10-12-2024

TAREA II: TÉCNICAS PARA REDUCIR LA DIMENSIÓN Y CLUSTERING

GESTIÓN GLOBAL DEL RIESGO. SCORING

GRUPO A

XINYUAN ZHENG
xinyuanz@ucm.es

ÍNDICE

Análisis De Componentes Principales y Factorial:

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

```
proc princomp DATA=series.euro n=7 plots=all outstat=series.euro_corr ;
  var T_UNI -- T_Mort_Accidente;
run;
```

Eigenvalues of the Correlation Matrix				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	6.06155021	3.12787391	0.3368	0.3368
2	2.93367630	0.24859523	0.1630	0.4997
3	2.68508107	1.07323408	0.1492	0.6489
4	1.61184698	0.12898486	0.0895	0.7385
5	1.48286212	0.12711861	0.0824	0.8208
6	1.35574351	0.77500349	0.0753	0.8962
7	0.58074002		0.0323	0.9284

A partir de la matriz de correlación y siguiendo el criterio de autovalores mayores que 1, se determina que el número óptimo de componentes principales es 6. Estos componentes cumplen con la condición de tener un autovalor superior a 1 y, en conjunto, explicar el 89.62% de la varianza total del modelo, lo que indica que capturan la mayor parte de la información contenida en las variables originales.

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

```
proc princomp DATA=series.euro n=6 plots=all outstat=series.euro_corr_f out=series.euro_data;
  var T_UNI -- T_Mort_Accidente;
  id PAIS;
run;
```

Una vez determinado el número de componentes a retener, se establece en el código el valor igual a 6 mediante el parámetro n=6. Esto permite restringir las salidas del análisis únicamente a las seis componentes principales retenidas, las cuales explican la mayor parte de la varianza del modelo. Además, se utiliza la columna “PAIS” como identificador; su inclusión será clave para responder a las preguntas posteriores que requieren identificar los países específicos relacionados con los valores de las componentes.

2.1. ¿Cómo se calcula la Componente 2?

Eigenvectors							
		Prin1	Prin2	**Prin3**	**Prin4**	**Prin5**	**Prin6**
T_UNI	T_UNI	-.102801	0.342253	-.299924	0.223998	0.210802	-.249626
IPC	IPC	0.067527	-.066681	0.115405	0.156478	0.061646	0.690982
_0_14_YEARS	0-14 YEARS	0.281146	-.038751	-.294230	-.230938	0.346495	-.000011
_15_64_YEARS	15-64 YEARS	0.095593	-.207328	0.014712	0.356474	-.564897	-.243922
_65_YEARS	65-YEARS	-.312251	0.204701	0.225664	-.099694	0.181807	0.197918
Renta	Renta	0.206522	0.119687	-.308349	-.070373	-.260994	0.306902
Edad_Media	Edad Media	-.331398	0.193067	0.226438	0.060709	0.031575	0.114126
Tasa_Nac	Tasa_Nac	0.283981	-.162567	-.244824	-.007236	0.374614	0.118113
T_Mort	T_Mort	-.254264	-.278015	0.238646	0.177419	0.224397	-.122708
I_Produc_M	I_Produc_M	0.178412	-.355188	-.080733	0.278586	0.250344	-.203012
T_Migra_A	T_Migra_A	-.140441	0.246172	-.249117	0.434502	0.001619	0.214002
Pob_T	Pob_T	-.157255	0.376675	-.269331	0.281174	0.151825	-.180422
T_Desemp	T_Desemp	-.152843	0.130616	0.001539	-.533283	0.025778	-.253279
I_Satisf_Vida	I_Satisf_Vida	0.325468	0.265298	0.223074	0.056779	-.015236	-.037121
ISV_Cities	ISV_Cities	0.291670	0.230510	0.289740	0.106789	0.059530	-.191513
ISV_Small_Towns	ISV_Small Towns	0.324877	0.257018	0.232345	0.011046	-.017177	-.048659
ISV_Rural_areas	ISV_Rural areas	0.316106	0.303893	0.191334	-.004782	-.011783	0.023150
T_Mort_Accidente	T_Mort_Accidente	0.072850	-.095272	0.361037	0.224060	0.359357	-.025853

La Componente 2 (Prin2) se calcula como una combinación lineal de las variables originales, ponderadas por los coeficientes que aparecen en la columna Prin2. La fórmula general, omitiendo las variables intermedias por simplicidad, se expresa de la siguiente manera:

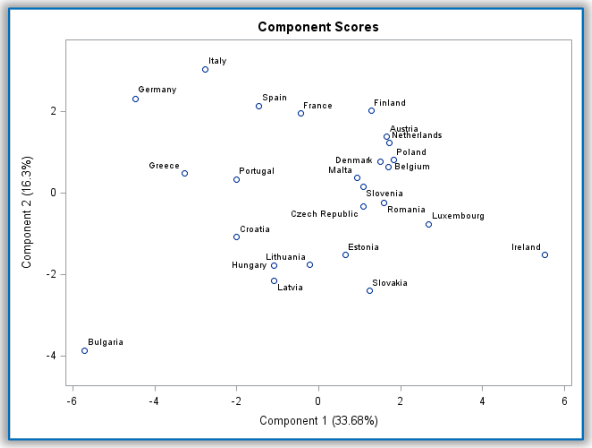
*Componente 2= 0.342*T_UNI -0.067*IPC + …… + 0.304*ISV_Rural_areas -0.095*T_Mort_Accidente*

2.2. ¿Con qué variables está más correlada?

Revisando los pesos distribuidos entre las variables para la Componente Principal 2 (Prin2), observamos que la variable con mayor correlación en valor absoluto es Pob_T. Esta variable tiene un impacto positivo de 0.377 sobre la Componente 2, siendo la principal contribuyente en su construcción.

2.3. ¿Qué país tiene mayor valor en dicha componente?

```
data series.euro_data_abs;
  set series.euro_data;
  Abs_Prin2 = abs(Prin2); /* Crear columna con el valor absoluto */
run;
proc sort data=series.euro_data_abs out=series.prin2_abs;
  by descending Abs_Prin2; /* Ordenar por el valor absoluto en orden descendente */
run;
proc print data=series.prin2_abs (obs=5);
  var PAIS Prin2 Abs_Prin2;
  title "Pais con el mayor valor en Prin2";
run;
```



Pais con el mayor valor en Prin2

Obs	PAIS	Prin2	Abs_Prin2
1	Bulgaria	-3.86721	3.86721
2	Italy	3.01736	3.01736
3	Slovakia	-2.39689	2.39689
4	Germany	2.30507	2.30507
5	Latvia	-2.15918	2.15918

Para identificar el país con el valor más alto en el Componente Principal 2 (Prin2), podemos hacerlo de dos maneras principales:

1) Visualizando un gráfico de componentes principales, donde los países se distribuyen en función de sus valores sobre los 2 componentes principales elegidos.
2) De forma más precisa, calculando los valores absolutos del Componente 2 y ordenándolos en forma descendente para determinar las observaciones con mayor magnitud (positiva o negativa).

Siguiendo este procedimiento, encontramos que:

El país con mayor valor absoluto en el Componente 2 es Bulgaria, con un valor negativo de -3.867, indicando una fuerte influencia negativa sobre esta componente.
Y si queremos analizarlo en términos de valores positivos, el país con el mayor valor será Italia, con un valor de 3.017, reflejando su mayor contribución positiva al eje del Componente 2.

3. Realizar un análisis Factorial.

```
proc factor data=series.euro corr outstat=series.factor_stats out=series.factor
        residuals msa nfact=10;
   var T_UNI -- T_Mort_Accidente;
        title "Análisis Factorial";
run;
```

A partir de este código, comenzamos el análisis factorial configurando inicialmente el número de factores con nfact=10 como punto de partida. Esto nos permite explorar cómo las variables contribuyen a los factores y evaluar los valores de MSA (Medida de Adecuación Muestral) tanto a nivel global como individual. El objetivo es identificar y eliminar de manera iterativa aquellas variables que presentan valores individuales inferiores a 0.5, ya que estas no aportan significativamente al modelo factorial. Este proceso continúa hasta que el MSA promedio del conjunto de datos alcance al menos 0.6, garantizando que las variables restantes sean adecuadas para el análisis factorial.

3.1. ¿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).

Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.54912727

T_UNI	IPC	_0_14_YEARS	_15_64_YEARS	_65_YEARS	Renta	Edad_Media	Tasa_Nac	T_Mort	I_Produc_M	T_Migra_A	Pob_T	T_Desemp	I_Satisf_Vida	ISV_Cities	ISV_Small_Towns	ISV_Rural_areas	T_Mort_Accidente
0.35277699	0.14421600	0.53326583	0.25626646	0.58691820	0.54911008	0.67792515	0.64578012	0.74338939	0.54173696	0.40047625	0.56255179	0.33033737	0.53901286	0.74898883	0.59954172	0.79737588	0.40317090

Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.54912727

Partiendo del análisis inicial, observamos un MSA promedio de 0.549, el cual es inferior al límite mínimo de 0.6 establecido para considerar el conjunto de datos adecuado para el análisis factorial. Además, identificamos variables con un MSA individual significativamente bajo, como en el caso de IPC, que presenta el menor valor de MSA_j = 0.144.

Dado que esta variable no contribuye de manera significativa al modelo factorial, procedemos a eliminarla del conjunto de datos con el objetivo de mejorar el MSA promedio y garantizar la calidad del análisis factorial.

```
proc factor data=series.euro (drop=IPC) corr outstat=series.factor_stats out=series.factor
        residuals msa nfact=10;
   var T_UNI -- T_Mort_Accidente;
        title "Analisis Factorial";
run;
```

Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.61180444

T_UNI	_0_14_YEARS	_15_64_YEARS	_65_YEARS	Renta	Edad_Media	Tasa_Nac	T_Mort	I_Produc_M	T_Migra_A	Pob_T	T_Desemp	I_Satisf_Vida	ISV_Cities	ISV_Small_Towns	ISV_Rural_areas	T_Mort_Accidente
0.38645110	0.54685184	0.26148437	0.58735836	0.54943196	0.79843395	0.68053899	0.72534766	0.51943104	0.51001596	0.56919318	0.38208023	0.69000122	0.85233646	0.70192064	0.79202644	0.41786202

Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.61180444

Después de eliminar la variable IPC del conjunto de datos debido a su baja contribución (MSA_j=0.144), el índice MSA global mejora, alcanzando un valor de 0.612. Este resultado se considera aceptable para proceder con el análisis factorial, ya que supera el umbral mínimo recomendado de 0.6 para garantizar la adecuación del conjunto de datos.

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

Eigenvalues of the Correlation Matrix: Total = 17 Average = 1				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	6.03862067	3.11269223	0.3552	0.3552
2	2.92592844	0.26331889	0.1721	0.5273
3	2.66260955	1.06100343	0.1566	0.6840
4	1.60160612	0.11967807	0.0942	0.7782
5	1.48192805	0.54887757	0.0872	0.8653
6	0.93305048	0.37350302	0.0549	0.9202
7	0.55954746	0.19169393	0.0329	0.9531
8	0.36785353	0.19714334	0.0216	0.9748
9	0.17071019	0.06367859	0.0100	0.9848
10	0.10703160	0.04081115	0.0063	0.9911
11	0.06622044	0.02849265	0.0039	0.9950
12	0.03772779	0.01723799	0.0022	0.9972
13	0.02048980	0.00629237	0.0012	0.9984
14	0.01419743	0.00565360	0.0008	0.9993
15	0.00854383	0.00478647	0.0005	0.9998
16	0.00375736	0.00358010	0.0002	1.0000
17	0.00017726		0.0000	1.0000

Tras eliminar la variable IPC, se observa en la salida del análisis factorial que el número de factores con autovalores superiores a 1 corresponde a los primeros 5 factores. Este criterio indica que estos factores explican más varianza que una variable promedio y, por lo tanto, se consideran significativos. En consecuencia, establecemos el número de factores para el análisis como 5 (nfact=5).

El código para el posterior análisis seria de esta forma:

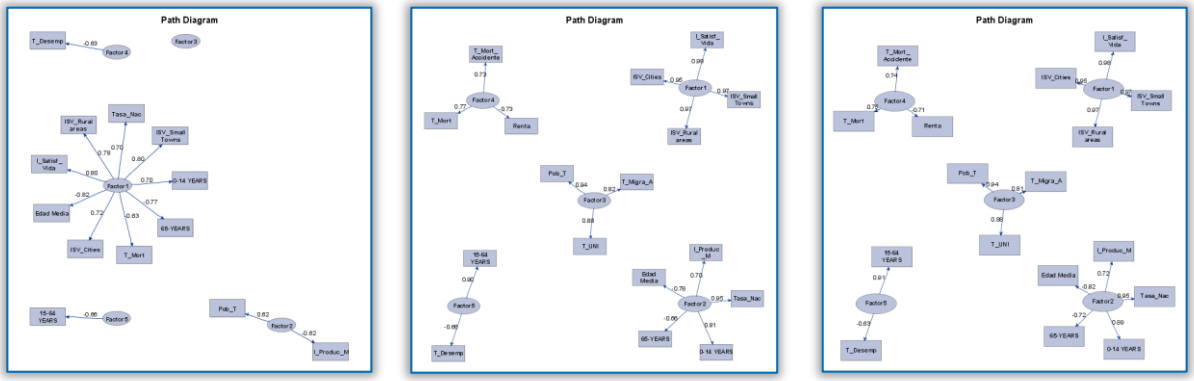
```
proc factor data=series.euro (drop=IPC) corr outstat=series.factor_stats out=series.factor
        residuals msa nfact=5 plots=all;
  var T_UNI -- T_Mort_Accidente;
        pathdiagram fuzz=0.6 scale=0.7 factorsize=0.8 novariance;
run;
```

4.1. 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.

```
proc factor data=series.euro (drop=IPC) corr outstat=series.factor_v_stats out=series.factor_v
        residuals msa nfact=5 rotate=varimax plots=all;
   var T_UNI -- T_Mort_Accidente;
        pathdiagram fuzz=0.6 scale=0.8 factorsize=1 novariance;
run;
```

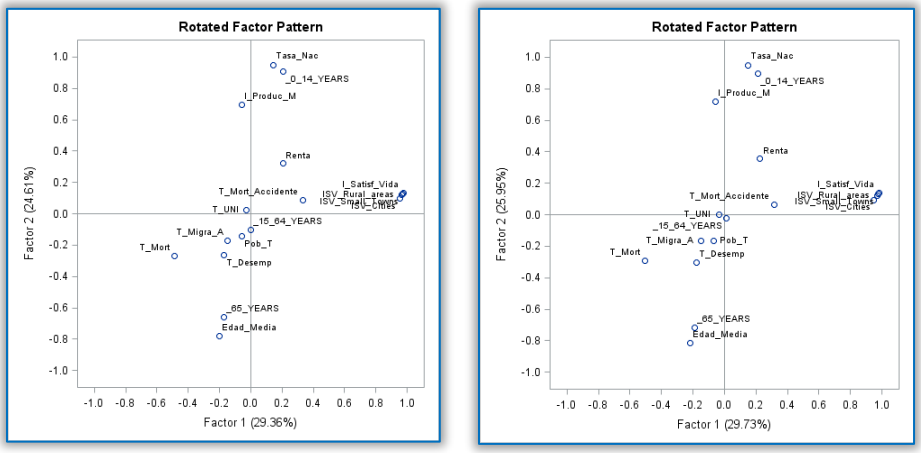
```
proc factor data=series.euro (drop=IPC) corr outstat=series.factor_q_stats out=series.factor_q
        residuals msa nfact=5 rotate=quartimax plots=all;
   var T_UNI -- T_Mort_Accidente;
        pathdiagram fuzz=0.6 scale=1 factorsize=1 novariance;
run;
```

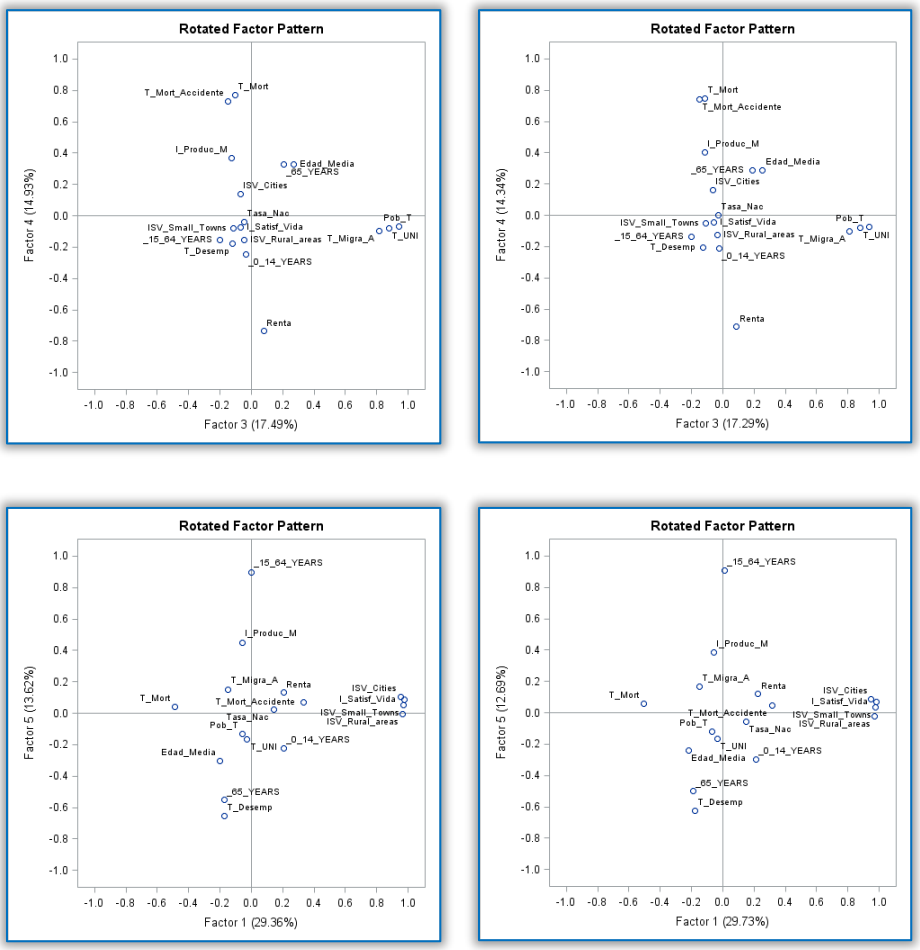
ORIGINAL **VARIMAX** **QUARTIMAX**



A partir del análisis realizado previamente y teniendo en cuenta que ya hemos establecido el número de factores en 5, podemos comparar las soluciones obtenidas con rotación Varimax y Quartimax. Ambas soluciones muestran distribuciones similares de las variables en los diferentes factores, tal como se aprecia en las gráficas del Pathdiagram, donde la configuración global de los factores es prácticamente equivalente.

VARIMAX **QUARTIMAX**





Esta similitud se extiende también a las gráficas de las variables en los planos factoriales (1-2, 3-4 y 1-5). En dichos gráficos, no se aprecia ningún patrón claro que permita diferenciar con nitidez el comportamiento de una rotación frente a la otra. Por lo tanto, ambas rotaciones ofrecen un resultado tan semejante que resulta complicado distinguir cuál es más adecuada bajo estas condiciones.

Sin embargo, la diferencia radica en la facilidad de interpretación de la estructura factorial. La rotación Varimax, al maximizar la varianza de las cargas factoriales dentro de cada factor, tiende a generar factores en los que las variables presentan cargas elevadas principalmente en un solo factor. Esto produce una solución más “limpia” y facilita comprender qué variables definen cada factor, reduciendo la complejidad al asignar significado teórico a cada dimensión extraída.

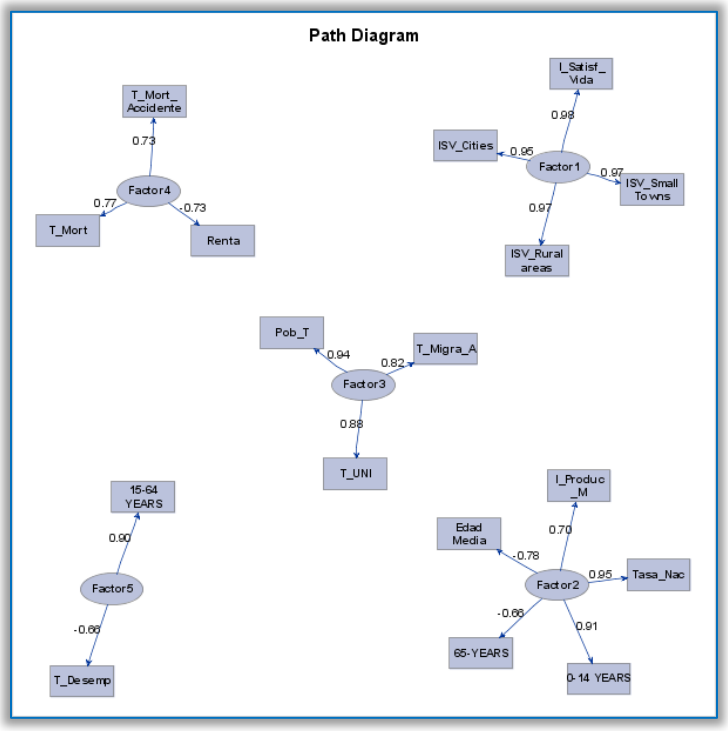
Dado que la configuración de 5 factores ya ha sido definida y las gráficas no muestran diferencias sustanciales en la distribución de variables entre Varimax y Quartimax, resulta más conveniente elegir la solución con rotación Varimax. Esta decisión se justifica por la simplicidad interpretativa, ya que Varimax proporciona una estructura factorial más clara, lo que agiliza la identificación y descripción conceptual de cada uno de los cinco factores retenidos.

CON EL MODELO FACTORIAL ROTADO

```
proc factor data=series.euro (drop=IPC) corr outstat=series.factor_v_stats out=series.factor_v
        residuals msa nfact=5 rotate=varimax plots=all;
  var T_UNI -- T_Mort_Accidente;
        pathdiagram fuzz=0.6 scale=0.8 factorsize=1 novariance;
run;
```

A partir de este código, procederemos a analizar los resultados de salidas para el modelo factorial utilizando la rotación “VARIMAX”.

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



Para el Path Diagram, se ha establecido un límite mínimo de cargas factoriales de 0.6 para las variables en relación con los factores (fuzz=0.6), mostrando únicamente aquellas variables con las mayores contribuciones a cada factor.

Factor 1:

Este factor está definido por variables como: I_Satisf_Vida (0.98), ISV_Small_Towns (0.97), ISV_Rural_Areas (0.97) y ISV_Cities (0.95). Estas variables están asociadas al índice de satisfacción de vida y a las percepciones de calidad en distintos tipos de áreas geográficas (ciudades, pueblos pequeños y zonas rurales). Por lo tanto, podemos considerar este componente como un Factor de Satisfacción de Vida.

Factor 2:

En este factor destacan variables como Tasa_Nac (0.95), 0-14_YEARS (0.91), Edad_Media (0.78), I_Produc_M (0.70) y 65_YEARS (-0.66). Estas variables están relacionadas con la estructura demográfica y la productividad económica. Las variables Tasa_Nac y 0-14_YEARS reflejan una población joven y dinámica, mientras que la relación negativa con 65_YEARS indica una menor proporción de población mayor. Este factor puede interpretarse como un Factor de Juventud y Productividad.

Factor 3:

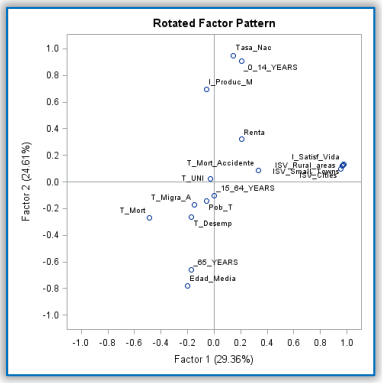
Las variables principales de este factor son Pob_T (0.94), T_UNI (0.88) y T_Migra_A (0.82). Estas variables reflejan aspectos como la población total, la tasa de urbanización y la migración neta, lo que indica un perfil vinculado al crecimiento urbano y los movimientos migratorios. Por ello, este componente puede considerarse como un Factor de Migración y Urbanización.

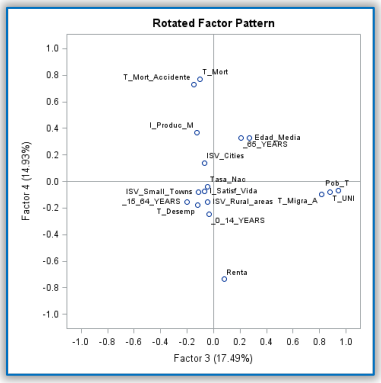
Factor 4:

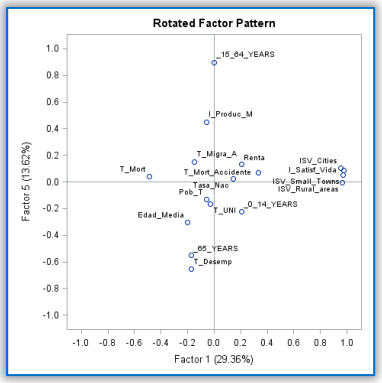
En este factor predominan variables como T_Mort (0.77), T_Mort_Accidente (0.73) y Renta (0.73). Estas variables están relacionadas con las tasas de mortalidad (general y por accidentes) y con la renta. Este factor podría estar asociado a desigualdades y riesgos vinculados al bienestar, por lo que lo denominamos Factor de Mortalidad y Renta.

Factor 5:

Finalmente, este factor está definido por variables como 15-64_YEARS (0.90) y T_Desemp (-0.66). Estas variables están estrechamente relacionadas con el empleo y la población activa. La relación negativa con el desempleo indica un entorno laboral favorable. Por lo tanto, este componente puede considerarse un Factor de Empleo y Actividad Laboral.







Observando las gráficas de las variables en los planos factoriales, los resultados obtenidos son consistentes con los descritos previamente en el Path Diagram. De manera resumida, cada factor puede categorizarse de la siguiente forma resumida:

- Factor 1: Satisfacción de vida.
- Factor 2: Población joven.
- Factor 3: Composición demográfica.
- Factor 4: Mortalidad.
- Factor 5: Empleo.

4.3. 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.

Rotated Factor Pattern						
		Factor1	Factor2	Factor3	Factor4	Factor5
T_UNI	T_UNI	-0.02566	0.02156	0.87913	-0.08162	-0.16370
_0_14_YEARS	0-14 YEARS	0.20582	0.90667	-0.03652	-0.24732	-0.22569
_15_64_YEARS	15-64 YEARS	-0.00081	-0.10529	-0.19987	-0.15277	0.89795
_65_YEARS	65-YEARS	-0.17302	-0.65924	0.20482	0.32910	-0.55344
Renta	Renta	0.20839	0.31915	0.07880	-0.73295	0.13453
Edad_Media	Edad Media	-0.19924	-0.77815	0.27089	0.32668	-0.30287
Tasa_Nac	Tasa_Nac	0.14299	0.94996	-0.04372	-0.03857	0.02228
T_Mort	T_Mort	-0.49024	-0.27031	-0.10582	0.76720	0.04258
I_Produc_M	I_Produc_M	-0.05896	0.69514	-0.12403	0.36531	0.45094
T_Migra_A	T_Migra_A	-0.14732	-0.17025	0.81505	-0.09906	0.15194
Pob_T	Pob_T	-0.05872	-0.14552	0.93953	-0.06689	-0.13373
T_Desemp	T_Desemp	-0.17234	-0.26389	-0.11743	-0.17977	-0.65585
I_Satisf_Vida	I_Satisf_Vida	0.97637	0.13139	-0.06909	-0.07225	0.08859
ISV_Cities	ISV_Cities	0.95139	0.09698	-0.06867	0.13623	0.10560
ISV_Small_Towns	ISV_Small Towns	0.97065	0.12860	-0.11613	-0.07982	0.05413
ISV_Rural_areas	ISV_Rural areas	0.96558	0.12312	-0.04456	-0.15345	-0.00540
T_Mort_Accidente	T_Mort_Accidente	0.33217	0.08567	-0.14908	0.72762	0.06824

En esta tabla se presentan las cargas de los factores sobre cada variable. Una carga con signo positivo indica una relación directa entre la variable y el factor, mientras que una carga negativa refleja una relación inversa. Por lo tanto, las variables con los valores absolutos más altos en las cargas son las que más contribuyen a definir el significado de cada factor.

La ecuación de la primera variable del modelo factorial “T_UNI” se puede expresarse como una combinación lineal de los 5 factores ponderados por sus cargas respectivas en cada factor, mostrado de la siguiente manera:

*T_UNI = -0.026*Factor1 +0.022*Factor2 +0.879*Factor3 -0.082*Factor4 -0.164*Factor5*

4.4. ¿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.

La ecuación del Factor 1, se calcula como una combinación lineal de las variables ponderadas por sus cargas factoriales en el Factor 1. Para simplificar, se omiten las variables intermedias, y la fórmula general se presenta de la siguiente manera:

*Factor 1= -0.026*T_UNI +0.206 *_0_14_YEARS + …… + 0.966*ISV_Rural_areas +0.332*T_Mort_Accidente*

4.5. 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

```
proc print data=series.factor_v;
  var PAIS Factor1--Factor5;
  where PAIS='Spain' or PAIS='Germany' or PAIS='Greece';
run;
```

Obs	PAIS	Factor1	Factor2	Factor3	Factor4	Factor5
10	Germany	-1.04821	-0.78697	3.57009	0.05642	1.07565
11	Greece	-0.80168	-1.10362	-0.95839	-0.63143	-1.96113
25	Spain	-0.06621	-0.70314	0.63880	-1.18462	-0.82193

Los valores de cada factor representan la carga del país en ese factor, indicando la intensidad con la que el país se asocia con las características representada por ese factor, siendo una medida estandarizada que refleja el nivel relativo del país en cada dimensión definida por el análisis factorial.

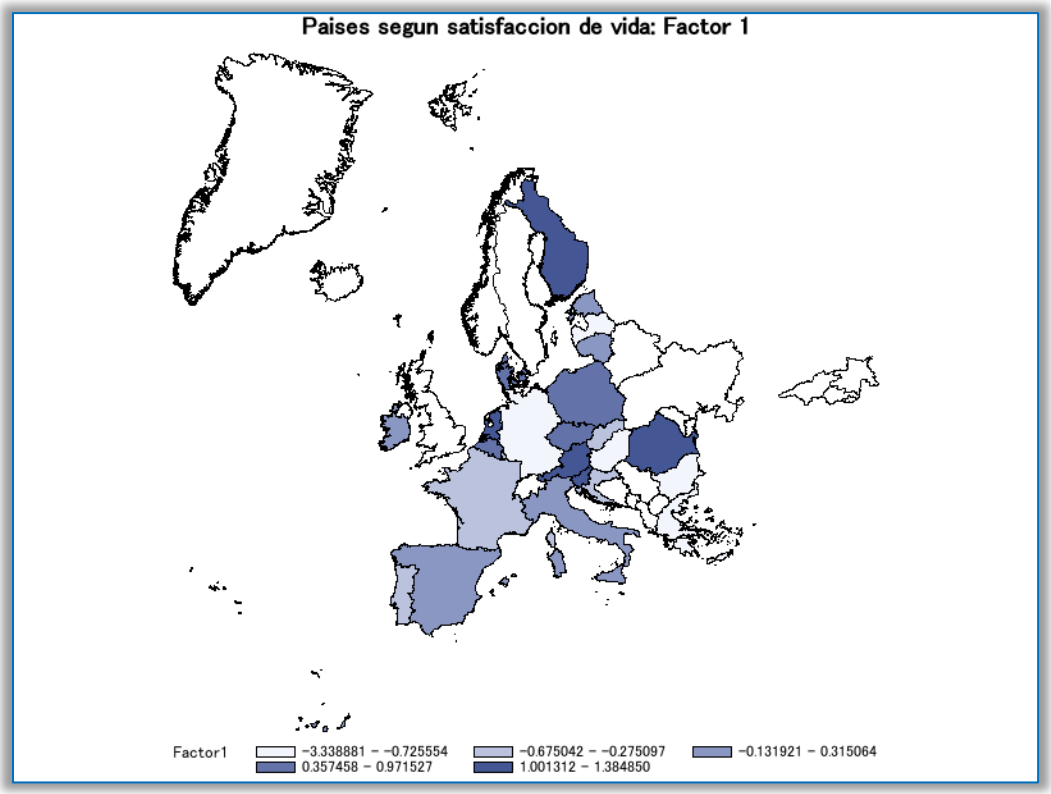
Por ejemplo, Alemania muestra un alto valor en el Factor3= 3.570, el cual está asociado con variables relacionadas con la "composición demográfica", como urbanización y migración. Esto indica que Alemania presenta características urbanas y migratorias significativamente superiores al promedio de los países europeos incluidos en el análisis. Además, Alemania tiene cargas positivas en el Factor4= 0.056 (Mortalidad) y Factor5= 1.076 (Empleo), lo que sugiere condiciones laborales favorables y niveles promedio en términos de mortalidad y renta. Sin embargo, tiene valores negativos en el Factor1= -1.048 (Satisfacción de Vida) y Factor2= -0.787 (Población Joven), lo que refleja una menor relación con las variables que definen estos factores, como satisfacción de vida y proporción de población joven. De forma resumida se puede interpretar como que Alemania sobresale por su fuerte urbanización y buenas condiciones laborales, pero enfrenta retos relacionados con menores niveles de satisfacción de vida y un envejecimiento demográfico, lo que también se refleja en su conexión con variables relacionadas con mortalidad.

En cuanto a Grecia, muestra valores negativos en todos los factores, en la cual se muestra a continuación: Factor1= -0.802 (Satisfacción de Vida), Factor2= -1.104 (Población Joven), Factor3= -0.958 (Composición Demográfica), Factor4= -0.631 (Mortalidad) y el Factor5= -1.961 (Empleo). Es decir que Grecia enfrenta retos significativos en empleo, envejecimiento poblacional y satisfacción de vida, aunque presenta aspectos más favorables en cuanto a la estabilidad demográfica y la mortalidad debido a su carga en los valores negativos en estos factores.

España presenta una combinación de valores positivos y negativos en los factores: Factor1= -0.066 (Satisfacción de Vida), Factor2= -0.703 (Población Joven), Factor3= 0.639 (Composición Demográfica), Factor4= -1.185 (Mortalidad) y Factor5= -0.822 (Empleo). Esto refleja fortalezas en su composición demográfica, con características urbanas y migratorias dinámicas, así como una menor tasa de mortalidad, presentando un perfil favorable. Sin embargo, enfrenta retos en empleo, satisfacción de vida y rejuvenecimiento poblacional, situándose por debajo del promedio europeo en estas dimensiones.

4.6. 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.

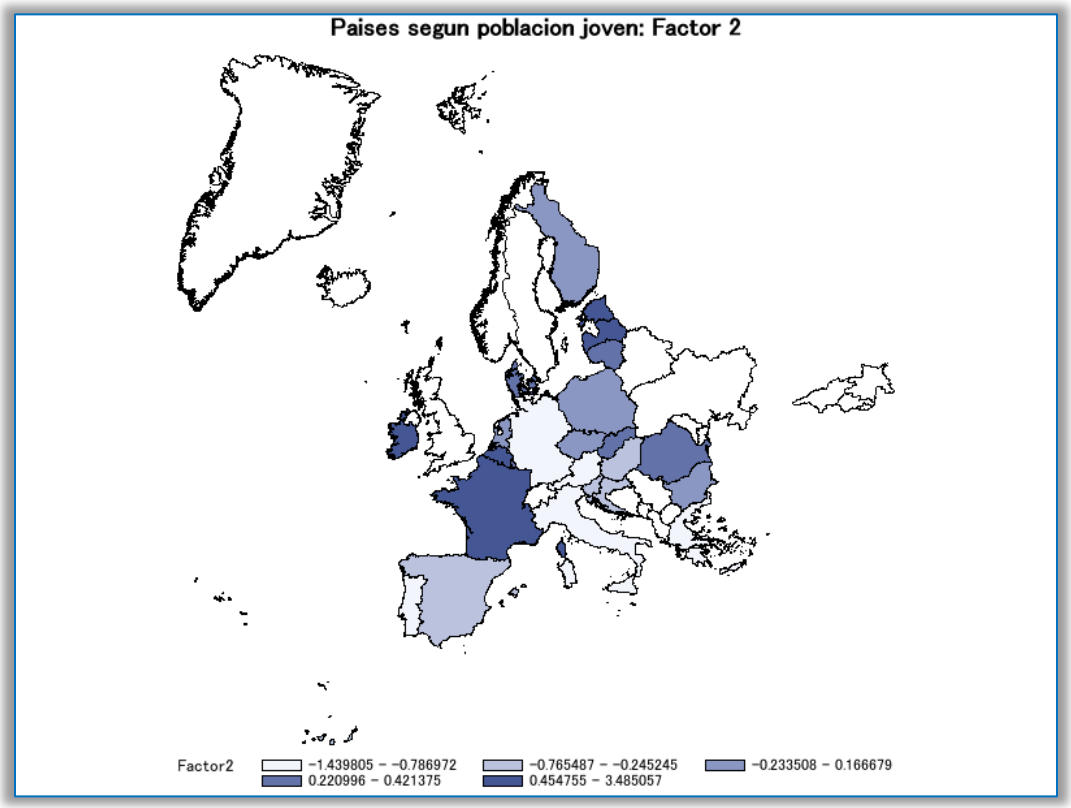
```
goptions reset=all  border;
title1 'Paises segun satisfaccion de vida: Factor 1';
proc gmap map=maps.europe
        data=series.factor_v all;
           id id;
           choro Factor1/ ;
run;
quit;
```



En esta gráfica se representan los países europeos en función del Factor 1, relacionado con el índice de satisfacción de vida y la calidad de vida percibida. La escala, dividida en cinco categorías, muestra valores entre -3.339 (más bajos) y 1.385 (más altos).

Los países con puntuaciones más altas, como Finlandia, Países Bajos, Austria y Eslovenia, se concentran en el centro y norte de Europa, probablemente debido a sistemas de bienestar robustos y altos estándares de vida. En contraste, países como Bulgaria, Alemania y Hungría, con valores negativos, se ubican mayormente en el este y sureste de Europa, lo que podría reflejar desafíos económicos, sociales o laborales que afectan su percepción de satisfacción.

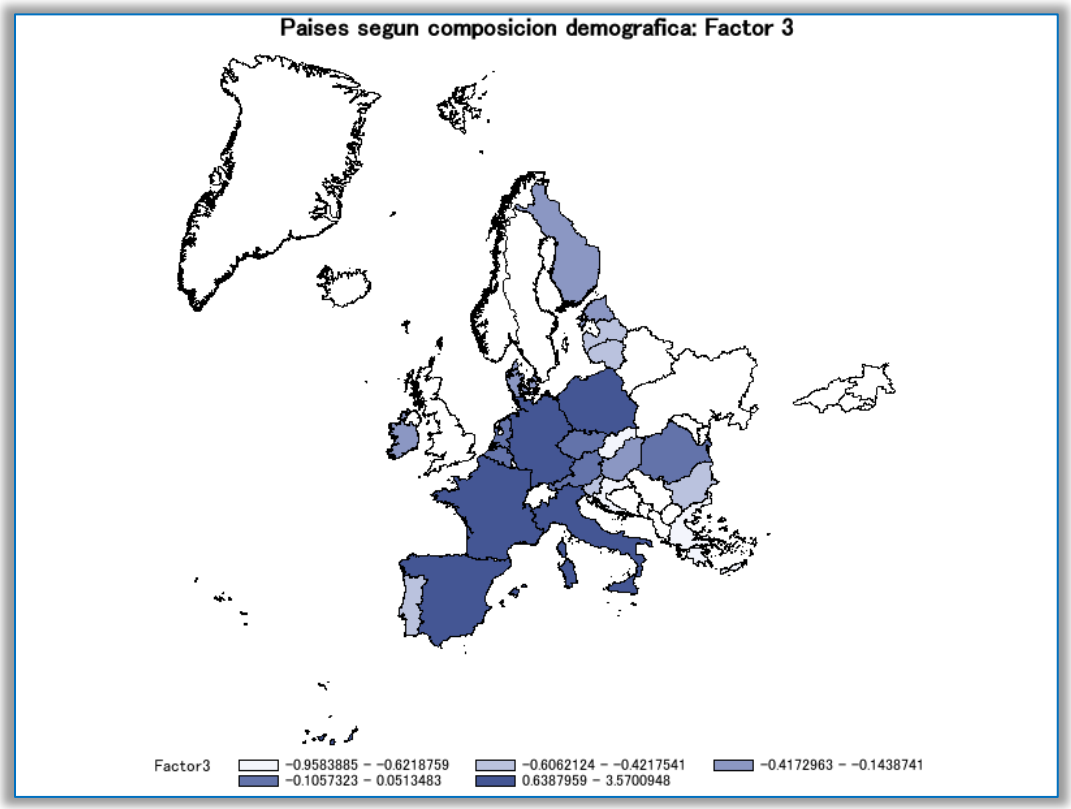
```
goptions reset=all  border;
title1 'Paises segun poblacion joven: Factor 2';
proc gmap map=maps.europe
        data=series.factor_v all;
            id id;
            choro Factor2/ ;
run;
quit;
```



En esta gráfica, se representa a los países europeos en función del Factor 2, asociado con la tasa de natalidad y la proporción de población joven (0-14 años), reflejando la juventud y el crecimiento poblacional. La escala, dividida en cinco categorías, abarca valores desde -1.439 (menor proporción de población joven) hasta 3.485 (mayor proporción).

Países como Irlanda, Francia, Estonia y Bélgica destacan con los valores más altos, indicando un perfil demográfico juvenil. En contraste, países como Italia, Grecia y Portugal presentan valores negativos, asociados a un envejecimiento poblacional y bajas tasas de natalidad.

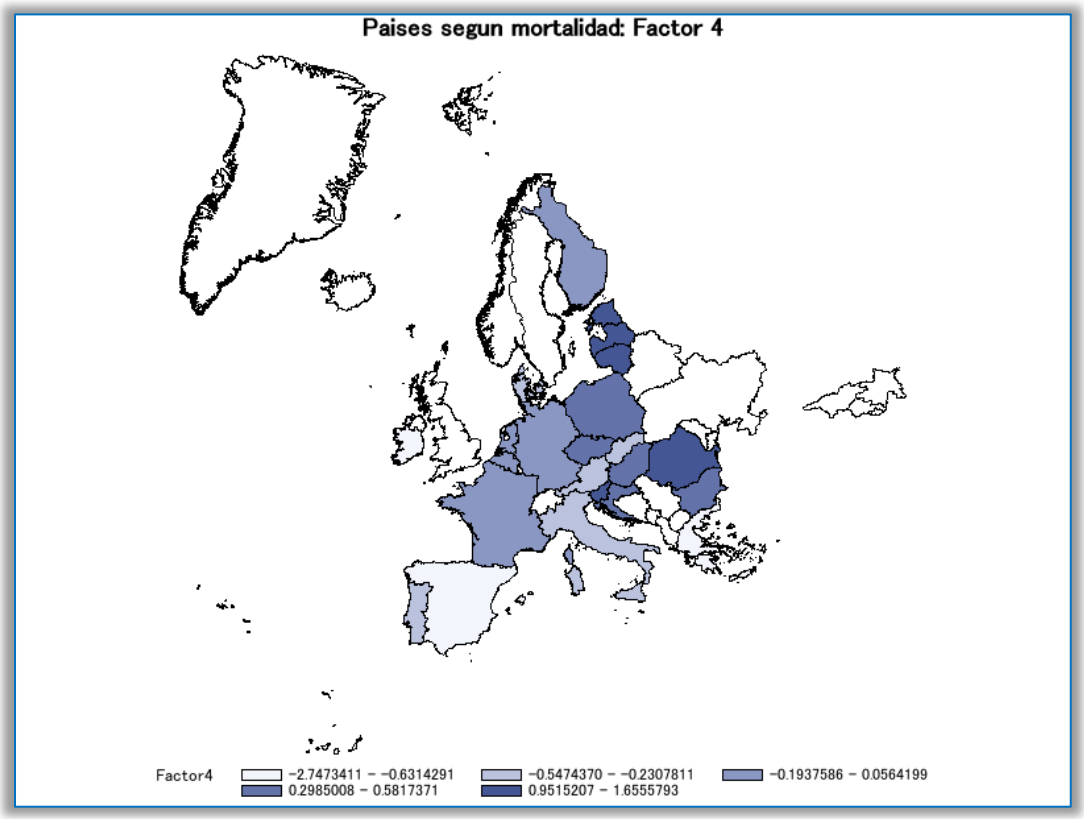
```
goptions reset=all  border;
title1 'Paises segun composicion demografica: Factor 3';
proc gmap map=maps.europe
        data=series.factor_v all;
            id id;
            choro Factor3/ ;
run;
quit;
```



La gráfica superior, representa a los países europeos en función del Factor 3, asociado con variables relacionadas con la migración, urbanización y concentración poblacional. Este factor refleja dinámicas demográficas, como la densidad poblacional y los patrones migratorios hacia áreas urbanas. La escala, dividida en cinco categorías, abarca valores desde -0.958 (los más bajos, indicando menor urbanización o migración) hasta 3.570 (los más altos, asociados a mayor urbanización y migración).

Países como Alemania, Francia, España e Italia, destacan con valores altos, reflejando patrones intensos de urbanización y migración interna o externa. En contraste, países como Grecia, Croacia y Malta presentan valores bajos, lo que puede estar relacionado con estructuras demográficas más estables o una menor dinámica migratoria.

```
goptions reset=all  border;
title1 'Paises segun mortalidad: Factor 4';
proc gmap map=maps.europe
        data=series.factor_v all;
           id id;
           choro Factor4/ ;
run;
quit;
```



Esta gráfica representa a los países europeos en función del Factor 4, asociado con variables relacionadas con la mortalidad general, mortalidad por accidentes y renta. Este factor refleja disparidades en términos de salud y bienestar, así como riesgos relacionados con la mortalidad. La escala está dividida en cinco categorías, con valores que oscilan entre -2.747 (los más bajos, asociados a menores tasas de mortalidad) y 1.656 (los más altos, vinculados a mayores tasas de mortalidad).

Países como Estonia, Letonia y Lituania presentan los valores más altos, lo que sugiere desafíos importantes en términos de salud pública y mortalidad. En contraste, países como Luxemburgo, España y Irlanda se sitúan en los valores más bajos, reflejando mejores condiciones de salud y menor exposición a riesgos de mortalidad.

5. ¿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?

```
proc factor data=series.euro (drop=IPC) corr outstat=series.factor_v_stats out=series.factor_v
            residuals msa nfact=5 rotate=varimax plots=all;
   var T_UNI -- T_Mort_Accidente;
            pathdiagram fuzz=0.6 scale=0.8 factorsize=1 novariance;
run;
```

Root Mean Square Off-Diagonal Residuals: Overall = 0.05483867

T_UNI	_0_14_YEARS	_15_64_YEARS	_65_YEARS	Renta	Edad_Media	Tasa_Nac	T_Mort	I_Produc_M	T_Migra_A	Pob_T	T_Desemp	I_Satisf_Vida	ISV_Cities	ISV_Small_Towns	ISV_Rural_areas	T_Mort_Accidente
0.08401909	0.01009838	0.07274255	0.05791225	0.08132579	0.05062340	0.03245778	0.01916459	0.04973568	0.07496751	0.05143365	0.08878856	0.01322368	0.02809118	0.01678210	0.01544770	0.06791925

El valor de RMSR es de 0.0544, y se encuentra entre el rango de 0.05 a 0.10; lo que indica que el modelo factorial tiene un buen ajuste general. El RMSR mide el error promedio en las correlaciones entre las variables después de ajustar los factores retenidos, y su bajo nivel sugiere que los factores seleccionados explican de manera adecuada las relaciones entre la mayoría de las variables, con errores residuales pequeños.

La variable con mayor suma de residuos es T_Desemp (0.089); al ser superior al RMSR promedio, indica que no está siendo bien explicada por los cinco factores retenidos en el modelo. Esto puede deberse a que T_Desemp tiene relaciones particulares con las demás variables que no son capturadas adecuadamente, o a que su varianza podría estar influida por factores externos no considerados en el análisis factorial.

Análisis Cluster

1. Realizar un análisis clúster jerárquico del conjunto de datos.

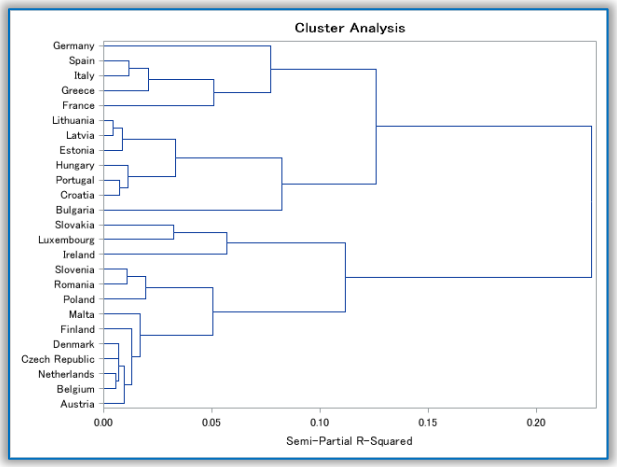
El siguiente análisis clúster jerárquico se ha llevado a cabo utilizando el conjunto de datos **"eurostat_A"**, con el objetivo de identificar patrones de similitud entre los países europeos presentes en el estudio. Este método permite clasificar las observaciones en grupos homogéneos basados en las características seleccionadas. El procedimiento jerárquico es útil para explorar la estructura de los datos y determinar cómo se agrupan progresivamente las observaciones, proporcionando una visión inicial de la segmentación natural dentro del conjunto de datos.

1.1. Utilizar el método de Ward. Incluir la opción Standard en la sentencia inicial del procedimiento para trabajar con las variables estandarizadas.

```
proc cluster data=series.euro method=ward STANDARD RSQUARE PSEUDO PRINT= 15 SIMPLE
outtree=series.euro_ward plots=all;
var T_UNI -- T_Mort_Accidente;
id PAIS;
copy PAIS id;
run;
```

Se ha empleado el **método de Ward**, conocido por minimizar la varianza dentro de los clústeres, y se ha estandarizado previamente las variables con la opción **STANDARD** para garantizar que todas tengan la misma escala. Además, se han evaluado las primeras **15 combinaciones de clústeres**, lo que ha permitido identificar patrones iniciales y establecer un punto de partida para determinar el número óptimo de agrupaciones.

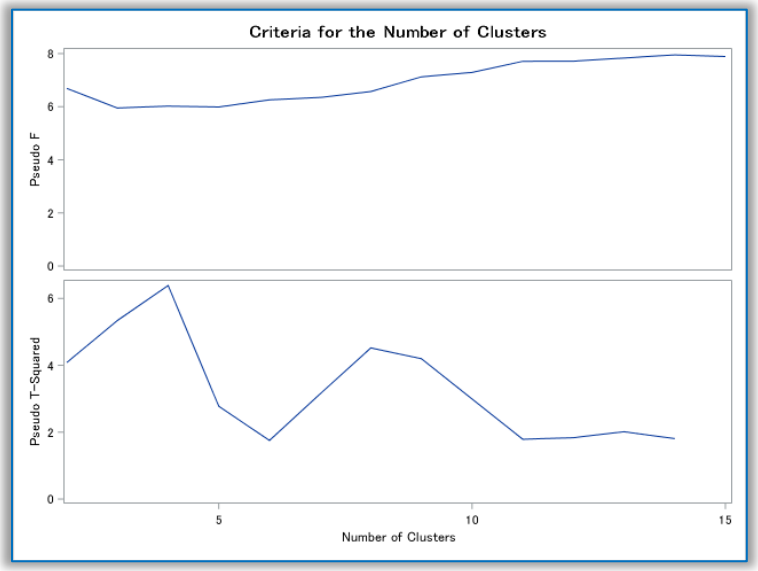
1.2. Representar el dendrograma. ¿Qué número de clústeres se intuye?



Para determinar el número óptimo de clústeres, se consideró la distancia más grande entre los grandes clústeres, indicada por las líneas verticales del dendrograma. Basándonos en esta distancia, se podría establecer el punto de corte en un valor cercano a $R^2 = 0.12$. Si se realiza el corte un poco antes, el número óptimo de clústeres para comenzar sería k = 4.

1.3. Observando y la gráfica de los estadísticos pseudo F y pseudo T ¿Qué número de clústeres nos recomiendan estos criterios? (R cuadrado 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	**Italy**	**Spain**	2	0.0116	.917	7.9	.	
14	**CL18**	**Finland**	6	0.0132	.904	8.0	1.8	
13	**CL14**	**Malta**	7	0.0170	.887	7.8	2.0	
12	**Poland**	**CL17**	3	0.0196	.867	7.7	1.8	
11	**Greece**	**CL15**	3	0.0208	.846	7.7	1.8	
10	**Luxembourg**	**Slovakia**	2	0.0324	.814	7.3	.	
9	**CL16**	**CL19**	6	0.0331	.781	7.1	4.2	
8	**CL13**	**CL12**	10	0.0507	.730	6.6	4.5	
7	**France**	**CL11**	4	0.0511	.679	6.3	3.1	
6	**Ireland**	**CL10**	3	0.0568	.622	6.3	1.8	
5	**CL7**	**Germany**	5	0.0773	.545	6.0	2.8	
4	**Bulgaria**	**CL9**	7	0.0825	.463	6.0	6.4	
3	**CL8**	**CL6**	13	0.1115	.351	5.9	5.3	
2	**CL4**	**CL5**	12	0.1257	.225	6.7	4.1	
1	**CL3**	**CL2**	25	0.2253	.000	.	6.7	



Usando los criterios de “Pseudo F” y “Pseudo T”, para contrastar el número de cluster óptimo para el análisis, se pretende buscar el punto de corte en la cual el “Pseudo F” se encuentre en su punto máximo mientras que el “Pseudo T” se encuentre en su punto mínimo. En este caso, el punto de corte podría ser para el k=6, en la cual el “Pseudo T” se encuentra en su valor mínimo y el “Pseudo F” presenta un valor razonable, aunque no máximo, indicando una partición equilibrada. Por otra parte, si revisamos el valor del R^2 , observamos que para k=6 muestra un $R^2 = 0.622$, lo que cumple con la condición establecida de $R^2 > 0.5$ establecido. Esto respalda la elección de k=6 como un número adecuado de clústeres para el análisis.

Por lo tanto, tras el análisis realizado utilizando el dendrograma y los criterios de Pseudo F y Pseudo T, se establece que el número óptimo de clústeres para el análisis es de (k=6).

1.4. 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.

```
proc tree data=series.euro_ward  out=series.euro_ward_pais n=6;
copy T_UNI -- T_Mort_Accidente PAIS id;
run;
proc sort data=series.euro_ward_pais;
   by cluster;
run;
proc print data=series.euro_ward_pais;
   var PAIS cluster;
   by cluster;
run;
```

CLUSTER=1

Obs	PAIS	CLUSTER
1	Latvia	1
2	Lithuania	1
3	Croatia	1
4	Portugal	1
5	Estonia	1
6	Hungary	1

CLUSTER=2

Obs	PAIS	CLUSTER
7	Belgium	2
8	Netherlands	2
9	Czech Republic	2
10	Denmark	2
11	Austria	2
12	Romania	2
13	Slovenia	2
14	Finland	2
15	Malta	2
16	Poland	2

CLUSTER=3

Obs	PAIS	CLUSTER
17	Italy	3
18	Spain	3
19	Greece	3
20	France	3

CLUSTER=4

Obs	PAIS	CLUSTER
21	Luxembourg	4
22	Slovakia	4
23	Ireland	4

CLUSTER=5

Obs	PAIS	CLUSTER
24	Germany	5

CLUSTER=6

Obs	PAIS	CLUSTER
25	Bulgaria	6

En estas tablas se muestra la distribución de los países según los clústeres obtenidos mediante el análisis jerárquico. Se observa que el Clúster 2 concentra la mayor cantidad de países, con un total de 10 países, lo que indica que este grupo tiene características más comunes entre un mayor número de observaciones. Por otro lado, los Clústeres 5 y 6 contienen únicamente un país cada uno (Alemania y Bulgaria, respectivamente), lo que sugiere que estos presentan características únicas o menos compartidas en comparación con los demás países del análisis.

2. Realizar un análisis clúster no jerárquico utilizando el número de clústeres elegido sobre los datos estandarizados

```
proc stdize data=series.euro method=std out=series.euro_std;
  var T_UNI -- T_Mort_Accidente;
run;

proc fastclus data=series.euro_std maxc=6 out=series.euro_std_pais outstat=series.std_pais_stat distance;
  var T_UNI -- T_Mort_Accidente;
  id PAIS;
run;
```

A partir de estas sintaxis, se realizará un análisis clúster no jerárquico mediante PROC FASTCLUS, utilizando el método de K-means con 6 clústeres, basados en los resultados del análisis jerárquico previo. Los datos fueron previamente estandarizados con PROC STDIZE para asegurar comparabilidad entre variables.

Este análisis asignó a cada país un clúster según su proximidad a los centroides calculados, permitiendo evaluar la consistencia con las agrupaciones jerárquicas y validar la robustez de los patrones identificados.

2.1. ¿Qué países forman cada uno de los clústeres? Mostrar una tabla con los países para cada clúster. Comparar con los obtenidos en el jerárquico.

```
proc sort data=series.euro_std_pais;
  by cluster;
run;

proc print data=series.euro_std_pais;
  var PAIS cluster;
  by cluster;
  title "Paises Agrupados por Cluster (Analisis No Jerarquico)";
run;
```

Cluster=1

Obs	PAIS	CLUSTER
1	Germany	1

Cluster=2

Obs	PAIS	CLUSTER
2	France	2
3	Italy	2
4	Spain	2

Cluster=3

Obs	PAIS	CLUSTER
5	Bulgaria	3
6	Greece	3

Cluster=4

Obs	PAIS	CLUSTER
7	Luxembourg	4
8	Malta	4

Cluster=5

Obs	PAIS	CLUSTER
9	Ireland	5

Cluster=6

Obs	PAIS	CLUSTER
10	Austria	6
11	Belgium	6
12	Croatia	6
13	Czech Republic	6
14	Denmark	6

15	Estonia	6
16	Finland	6
17	Hungary	6
18	Latvia	6
19	Lithuania	6
20	Netherlands	6
21	Poland	6
22	Portugal	6
23	Romania	6
24	Slovakia	6
25	Slovenia	6

El análisis clúster no jerárquico basado en 6 clústeres muestra una distribución notablemente desigual de países entre los grupos. El Clúster 6 concentra la mayoría de los países, con un total de 16 países, lo que representa un aumento significativo en comparación con los resultados del análisis jerárquico, donde el Clúster 2 contenía únicamente 10 países. Esto refleja que el método no jerárquico tiende a agrupar más observaciones en un único clúster, generando una mayor concentración y homogeneidad dentro de ciertos grupos.

En contraste, el método jerárquico distribuye mejor las observaciones entre los diferentes clústeres, lo que proporciona una diferenciación más clara entre subgrupos. Por ejemplo, en el análisis jerárquico, los Clústeres 5 y 6 contenían un solo país cada uno (Alemania y Bulgaria, respectivamente), mientras que en el análisis no jerárquico, Irlanda se asignó como único miembro del Clúster 5 y Alemania permaneció como único miembro del Clúster 1. Este contraste resalta cómo el método jerárquico puede identificar características únicas de países específicos, mientras que el no jerárquico agrupa de manera más generalizada países con similitudes globales.

En resumen, aunque ambos métodos capturan patrones relevantes, el análisis jerárquico parece ofrecer una distribución más equilibrada, permitiendo un análisis más detallado y segmentado de las diferencias entre países, mientras que el método no jerárquico proporciona una visión más general y menos diferenciada.

3. 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.

```
proc print data=series.std_pais_stat;
run;
```

Empleando este código, podemos obtener el valor de R^2 del análisis de clúster no jerárquico a través de las estadísticas generadas. Por otro lado, el valor de R^2 del análisis de clúster jerárquico se puede consultar directamente en los resultados de salida generados por el código correspondiente al punto 2.1. Esto permite una comparación entre ambos enfoques, evaluando la calidad de las agrupaciones en cada método.

R² del análisis de clúster jerárquico

Cluster History

Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Pseudo F Statistic	Pseudo t-Squared	Tie
15	Italy	Spain	2	0.0116	.917	7.9	.	
14	CL18	Finland	6	0.0132	.904	8.0	1.8	
13	CL14	Malta	7	0.0170	.887	7.8	2.0	
12	Poland	CL17	3	0.0196	.867	7.7	1.8	
11	Greece	CL15	3	0.0208	.846	7.7	1.8	
10	Luxembourg	Slovakia	2	0.0324	.814	7.3	.	
9	CL16	CL19	6	0.0331	.781	7.1	4.2	
8	CL13	CL12	10	0.0507	.730	6.6	4.5	
7	France	CL11	4	0.0511	.679	6.3	3.1	
6	Ireland	CL10	3	0.0568	.622	6.3	1.8	
5	CL7	Germany	5	0.0773	.545	6.0	2.8	
4	Bulgaria	CL9	7	0.0825	.463	6.0	6.4	
3	CL8	CL6	13	0.1115	.351	5.9	5.3	
2	CL4	CL5	12	0.1257	.225	6.7	4.1	
1	CL3	CL2	25	0.2253	.000	.	6.7	

R² del análisis de clúster no jerárquico

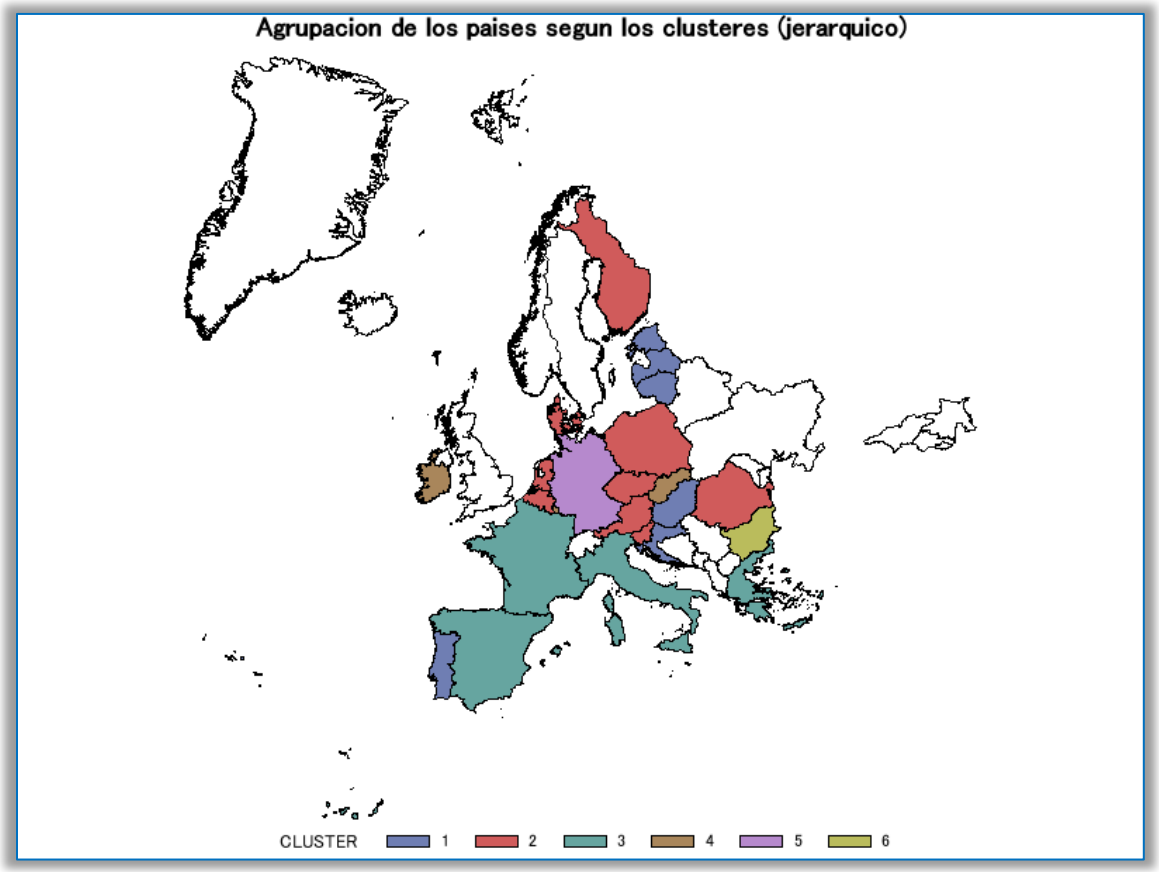
' proc print data=series.std_pais_stat

Obs	_TYPE_	CLUSTER	OVER_ALL	T_UNI	IPC	_0_14_YEARS	_15_64_YEARS	_65_YEARS	Renta	Edad_Media	Tasa_Nac	T_Mort	I_Pr
1	INITIAL	1	.	1.9922	0.2123	-1.2824	-0.1324	1.1850	0.5303	1.8435	-0.8356	0.2962	
2	INITIAL	2	.	3.0511	-0.1950	1.6676	-1.7588	0.1253	0.1365	-0.5173	1.5747	-0.7555	
3	INITIAL	3	.	-0.4417	-1.2505	-0.8371	-0.1884	0.8625	-1.1104	0.8170	-0.6749	2.0492	
4	INITIAL	4	.	-0.7295	0.3234	0.5544	1.8867	-1.9940	3.6804	-1.4411	0.5303	-1.5882	
5	INITIAL	5	.	-0.3801	-0.7752	3.5600	-0.5810	-2.4547	1.7116	-2.7755	3.1816	-1.8511	
6	INITIAL	6	.	-0.1406	-1.6641	-0.0022	0.5967	-0.5197	-0.8698	-0.3120	-0.0321	1.1289	
7	LEAST	.	2.0000	.	.	.	.	.	.	.	.	.	
8	CRITERION	.	0.6579	.	.	.	.	.	.	.	.	.	
9	MEAN	.	.	0.0000	-0.0000	-0.0000	0.0000	0.0000	-0.0000	-0.0000	0.0000	-0.0000	
10	STD	.	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	
11	WITHIN_STD	.	0.7547	0.5938	0.9575	0.6679	0.9694	0.7803	0.7848	0.6777	0.7594	0.8185	
12	RSQ	.	0.5491	0.7209	0.2742	0.6469	0.2561	0.5180	0.5124	0.6364	0.5435	0.4696	
13	RSQ_RATIO	.	1.2179	2.5829	0.3777	1.8319	0.3442	1.0746	1.0507	1.7502	1.1904	0.8854	
14	PSEUDO_F	.	4.6282	.	.	.	.	.	.	.	.	.	
15	ERSQ	.	.	.	.	.	.	.	.	.	.	.	
16	CCC	.	.	.	.	.	.	.	.	.	.	.	

Al comparar los valores de R² obtenidos en los análisis de clúster jerárquico y no jerárquico, se observa una diferencia notable en la capacidad explicativa de cada método. El análisis jerárquico presenta un R² de 0.622, lo que indica que este método explica el 62.2% de la variabilidad total en los datos mediante la agrupación realizada. Por otro lado, el análisis no jerárquico muestra un R² de 0.5491, explicando el 54.91% de la variabilidad.

Este contraste sugiere que el análisis jerárquico ofrece una mejor capacidad para capturar la estructura subyacente de los datos, generando clústeres que explican más variabilidad. Por el contrario, el análisis no jerárquico tiende a simplificar las agrupaciones, lo que puede ser útil para interpretaciones más generales, pero a costa de perder precisión en la representación de las características de los datos. Por tanto, en términos de calidad de las agrupaciones, el análisis jerárquico resulta más robusto en este caso.

```
goptions reset=all  border;
title1 'Agrupacion de los paises segun los clusteres (jerarquico)';
proc gmap map=maps.europe
        data=series.euro_ward_pais all;
           id id;
           choro CLUSTER/discrete;
run;
```



En esta gráfica se muestra la distribución de los países europeos del dataset, coloreados según el clúster al que pertenecen tras el análisis jerárquico. Se destacan clústeres que contienen un único país, como el clúster 5 (Alemania) y el clúster 6 (Bulgaria). Los países mediterráneos, como España, Francia, Italia y Grecia, conforman el clúster 3, mientras que los países del norte y centro de Europa están mayormente asociados al clúster 2. Por otro lado, países como Luxemburgo, Irlanda y Eslovaquia forman el clúster 4, reflejando características que los diferencian del resto.

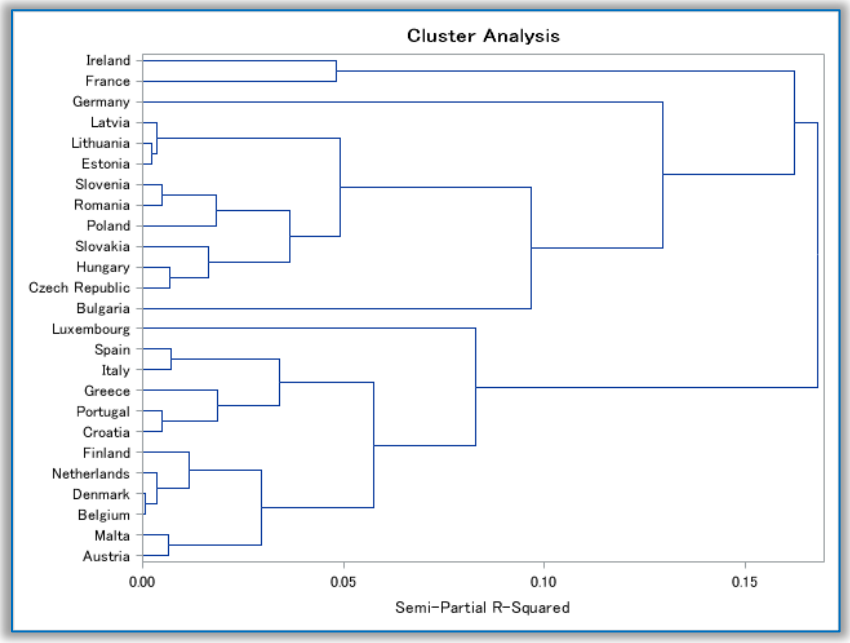
4. Utilizando como variables los factores rotados obtenidos en el apartado 4 de la primera parte

```
proc factor data=series.euro (drop=IPC) corr outstat=series.factor_v_stats out=series.factor_v
        residuals msa nfact=5 rotate=varimax plots=all;
  var T_UNI -- T_Mort_Accidente;
        pathdiagram fuzz=0.6 scale=0.8 factorsize=1 novariance;
run;
```

A partir del análisis factorial realizado en el apartado 4, utilizando rotación VARIMAX sobre las variables seleccionadas, se generó un nuevo conjunto de datos que contiene los valores de los países en los cinco factores identificados. Este dataset de salida, será la base para los análisis subsecuentes, permitiendo agrupar a los países según sus características principales en estos factores y explorar patrones de similitud mediante técnicas de clúster.

4.1. 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 recomendarías?

```
proc cluster data=series.factor_v  method=ward STANDARD RSQUARE PSEUDO PRINT= 15 SIMPLE
outtree=series.factor_v_ward plots=all;
var Factor1 -- Factor5;
id PAIS;
copy PAIS;
run;
```



Observando únicamente el dendrograma, sin considerar los criterios adicionales como “Pseudo F”, “Pseudo T” o el R^2 en detalle, el número óptimo de clústeres puede determinarse identificando las líneas verticales más largas que representan la mayor distancia entre grupos antes de que se unan en conglomerados mayores. En este caso, se observa una distancia significativa alrededor de un R^2=0.07, donde la separación entre los clústeres sigue siendo clara y los grupos mantienen homogeneidad interna.

Por lo tanto, el punto de corte óptimo basado exclusivamente en el análisis del dendrograma se establece en k=6, equilibrando la separación entre clústeres y la cohesión dentro de cada grupo.

4.2. 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=series.factor_v maxc=6 out=series.factor_v_no outstat=series.factor_v_no_stat distance;
  var Factor1 -- Factor5;
  id PAIS;
run;
```

En este apartado se realizará un análisis clúster no jerárquico utilizando las puntuaciones factoriales rotadas y el método “fastclus”, con el número de clústeres previamente determinado (k=6). Esto permitirá agrupar los países según sus similitudes en los factores, optimizando la distancia intra-clúster y maximizando las diferencias entre ellos.

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

```
proc sort data=series.factor_v_no;
  by cluster;
run;
proc print data=series.factor_v_no;
  var PAIS cluster;
  by cluster;
  title "Paises Agrupados por Cluster (Analisis No Jerarquico)";
run;
```

A partir de este punto, y para el análisis del apartado posterior, se trabajará con el modelo no jerárquico con k=6, obtenido del análisis realizado en el apartado anterior. A continuación, se presentan las tablas que detallan la clasificación de los países en los diferentes clústeres definidos por este modelo.

Países Agrupados por Cluster (Análisis No Jerarquico)

Cluster=1

Obs	PAIS	CLUSTER
1	Austria	1
2	Belgium	1
3	Czech Republic	1
4	Estonia	1
5	Hungary	1
6	Latvia	1
7	Lithuania	1
8	Netherlands	1
9	Poland	1
10	Romania	1
11	Slovenia	1

Cluster=2

Obs	PAIS	CLUSTER
12	Germany	2

Cluster=3

Obs	PAIS	CLUSTER
13	Bulgaria	3

Cluster=4

Obs	PAIS	CLUSTER
14	Croatia	4
15	Denmark	4
16	Finland	4
17	Greece	4
18	Italy	4
19	Portugal	4
20	Spain	4

Cluster=5

Obs	PAIS	CLUSTER
21	France	5
22	Ireland	5

Cluster=6

Obs	PAIS	CLUSTER
23	Luxembourg	6
24	Malta	6
25	Slovakia	6

Para el Cluster 1, se reúne a 11 países del este y centro de Europa, como Austria, Bélgica, República Checa, Estonia, Hungría, Letonia, Lituania, Países Bajos, Polonia, Rumanía y Eslovenia. El Cluster 2 aísla a Alemania, reflejando sus características únicas. De forma similar, el Cluster 3 separa a Bulgaria como un grupo individual. Por otro lado, el Cluster 4 agrupa a 7 países mediterráneos y nórdicos, como Croacia, Dinamarca, Finlandia, Grecia, Italia, Portugal y España. El Cluster 5 incluye a Francia e Irlanda como un grupo diferenciado, mientras que el Cluster 6 reúne a Luxemburgo, Malta y Eslovaquia.

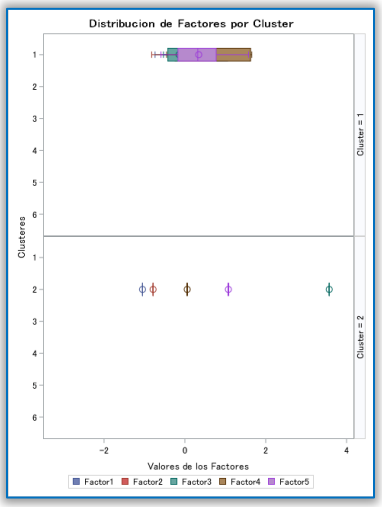
4.4. Representar los gráficos caja o los histogramas (lo que resulte más interpretable en cada caso) de los factores para cada uno de los clústeres. Teniendo en cuenta las variables que representa cada factor interpretar las diferencias entre los clústeres de países según los valores que presentan en los factores.

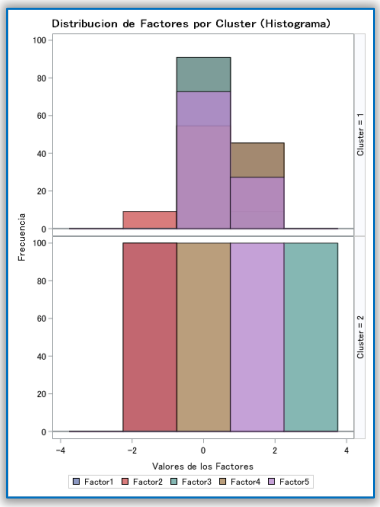
```
proc sgpanel data=series.factor_v_no;
  panelby cluster / layout=rowlattice;
  hbox Factor1 / category=cluster;
  hbox Factor2 / category=cluster;
  hbox Factor3 / category=cluster;
  hbox Factor4 / category=cluster;
  hbox Factor5 / category=cluster;
  colaxis label="Valores de los Factores";
  rowaxis label="Clusteres";
  title "Distribucion de Factores por Cluster";
run;
```

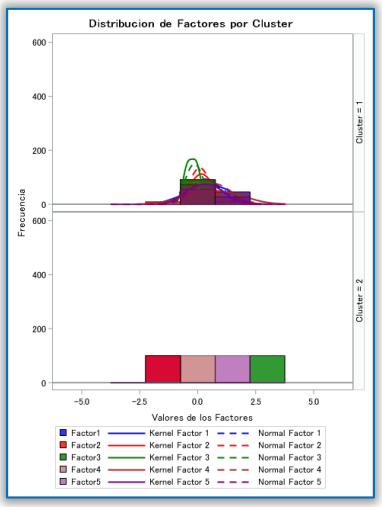
```
proc sgpanel data=series.factor_v_no;
  panelby cluster / layout=rowlattice;
  histogram Factor1 / transparency=0.2;
  histogram Factor2 / transparency=0.2;
  histogram Factor3 / transparency=0.2;
  histogram Factor4 / transparency=0.2;
  histogram Factor5 / transparency=0.2;
  colaxis label="Valores de los Factores";
  rowaxis label="Frecuencia";
  title "Distribucion de Factores por Cluster (Histograma)";
run;
```

```
proc sgpanel data=series.factor_v_no;
  panelby cluster / layout=rowlattice;

  /* Histogramas y densidades */
  histogram Factor1 / transparency=0.2 fillattrs=(color=blue) name="Factor1";
  density Factor1 / type=kernel lineattrs=(pattern=solid color=blue) legendlabel="Kernel Factor 1";
  density Factor1 / lineattrs=(pattern=dash color=blue) legendlabel="Normal Factor 1";

  histogram Factor2 / transparency=0.2 fillattrs=(color=red) name="Factor2";
  density Factor2 / type=kernel lineattrs=(pattern=solid color=red) legendlabel="Kernel Factor 2";
  density Factor2 / lineattrs=(pattern=dash color=red) legendlabel="Normal Factor 2";

  histogram Factor3 / transparency=0.2 fillattrs=(color=green) name="Factor3";
  density Factor3 / type=kernel lineattrs=(pattern=solid color=green) legendlabel="Kernel Factor 3";
  density Factor3 / lineattrs=(pattern=dash color=green) legendlabel="Normal Factor 3";

  histogram Factor4 / transparency=0.5 fillattrs=(color=brown) name="Factor4";
  density Factor4 / type=kernel lineattrs=(pattern=solid color=brown) legendlabel="Kernel Factor 4";
  density Factor4 / lineattrs=(pattern=dash color=brown) legendlabel="Normal Factor 4";

  histogram Factor5 / transparency=0.5 fillattrs=(color=purple) name="Factor5";
  density Factor5 / type=kernel lineattrs=(pattern=solid color=purple) legendlabel="Kernel Factor 5";
  density Factor5 / lineattrs=(pattern=dash color=purple) legendlabel="Normal Factor 5";

  /* Ejes y titulos */
  colaxis label="Valores de los Factores";
  rowaxis label="Frecuencia";
  title "Distribucion de Factores por Cluster";

  /* Leyenda */
  keylegend / across=3 position=bottom;
run;
```

La representación propuesta consiste en establecer en el eje **Y** los números de los 6 clústeres obtenidos a partir del análisis y en el eje **X** los valores de cada factor asociados a cada clúster. Esto significa que cada punto o segmento representará el peso promedio o distribución de los factores dentro de cada clúster, permitiendo analizar las diferencias y similitudes entre los grupos en términos de los factores principales.





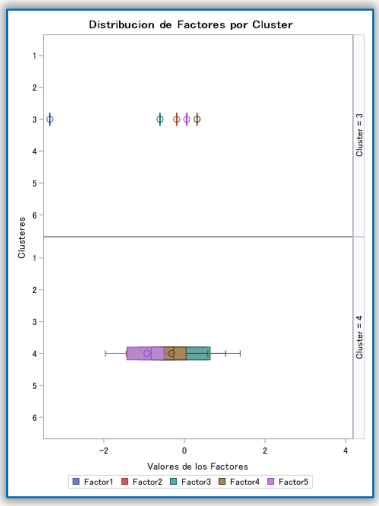


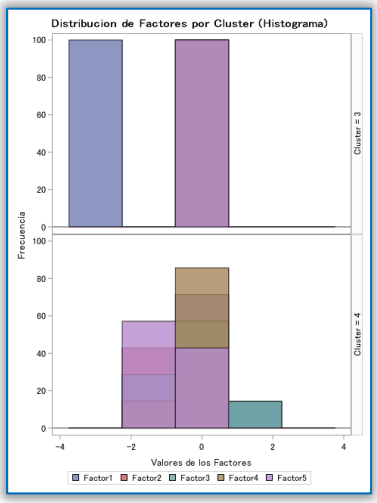
En el Cluster 1, se observa que los factores 1 (Satisfacción de Vida), 3 (Composición demográfica), 4 (Mortalidad) y 5 (Empleo) tienen un peso positivo (entre 0 y 2) sobre los países del grupo. Sin embargo, los factores 4 (Mortalidad) y 5 (Empleo) presentan las mayores dispersiones, lo que indica una menor homogeneidad en estos aspectos entre los países del clúster.

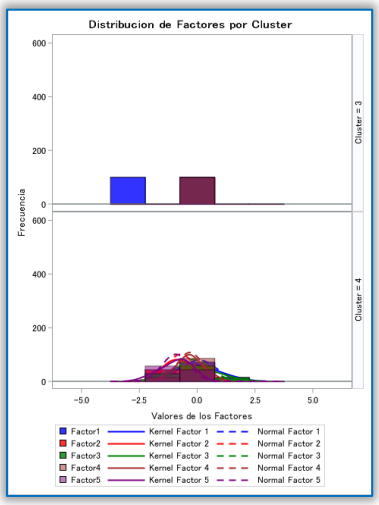
Esto sugiere que, para los 11 países que conforman este grupo (Austria, Bélgica, República Checa, Estonia, Hungría, Letonia, Lituania, Países Bajos, Polonia, Rumanía y Eslovenia), existe una relación positiva destacada con las variables asociadas al Factor 4 (T_Mort, T_Mort_Accidente y Renta) y al Factor 5 (15-64_YEARS y T_Desemp). Esto implica que estos países comparten similitudes en términos de índices de empleo y actividad laboral, así como en aspectos relacionados con las tasas de mortalidad y el bienestar económico, aunque con diferencias internas más marcadas en estos dos últimos factores.

Para el Cluster 2, compuesto únicamente por Alemania, se observa un peso positivo (aprox. 1) destacado en el Factor 5 (Empleo) y, en mayor medida(aprox. 3.5), en el Factor 3 (Composición demográfica). Esto indica que Alemania se caracteriza por variables asociadas al Factor 3 como (T_UNI), (T_Migra_A) y (Pob_T), , así como (15-64_YEARS) y (T_Desemp), reflejadas en el Factor 5.

Dado que este clúster solo incluye a Alemania, no se observa dispersión, ya que todos los valores se concentran exactamente en las medianas de los factores, destacando a este país como un caso particular que se diferencia significativamente de los otros clústeres.





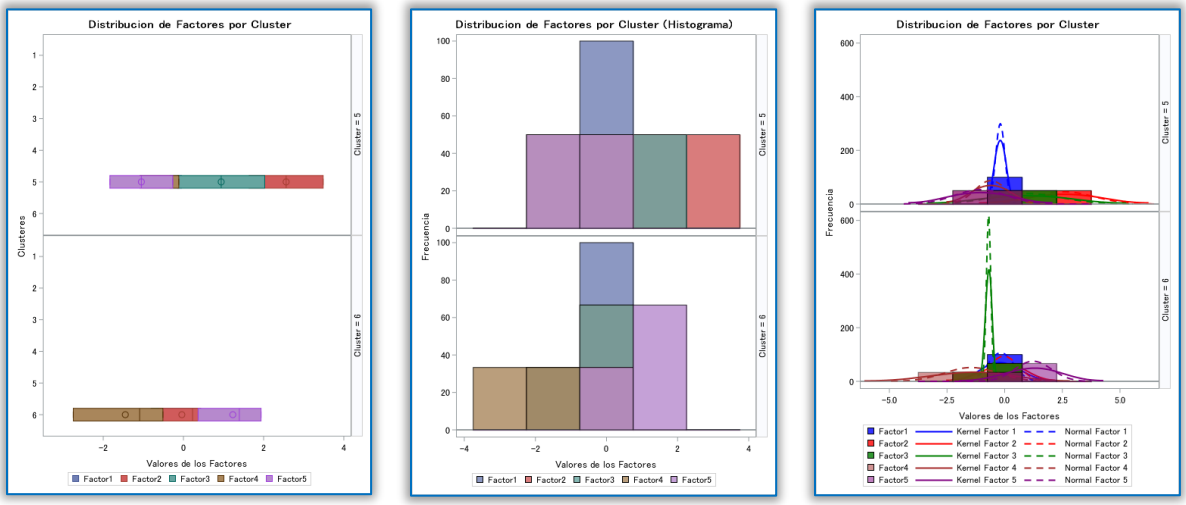


Para el Cluster 3, compuesto únicamente por Bulgaria, destaca un alto peso negativo(aprox. -4) con el Factor 1 (Satisfacción de Vida). Este factor está definido por variables como: ISV_Small_Towns, ISV_Rural_Areas, ISV_Cities e I_Satisf_Vida; que reflejan percepciones y niveles de satisfacción de vida en diferentes áreas geográficas. La relación negativa de Bulgaria con este factor sugiere niveles bajos de satisfacción de vida en comparación con otros países del análisis.

Además, dado que el clúster contiene solo un país, los valores están completamente centrados en la mediana, lo que significa que no hay dispersión en este grupo. Esta característica refuerza la posición singular de Bulgaria en el análisis, destacando como un país con indicadores de satisfacción de vida notablemente bajos en el contexto europeo.

En el Cluster 4, predomina el Factor 5 (Empleo), que presenta una alta dispersión en el rango negativo (entre -2 y 0). Esto indica que los países de este clúster comparten características relacionadas con un entorno laboral menos favorable, como una baja proporción de población activa (15-64 años) y altos

niveles de desempleo. La gran dispersión sugiere diferencias significativas entre los países en estos aspectos. Este clúster, compuesto por 7 países (Croacia, Dinamarca, Finlandia, Grecia, Italia, Portugal y España), refleja una conexión entre estos países mediterráneos y nórdicos en términos de desafíos en el empleo y la actividad económica, aunque con variaciones importantes dentro del grupo.



En el Cluster 5, se observa una gran dispersión en los factores, particularmente en el Factor 5 (Empleo), que muestra valores negativos significativos en el rango de -2 a -0.5. Este factor está definido por las variables 15-64_YEARS y T_Desemp, lo que sugiere que Francia e Irlanda, los países que conforman este clúster, presentan desafíos en términos de empleo y actividad laboral, con niveles elevados de desempleo o una menor proporción de población activa.

Por otro lado, los Factores 3 (Composición Demográfica) y Factor 4 (Mortalidad) también tienen un peso destacado. El Factor 3, definido por variables como Pob_T, T_UNI y T_Migra_A , muestra valores tanto positivos como muy elevados (rango de 2 a 3.5), lo que indica que estos países presentan características favorables en términos de urbanización y migración neta. El Factor 4 (Mortalidad), relacionado con las variables T_Mort, T_Mort_Accidente y Renta, aunque con menos dispersión, tiene un peso ligeramente negativo (-0.5), reflejando similitudes en términos de tasas de mortalidad general y renta en el grupo.

Esta combinación de factores revela que Francia e Irlanda comparten un perfil particular, con aspectos positivos asociados a la composición demográfica, urbanización y migración, pero con importantes retos en el ámbito del empleo y el bienestar económico.

En el Cluster 6, se observa un peso negativo en el Factor 4 (Mortalidad) con una amplia dispersión, que abarca valores entre -3 y -0.5. Este factor está relacionado con las variables T_Mort, T_Mort_Accidente y Renta. Los valores negativos sugieren que los países en este clúster, Luxemburgo, Malta y Eslovaquia tienen características diferenciadas en términos de bienestar y mortalidad, con tasas más bajas de mortalidad general y por accidentes en comparación con otros grupos.

Por otro lado, el Factor 5 (Empleo) presenta un peso positivo con una dispersión más moderada, en un rango de 0.5 a 2. Este factor está definido por las variables 15-64_YEARS y T_Desemp. Los valores positivos indican una relación favorable en términos de empleo y actividad laboral, reflejando una mayor proporción de población activa y tasas de desempleo más bajas.
En conjunto, este análisis revela que Luxemburgo, Malta y Eslovaquia comparten características relacionadas con un entorno laboral positivo y niveles relativamente bajos de mortalidad y desigualdad, aunque con diferencias internas significativas en el ámbito del bienestar y el empleo.