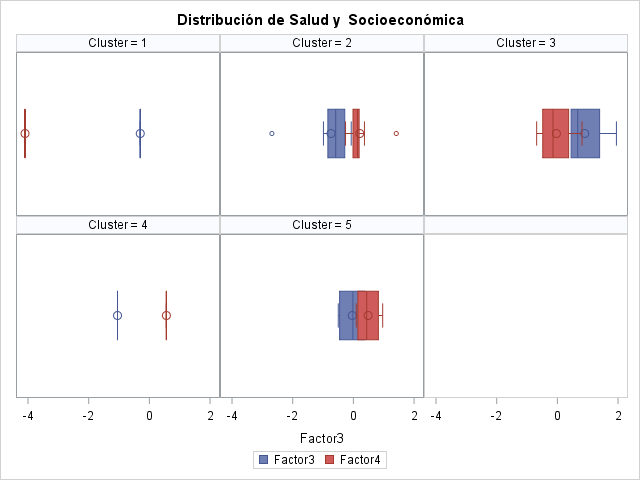
En el Cluster 1, vemos que los valores tanto de Factor 1 (Impacto Ambiental) como de Factor 2 (Demografía) se encuentran alrededor de 0, lo que indica niveles relativamente bajos en ambos aspectos para España Por otro lado, en el Cluster 2 observamos una mayor dispersión. Por ejemplo, países que tienden a tener baja demografia , así como bajo impacto ambiental. El Cluster 3, que incluye países como Bulgaria, Hungría y Rumania, tiene una distribución similar, con algunos miembros exhibiendo valores más altos en el Factor 1 (por encima de 1) y otros con valores más bajos (alrededor de -1). En el Cluster 4 vemos que Irlanda un Impacto Ambiental en el estandart normal , pero alta demografía . Por ultimo el Cluster 5 tiene a tener una distribución a alto impacto ambiental y una demografia



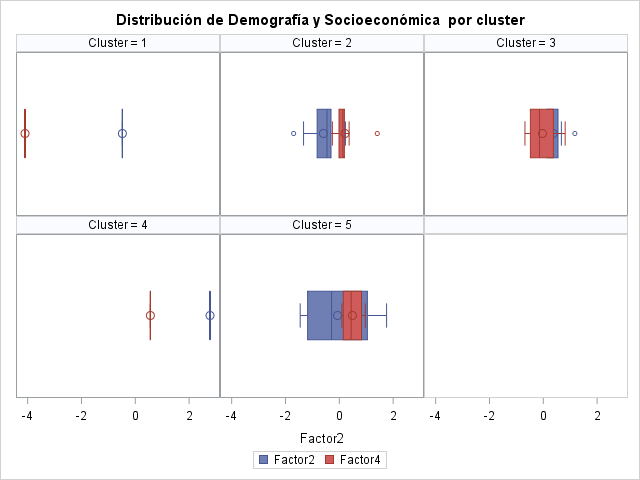
El análisis de los clústeres muestra que España (Clúster 1) tiene un perfil bajo en el nivel socioeconómico y un nivel moderado en Impacto Ambiental, mientras que el Clúster 2, compuesto por países como Austria, Bélgica y Dinamarca, destaca por valores equilibrado en Impacto Ambiental y un enfoque socioeconómico moderado, con menor dispersión. El Clúster 3, conformado por países de Europa del Este, presenta valores ligeramente negativos en ambos factores, con poca variabilidad, reflejando desafíos comunes en su desarrollo , pero en el factor demográfico , hay países con buen nivel socioeconómico moderado. Irlanda (Clúster 4) comparte un perfil balanceado en los dos factores. Finalmente, el Clúster 5, que incluye a Francia, Alemania, Italia y Polonia, es el más diverso, mostrando amplia dispersión y oscilación en ambos factores, indicando un mayor impacto ambiental , pero alto nivel socioeconómico



El clúster 1, representado por España, muestra que tanto los valores de salud como los de demografía son ligeramente negativos, aunque cercanos a 0. En el clúster 2, se observa una mayor dispersión en el factor salud, con los países de este grupo sin superar el valor de 0, lo que plantea importantes desafíos de mejora. En el ámbito demográfico, los valores también presentan dispersión significativa.El clúster 3 destaca por un buen desempeño en el factor salud, con valores positivos cercanos a 2. En cuanto a demografía, los valores están en torno al promedio, próximos a 0. Por su parte, el clúster 4, compuesto únicamente por Irlanda, presenta un buen nivel demográfico, aunque en salud muestra un desempeño negativo con un valor de -1. Finalmente, el clúster 5 refleja un desempeño equilibrado y homogéneo en términos de salud y demografía. Los valores están centrados en torno a 0, con una ligera inclinación hacia valores positivos. Esto sugiere que los países de este grupo tienen un desarrollo estable y promedio en ambas áreas, con poca variabilidad interna y sin valores atípicos significativos.



En los clústeres 1 y 4, que representan a un solo país cada uno, se observan diferencias marcadas entre los factores analizados. En el clúster 1, correspondiente a España, se identifica un bajo nivel socioeconómico junto con un nivel de salud estable. En contraste, el clúster 4, que representa a Irlanda, presenta un nivel de salud negativo pero destaca por su buen desempeño socioeconómico.Por su parte, los clústeres 2, 3 y 5, que agrupan a varios países, reflejan una mayor variabilidad. El clúster 2 muestra una distribución equilibrada alrededor del valor 0, con baja dispersión en los factores de salud y socioeconómico, y pocos casos atípicos. El clúster 3 evidencia una mayor diversidad: el factor salud tiene una tendencia positiva, mientras que el nivel socioeconómico tiende hacia valores negativos. Finalmente, el clúster 5 destaca por una ligera tendencia positiva en ambos factores; no obstante, algunos países en este grupo presentan desempeños subóptimos en el factor salud, aunque con posibilidad de mejora.



En los clústeres 1 y 4, que representan a un solo país cada uno, se observan diferencias significativas en los factores analizados. En el clúster 1, correspondiente a España, el factor demográfico presenta una ligera variación negativa, mientras que el factor socioeconómico muestra una tendencia más marcada hacia valores negativos. Por otro lado, en el clúster 4, representado por Irlanda, ambos factores tienen valores positivos, destacándose especialmente el desempeño en el aspecto socioeconómico. Los clústeres 2, 3 y 5, que agrupan a varios países, reflejan una mayor heterogeneidad en sus características. En el clúster 2, ambos factores están relativamente centrados en torno al valor 0, con baja dispersión. Sin embargo, el factor demográfico tiende hacia valores negativos. En el clúster 3, el factor socioeconómico muestra una variación más notable en comparación con el demográfico. Finalmente, el clúster 5 representa un grupo de países con valores demográficos y socioeconómicos moderadamente equilibrados y concentrados cerca del promedio, con menor variabilidad en el aspecto socioeconómico. Esto sugiere que estos países comparten características similares y homogéneas en ambos factores, lo que podría indicar un nivel de estabilidad o similitud estructural en estas dimensiones.