

1. Analizar la necesidad de trabajar con la información inicial bruta o transformarla en información neta.

Antes de proceder a reducir las dimensiones de las variables de la base de datos realizaremos un análisis descriptivo simple con el fin de determinar si se emplea o no la matriz de covarianza.

```
proc means data=Tmp1.Seprovincias_e mean std min max;
run;
```

Una vez ejecutado el código anterior, vemos que las escalas de las variables son muy distintas entre ellas y encontramos desde porcentajes hasta número de población. Obviamente, se aprecia que no todas las variables siguen una misma escala. Es por ello que emplearemos la matriz de correlación con el fin de que los factores no se vean sesgados por las grandes varianzas de los datos.

Variable	Label	Mean	Std Dev	Minimum	Maximum
PT	Población Total	899448.87	1154297.28	84509.00	6454440.00
P_M	Población Total (mujeres)	457416.63	595416.86	41434.00	3354799.00
T_M	Mortalidad	9.3790385	2.1206790	5.8200000	14.3600000
T_N	Tasa Bruta de Natalidad (nacidos por mil habitantes)	8.8388462	2.1387936	5.5500000	19.3300000
IPC	IPC	102.3501346	0.8193323	100.6000000	104.7620000
NE	Número de empresas	61286.12	90519.98	3749.00	508612.00
NE_C	Construcción (nº empresas)	7804.79	10473.73	309.0000000	59661.00
NE_CTH	Comercio, transporte y hostelería (nº empresas)	23741.21	30174.42	2030.00	158331.00
NE_I	Información y comunicaciones (nº empresas)	1131.88	2983.10	35.0000000	19058.00
AC_F	Actividades financieras y de seguros (nº empresas)	1378.31	2068.40	50.0000000	12357.00
AC_P	Actividades profesionales y técnicas (nº empresas)	10853.50	20403.75	504.0000000	123863.00
AC_S	Educación, sanidad y servicios sociales (nº empresas)	4641.94	7833.55	313.0000000	44730.00
T_AC	Tasa Actividad (%)	57.8400000	4.0424784	47.4100000	68.6900000
OC	Ocupados (miles de personas)	347.0903846	486.2721465	24.6000000	2806.40
PIB	PIB a precios de mercado (miles de euros)	20275144.96	32772009.04	1397441.00	198652445
CA_E	Censo Agrario Número de Explotaciones	19034.54	14267.64	3.0000000	68037.00
CA_S	Censo Agrario Superficie agrícola	456782.47	339925.46	16.7200000	1491594.85
CV_F	Censo 2011: Total viviendas familiares	484781.19	537872.14	26233.00	2894679.00
CV_P	Censo 2011: Total viviendas principales	347763.31	445461.00	24666.00	2469378.00
CV_V	Censo 2011: Viviendas vacías	66218.54	60511.10	1335.00	283155.00

Tras esto, aplicaremos una matriz de correlación para ver si existen variables correlacionadas y por ende saber si tiene sentido aplicar un análisis factorial. Para ello, nos ayudaremos del procedimiento corr y de Excel (para generar un mapa de calor), obteniendo de esta forma el siguiente mapa de calor:

```
proc corr data=Tmp1.Seprovincias_e;run;
```

	PT	P_M	T_M	T_N	IPC	NE	NE_C	NE_CTH	NE_I	AC_F	AC_P	AC_S	T_AC	OC	PIB	CA_E	CA_S	CV_F	CV_P	CV_V	
PT	1,000	1,000	-0,344	0,114	0,335	0,995	0,984	0,996	0,945	0,992	0,980	0,990	0,334	0,996	0,981	0,097	-0,166	0,991	0,999	0,909	Correlación
	0,000	0,000	0,013	0,419	0,015	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,016	0,000	0,000	0,493	0,240	0,000	0,000	0,000	P-valor
P_M	1,000	1,000	-0,338	0,112	0,335	0,995	0,985	0,995	0,948	0,993	0,982	0,991	0,329	0,996	0,983	0,093	-0,166	0,990	0,999	0,906	Correlación
	0,000	0,000	0,014	0,430	0,015	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,017	0,000	0,000	0,512	0,241	0,000	0,000	0,000	P-valor
T_M	-0,344	-0,338	1,000	-0,737	0,188	-0,309	-0,296	-0,329	-0,260	-0,305	-0,296	-0,292	-0,733	-0,328	-0,296	0,016	0,282	-0,332	-0,326	-0,315	Correlación
	0,013	0,014	0,000	0,000	0,000	0,000	0,000	0,017	0,063	0,028	0,033	0,036	0,000	0,018	0,033	0,908	0,043	0,016	0,018	0,023	P-valor
T_N	0,114	0,112	-0,737	1,000	-0,254	0,105	0,089	0,101	0,115	0,105	0,113	0,108	0,472	0,112	0,114	-0,123	-0,252	0,084	0,100	0,040	Correlación
	0,419	0,430	0,000	0,000	0,069	0,459	0,532	0,474	0,419	0,460	0,424	0,445	0,000	0,430	0,422	0,385	0,071	0,552	0,480	0,781	P-valor
IPC	0,335	0,335	0,188	-0,254	1,000	0,365	0,400	0,362	0,300	0,323	0,333	0,359	0,090	0,358	0,355	-0,191	-0,244	0,341	0,350	0,274	Correlación
	0,015	0,015	0,181	0,069	0,000	0,008	0,003	0,008	0,031	0,020	0,016	0,009	0,527	0,009	0,010	0,175	0,082	0,013	0,011	0,049	P-valor
NE	0,995	0,995	-0,309	0,105	0,365	1,000	0,995	0,994	0,963	0,993	0,991	0,997	0,332	0,998	0,990	0,045	-0,180	0,982	0,996	0,886	Correlación
	0,000	0,000	0,026	0,459	0,008	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,016	0,000	0,000	0,752	0,201	0,000	0,000	0,000	P-valor
NE_C	0,984	0,985	-0,296	0,089	0,400	0,995	1,000	0,986	0,958	0,984	0,985	0,989	0,341	0,993	0,985	0,030	-0,201	0,976	0,986	0,882	Correlación
	0,000	0,000	0,033	0,532	0,003	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,013	0,000	0,000	0,834	0,152	0,000	0,000	0,000	P-valor
NE_CTH	0,996	0,995	-0,329	0,101	0,362	0,994	0,986	1,000	0,929	0,983	0,971	0,986	0,326	0,991	0,972	0,094	-0,183	0,991	0,997	0,918	Correlación
	0,000	0,000	0,017	0,474	0,008	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,018	0,000	0,000	0,509	0,195	0,000	0,000	0,000	P-valor
NE_I	0,945	0,948	-0,260	0,115	0,300	0,963	0,958	0,929	1,000	0,969	0,990	0,974	0,308	0,964	0,985	-0,066	-0,155	0,912	0,944	0,761	Correlación
	0,000	0,000	0,063	0,419	0,031	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,026	0,000	0,000	0,640	0,273	0,000	0,000	0,000	P-valor
AC_F	0,992	0,993	-0,305	0,105	0,323	0,993	0,984	0,983	0,969	1,000	0,992	0,992	0,317	0,995	0,989	0,086	-0,154	0,979	0,991	0,885	Correlación
	0,000	0,000	0,028	0,460	0,020	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,022	0,000	0,000	0,546	0,277	0,000	0,000	0,000	P-valor
AC_P	0,980	0,982	-0,296	0,113	0,333	0,991	0,985	0,971	0,990	0,992	1,000	0,995	0,332	0,992	0,997	-0,006	-0,173	0,957	0,980	0,833	Correlación
	0,000	0,000	0,033	0,424	0,016	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,016	0,000	0,000	0,969	0,219	0,000	0,000	0,000	P-valor
AC_S	0,990	0,991	-0,292	0,108	0,359	0,997	0,989	0,986	0,974	0,992	0,995	1,000	0,318	0,996	0,995	0,018	-0,172	0,970	0,991	0,858	Correlación
	0,000	0,000	0,036	0,445	0,009	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,022	0,000	0,000	0,898	0,222	0,000	0,000	0,000	P-valor
T_AC	0,334	0,329	-0,733	0,472	0,090	0,332	0,341	0,326	0,308	0,317	0,332	0,318	1,000	0,354	0,331	-0,117	-0,192	0,327	0,324	0,266	Correlación
	0,016	0,017	0,000	0,000	0,527	0,016	0,013	0,018	0,026	0,022	0,016	0,022	0,000	0,010	0,017	0,410	0,172	0,018	0,019	0,057	P-valor
OC	0,996	0,996	-0,328	0,112	0,358	0,998	0,993	0,991	0,964	0,995	0,992	0,996	0,354	1,000	0,992	0,049	-0,181	0,982	0,996	0,883	Correlación
	0,000	0,000	0,018	0,430	0,009	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,010	0,000	0,000	0,732	0,199	0,000	0,000	0,000	P-valor
PIB	0,981	0,983	-0,296	0,114	0,355	0,990	0,985	0,972	0,985	0,989	0,997	0,995	0,331	0,992	1,000	-0,007	-0,176	0,957	0,982	0,830	Correlación
	0,000	0,000	0,033	0,422	0,010	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,017	0,000	0,000	0,960	0,211	0,000	0,000	0,000	P-valor
CA_E	0,097	0,093	0,016	-0,123	-0,191	0,045	0,030	0,094	-0,066	0,086	-0,006	0,018	-0,117	0,049	-0,007	1,000	0,438	0,153	0,095	0,338	Correlación
	0,493	0,512	0,908	0,385	0,175	0,752	0,834	0,509	0,640	0,546	0,969	0,898	0,410	0,732	0,960	0,000	0,001	0,279	0,502	0,014	P-valor
CA_S	-0,166	-0,166	0,282	-0,252	-0,244	-0,180	-0,201	-0,183	-0,155	-0,154	-0,173	-0,172	-0,192	-0,181	-0,176	0,438	1,000	-0,177	-0,170	-0,153	Correlación
	0,240	0,241	0,043	0,071	0,082	0,201	0,152	0,195	0,273	0,277	0,219	0,222	0,172	0,199	0,211	0,001	0,000	0,210	0,229	0,279	P-valor
CV_F	0,991	0,990	-0,332	0,084	0,341	0,982	0,976	0,991	0,912	0,979	0,957	0,970	0,327	0,982	0,957	0,153	-0,177	1,000	0,992	0,951	Correlación
	0,000	0,000	0,016	0,552	0,013	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,018	0,000	0,000	0,279	0,210	0,000	0,000	0,000	P-valor
CV_P	0,999	0,999	-0,326	0,100	0,350	0,996	0,986	0,997	0,944	0,991	0,980	0,991	0,324	0,996	0,982	0,095	-0,170	0,992	1,000	0,910	Correlación
	0,000	0,000	0,018	0,480	0,011	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,019	0,000	0,000	0,502	0,229	0,000	0,000	0,000	P-valor
CV_V	0,909	0,906	-0,315	0,040	0,274	0,886	0,882	0,918	0,761	0,885	0,833	0,858	0,266	0,883	0,830	0,338	-0,153	0,951	0,910	1,000	Correlación
	0,000	0,000	0,023	0,781	0,049	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,057	0,000	0,000	0,014	0,279	0,000	0,000	0,000	P-valor

Como se puede observar, existen muchas variables que están correlacionadas entre sí y con p-valores bajos, por lo que tendrá sentido aplicar un análisis factorial y por tanto reducir la dimensión de las variables originales a una dimensión menor.

2. De ser necesaria una transformación de la información inicial, nos quedaremos las puntuaciones factoriales de cada provincia proporcionadas por un modelo factorial subyacentes rotados con varimax.

Una vez seleccionado el método de cálculo del análisis de factores, ejecutaremos esta sintaxis:

```
PROC FACTOR DATA=Tmpl1.Seprovincias_e CORR OUTSTAT=ESTEJ1F
out=PROEJ1F RESIDUALS NFACT=3 MSA SCREE rotate=varimax;
RUN;
```

Aplicando en análisis factorial de los datos, obtenemos la siguiente tabla de pesos de los componentes:

Eigenvalues of the Correlation Matrix: Total = 20 Average = 1				
	Eigenvalue	Difference	Proportion	Cumulative
1	14.0145146	11.7132480	0.7007	0.7007
2	2.3012666	0.7328420	0.1151	0.8158
3	1.5684245	0.8270720	0.0784	0.8942
4	0.7413525	0.0704633	0.0371	0.9313
5	0.6708892	0.2870240	0.0335	0.9648
6	0.3838652	0.2195120	0.0192	0.9840
7	0.1643532	0.0608898	0.0082	0.9922
8	0.1034635	0.0792856	0.0052	0.9974
9	0.0241779	0.0099306	0.0012	0.9986
10	0.0142473	0.0095722	0.0007	0.9993
11	0.0046750	0.0011316	0.0002	0.9996
12	0.0035435	0.0012930	0.0002	0.9997
13	0.0022505	0.0010282	0.0001	0.9999
14	0.0012223	0.0003368	0.0001	0.9999
15	0.0008855	0.0003708	0.0000	1.0000
16	0.0005147	0.0003516	0.0000	1.0000
17	0.0001631	0.0000383	0.0000	1.0000
18	0.0001248	0.0000671	0.0000	1.0000
19	0.0000577	0.0000493	0.0000	1.0000
20	0.0000084		0.0000	1.0000

Aplicando el criterio de selección de los pesos con un valor superior a 1, nos quedaremos con los 3 primeros factores.

Eigenvalues of the Correlation Matrix: Total = 20 Average = 1				
	Eigenvalue	Difference	Proportion	Cumulative
1	14.0145146	11.7132480	0.7007	0.7007
2	2.3012666	0.7328420	0.1151	0.8158
3	1.5684245	0.8270720	0.0784	0.8942

Siendo la variabilidad total explicada de estas 3 componentes de: 0.8942 es decir de aproximadamente un 90%.

Por lo que ejecutamos el mismo código pero con la diferencia de que nos calcule todo con solo 3 factores.

```
PROC FACTOR DATA=Tmp1.Seprovincias_e CORR OUTSTAT=ESTEJ1F
out=PROEJ1F
RESIDUALS NFACT=3 MSA SCREE;
RUN;
```

En lo que a los factores respecta, obtenemos la siguiente tabla

Nota: con el fin de facilitar la comprensión de los datos se ha procedido a crear con la ayuda de Excel un mapa de calor con el fin de ver a simple vista que variables componen cada factor.

Variable	Etiqueta	Factor1	Factor2	Factor3
PT	Población Total	0,98974	0,12401	-0,02769
P_M	Población Total (mujeres)	0,99033	0,11918	-0,03045
T_M	Mortalidad	-0,22515	-0,92339	0,00931
T_N	Tasa Bruta de Natalidad (nacidos por mil habitantes)	0,00428	0,87079	-0,07123
IPC	IPC	0,37906	-0,37535	-0,53332
NE	Número de empresas	0,99029	0,09972	-0,07947
NE_C	Construcción (nº empresas)	0,98468	0,08525	-0,11093
NE_CTH	Comercio, transporte y hostelería (nº empresas)	0,98778	0,10683	-0,04453
NE_I	Información y comunicaciones (nº empresas)	0,94597	0,08911	-0,13071
AC_F	Actividades financieras y de seguros (nº empresas)	0,98944	0,10121	-0,03137
AC_P	Actividades profesionales y técnicas (nº empresas)	0,97735	0,10329	-0,1035
AC_S	Educación, sanidad y servicios sociales (nº empresas)	0,98608	0,09028	-0,09357
T_AC	Tasa Actividad (%)	0,25817	0,72841	-0,15933
OC	Ocupados (miles de personas)	0,98902	0,11741	-0,07597
PIB	PIB a precios de mercado (miles de euros)	0,97758	0,09837	-0,11353
CA_E	Censo Agrario Número de Explotaciones	0,13714	-0,07151	0,85648
CA_S	Censo Agrario Superficie agrícola	-0,10903	-0,26214	0,73156
CV_F	Censo 2011: Total viviendas familiares	0,98492	0,10766	0,00346
CV_P	Censo 2011: Total viviendas principales	0,99177	0,10569	-0,03587
CV_V	Censo 2011: Viviendas vacías	0,90795	0,09067	0,15322

Viendo la tabla anterior, vemos que:

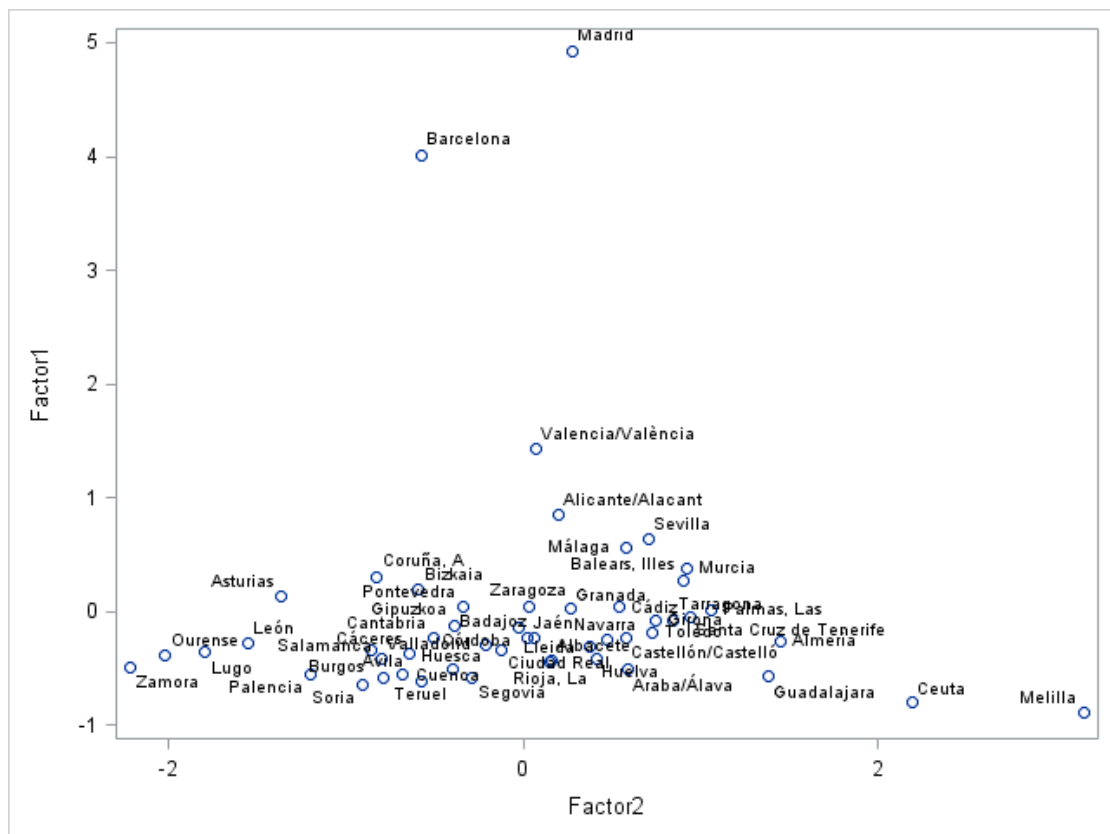
1. El **primer factor** se ve alimentado por variables referentes a población total, número de empresas, censos (salvo los referentes al entorno agrario), actividades económicas, PIB, etc... acompañado negativamente de la pérdida de población por defunciones y el capital humano del campo. En resumen, una posible interpretación de este factor sería el **poder económico que tiene cada provincia en relación al tejido industrial/financiero y el capital humano disponible**.
2. El **segundo factor** se ve alimentado por variables referentes a la natalidad y tasa de actividad siendo el principal aporte económico la mortalidad, por lo que podría hacer referencia a la **mano de obra trabajadora de la provincia**.

3. El **tercer factor** se ve alimentado por variables referentes a la **población rural de cada provincia**.

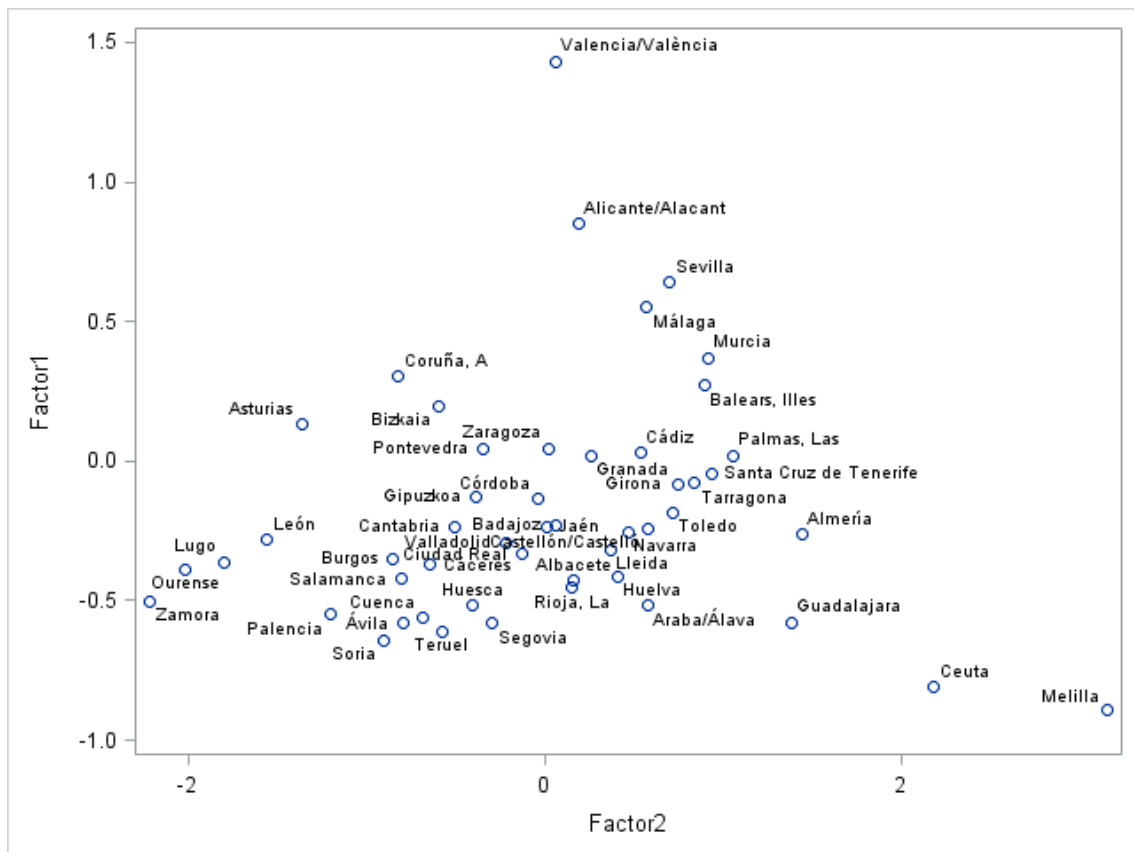
Si representamos en un gráfico de dispersión las proyecciones de cada provincia en función de cada factor obtendremos:

Poder económico (Factor 1) mano de obra trabajadora (Factor2)

Como se puede observar en el gráfico de dispersión, vemos que las provincias de Madrid y Barcelona toman valores extremos en lo que a poder económico respecta (es por esto, que el grafico sale tan compacto).



Por lo que si eliminamos estas observaciones con el único fin de ver cómo se comportan el resto de provincias en la nube de puntos compacta, obtendremos el siguiente gráfico:



Nota: esto se ha realizado con el fin de mejorar la visualización de los datos, en ningún momento se eliminaran permanentemente estas variables.

En la gráfica anterior (recordemos que no aparecen Madrid ni Barcelona) vemos ya una nube de puntos más dispersa. En lo que a los significados por sectores respecta podemos dividir la gráfica (mediante los ejes de coordenadas):

1. **Factor 1:** todos los valores que estén por encima de 0, significara que tendrán un mayor poder económico respecto a las provincias que estén en valores negativos. Por lo que cuanto mayor altura tenga la provincia, más poder económico tendrá.
2. **Factor 2:** todas las provincias que tomen valores positivos, significara que tienen una tasa de natalidad alta junto con una gran tasa de actividad, es decir, que disponen de una gran cantidad de mano de obra respecto a las provincias con valores negativos.

Poder económico (Factor 1) frente a población rural (Factor3)

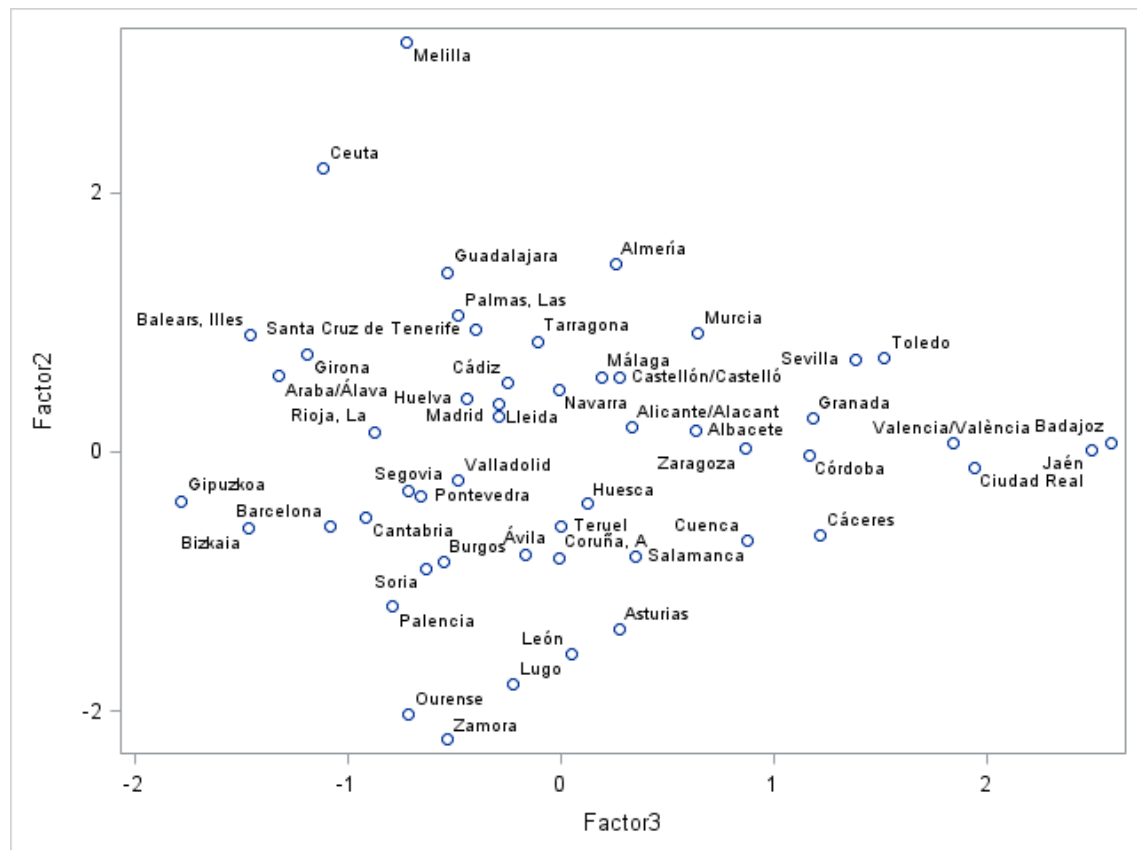
En este grafico vemos otra vez como tanto Madrid como Barcelona toman valores muy altos, por lo que procederemos de la misma forma que en el grafico anterior (eliminado temporalmente dichas variables para apreciar mejor la nube de puntos compactada por la escala).

Si dividimos el grafico mediante los ejes de coordenadas y creamos sectores imaginarios, podremos ver que:

1. **Factor 1:** todos los valores que estén por encima de 0, significara que tendrán un mayor poder económico respecto a las provincias que estén en valores negativos. Por lo que cuanto mayor altura tenga la provincia, más poder económico tendrá.
2. **Factor 3:** todas las provincias que tomen valores positivos significara que tendrán una población más rural mientras que cuanto más negativo sea esta, menor población rural tendrá. Es decir, que cuanto más a la izquierda está más población dedicada al sector rural tendrá.

Mano de obra trabajadora (Factor 2) frente a población rural (Factor3)

Finalmente en este grafico vemos que no existen datos con valores altos (como si pasaba en las anteriores graficas) y vemos que los datos están repartidos en una nube más concentrada. En lo que a la interpretación del significado de este grafico respecta, tenemos que más arriba se este, mas población nace junto con tasas de actividad altas y cuanto más a la derecha este, mayor población rural tendrá.



3. Detectar el/los posible/s valor/es para el número de clústeres a formar.

Una vez ya tenemos agrupadas las variables originales en 3 variables gracias al análisis factorial, procederemos a realizar el análisis clúster con las variables ya proyectadas en los diferentes factores.

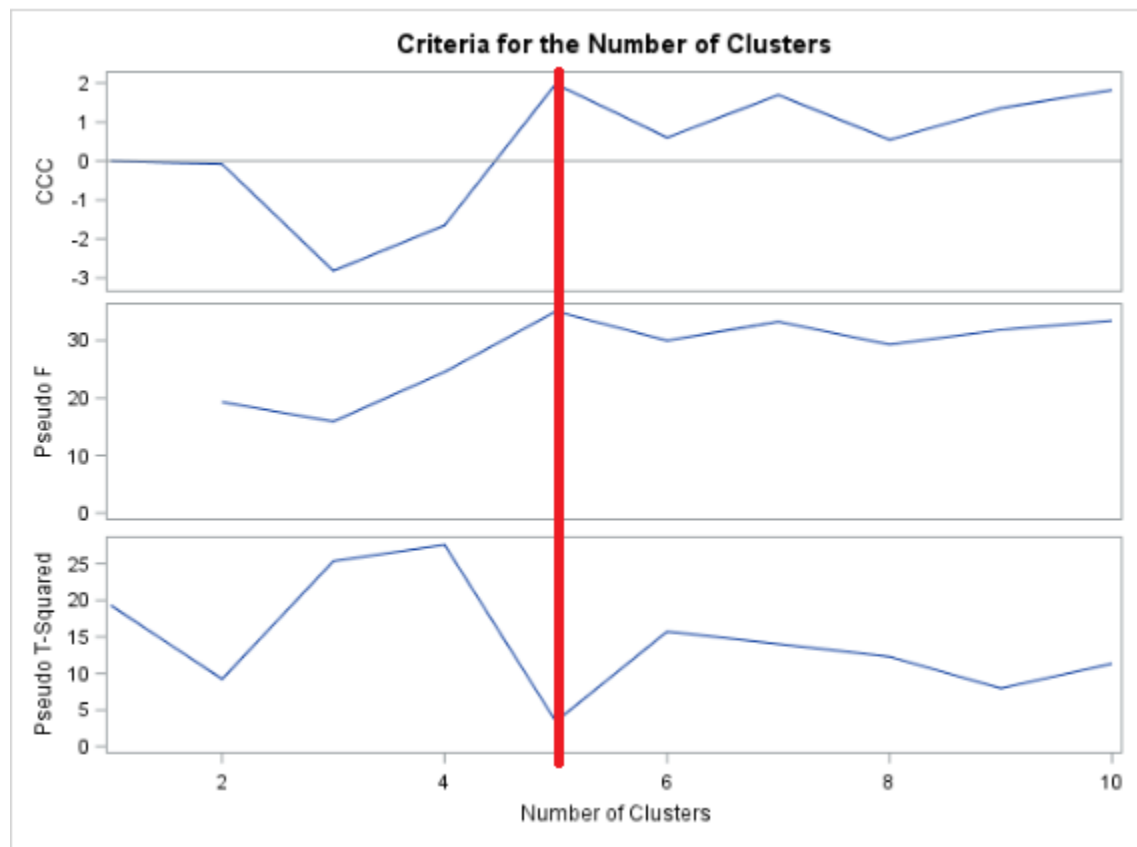
Para realizar en análisis clúster, aplicaremos la siguiente sentencia:

```
proc cluster DATA= PROEJ1F METHOD=AVE plots=all ccc pseudo;  
  VAR Factor1 Factor2 Factor3;  
  id Provincia;  
run;
```

Si nos fijamos en los estadísticos ccc y pseudo del procedimiento cluster, detectamos que la combinación óptima de forma que:

1. **CCC:** tome valores altos
2. **Pseudo F:** tome valores altos
3. **Pseudos T:** tome valores bajos

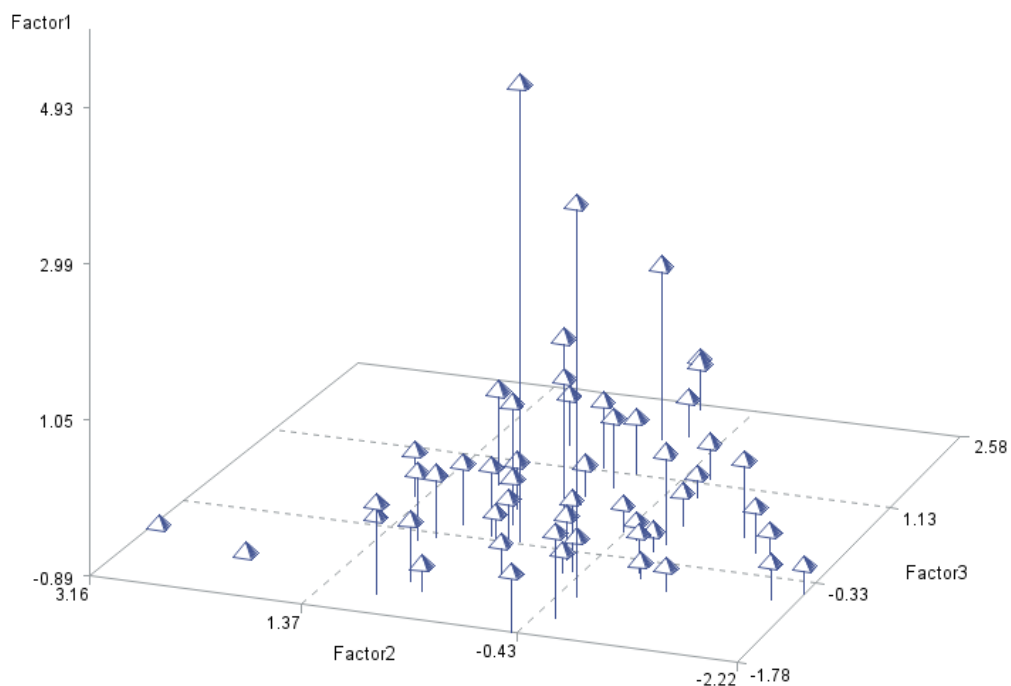
Por lo que si nos fijamos en el gráfico, nos quedaremos con un total de 5 clusters debido a que cumple los puntos anteriores.



Aplicando un gráfico 3d de los 3 factores mediante el siguiente código:

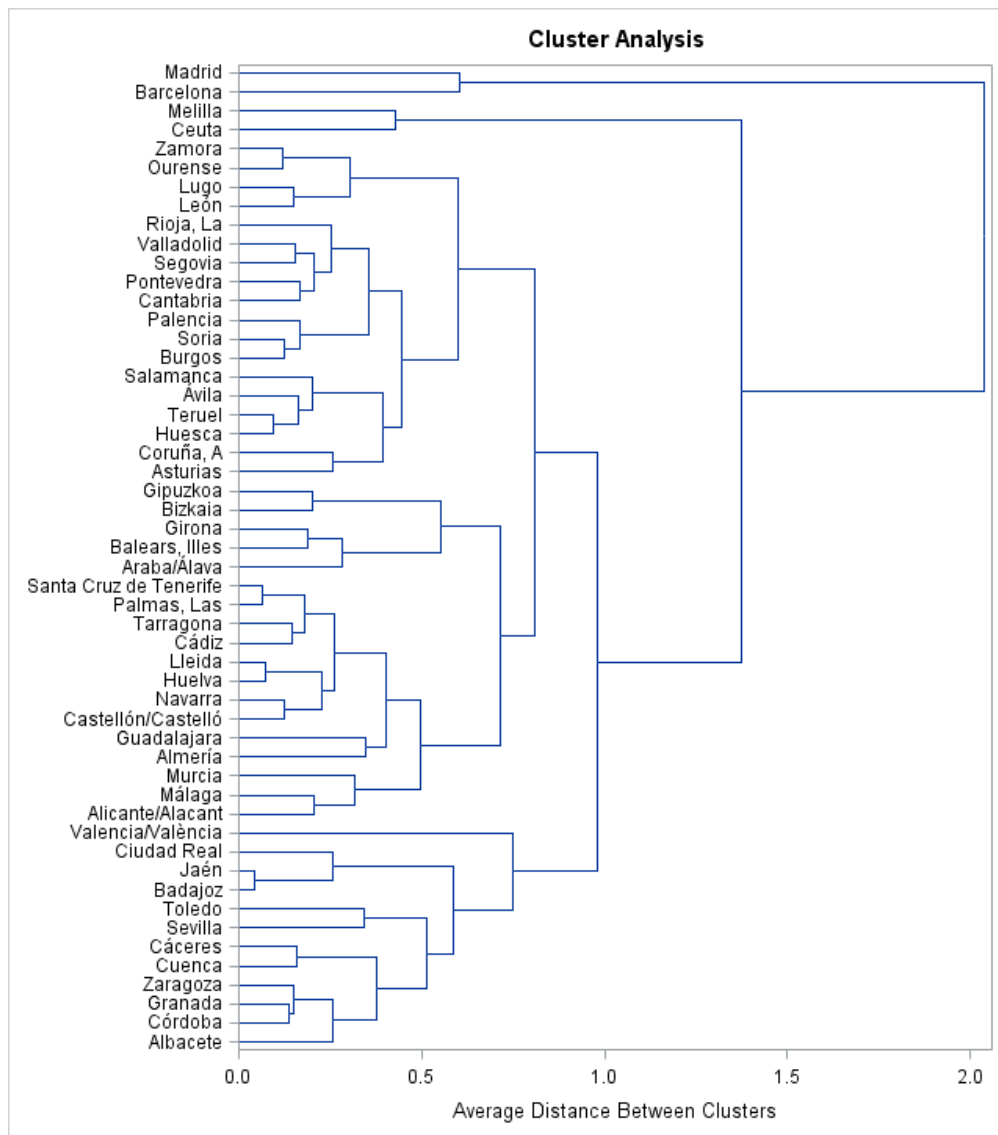
```
proc g3d data=PROEJ1F;  
  scatter Factor2*Factor3=Factor1;  
run;
```

Dando como resultado:



Una vez aplicado el anterior código, obtenemos el siguiente árbol de agrupación de clúster en un espacio de dimensión 3, donde recordemos que las dimensiones de esta nueva base de datos son:

1. Poder económico (Factor1)
2. Mano de obra trabajadora (Factor 2)
3. Población rural (Factor3)

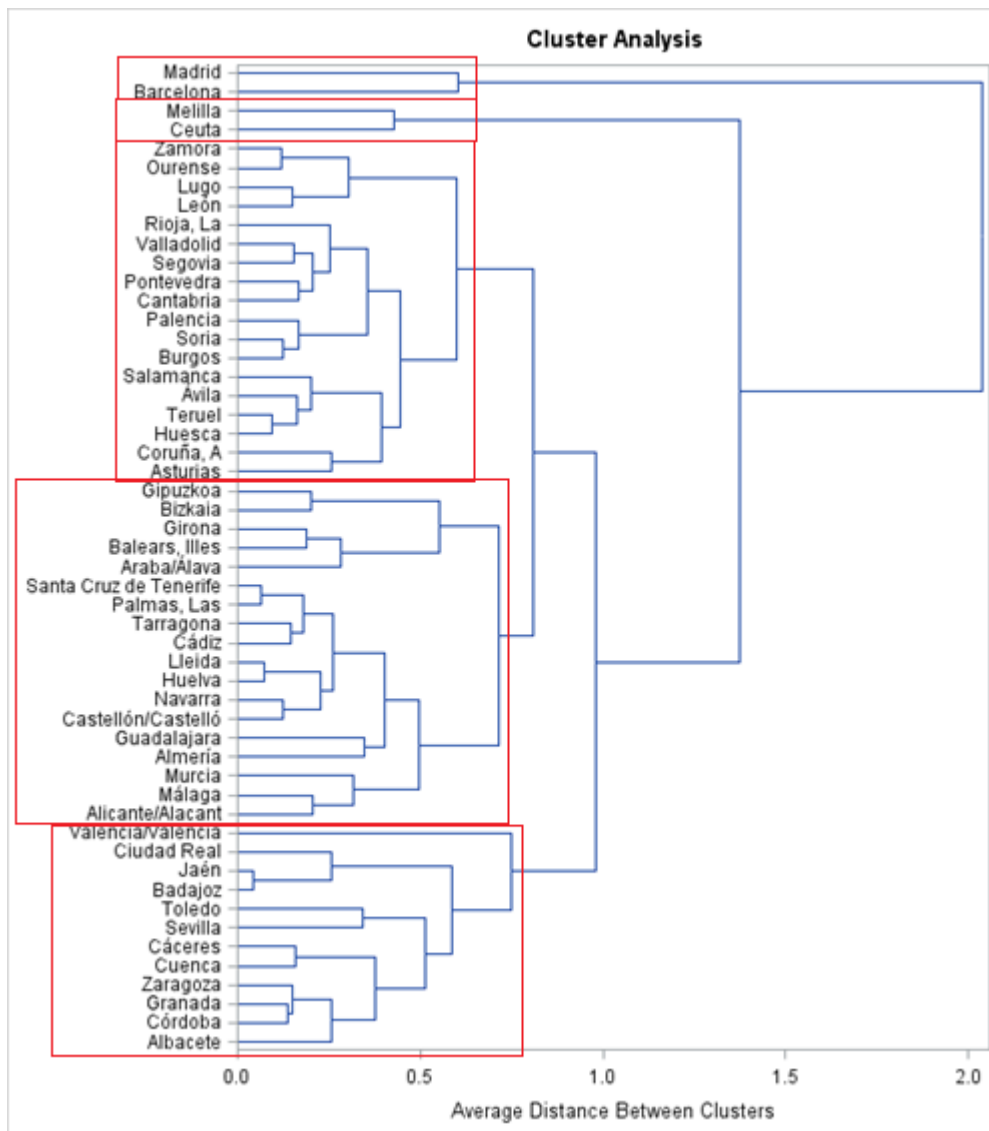


Cabe destacar que independientemente del número de clúster que se seleccionen, vemos que tanto Madrid como Barcelona permanecen juntos. Esto se debe a los valores altos que tienen dichas variables en 2 de las 3 factores, dando como resultado que las distancias de las coordenadas solo puedan ser agrupadas en un grupo compuesta por ellas.

Por otra parte, tanto Melilla como Ceuta también tienen un comportamiento muy similar y permanecen separadas del resto de provincias (como también le pasa a Madrid y Barcelona).

Finalmente, si nos fijamos en el resto de clúster observamos 3 grandes grupos (sin contar con Madrid y Barcelona).

Y junto a los estadísticos obtenidos anteriormente, dividiremos el cluster en un total de 5 grupos, siendo los grupos formados por el procedimiento cluster:



4. Conformer por un método no jerárquico la composición de los clústeres para el/los valor/es en el punto 3. Dar una interpretación a esta estructura de las provincias.

Por otra parte, si aplicamos un procedimiento de cluster no jerarquizado y ponemos como máximo un total de 5 cluster, obtendremos (mediante el siguiente código) la siguiente representación gráfica:

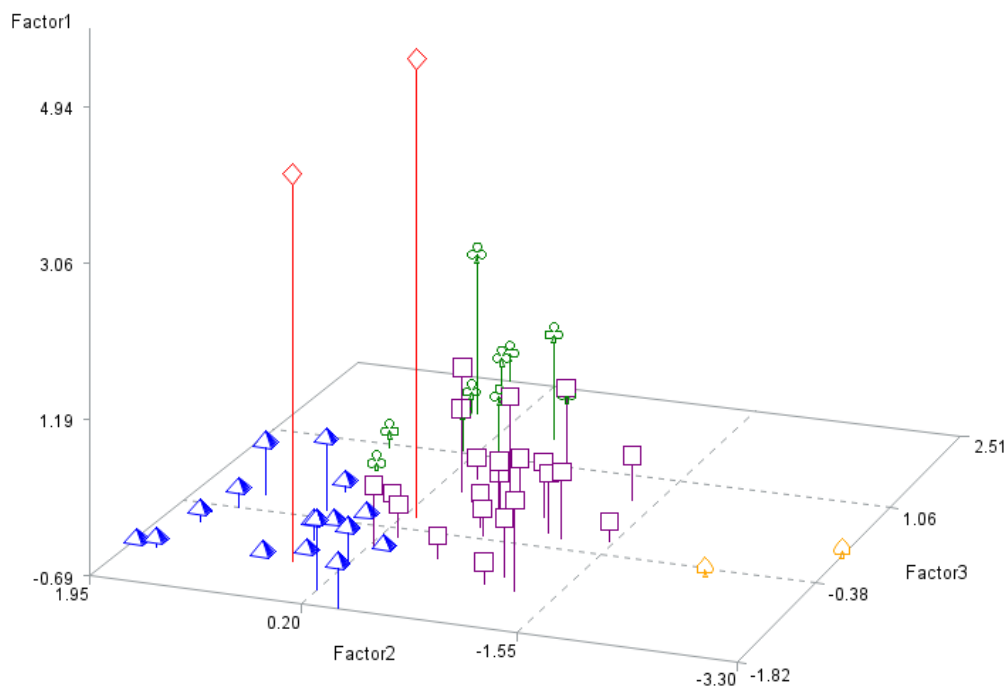
```
DATA cluster_no_jerar_edit;
set cluster_no_jerar;
length ClusterC $12. colorval $10. Shapeval $14. ;
label (data=cluster_no_jerar_edit, y=Poder_economico,
x=Eficacia_productiva,z=Poblacion_agricola, size=1.1,font=swissb,
pos=1, text=clusterC);
  if cluster=1then do;
    ClusterC="cluster1";
    shapeval="club";
    colorval="green";
  end;
  if cluster=2then do;
    ClusterC="cluster2";
    shapeval="diamond";
```

```

        colorval="red";
    end;
    if cluster=3 then do;
        ClusterC="cluster3";
        shapeval="spade";
        colorval="black";
    end;
    if cluster=4 then do;
        ClusterC="cluster4";
        shapeval="circlefill";
        colorval="blue";
    end;
    if cluster=5 then do;
        ClusterC="cluster5";
        shapeval="square";
        colorval="purple";
    end;
end;
run;

Proc g3d data=cluster_no_jerar_edit;
scatter Factor2*Factor3=Factor1/ color=colorval shape=shapeval;
run;

```



Para ver las características de cada cluster, ejecutaremos un box-plot para cada factor, siendo el código empleado:

```

PROC SGPLOT DATA = cluster_no_jerar;

```

```

VBOX Factor1
/ category = CLUSTER;

title 'Poder económico por cluster';
RUN;

PROC SGPLOT DATA = cluster_no_jerar;
VBOX Factor2
/ category = CLUSTER;

title 'Poblacion y trabajo por cluster';
RUN;

PROC SGPLOT DATA = cluster_no_jerar;
VBOX Factor2
/ category = CLUSTER;

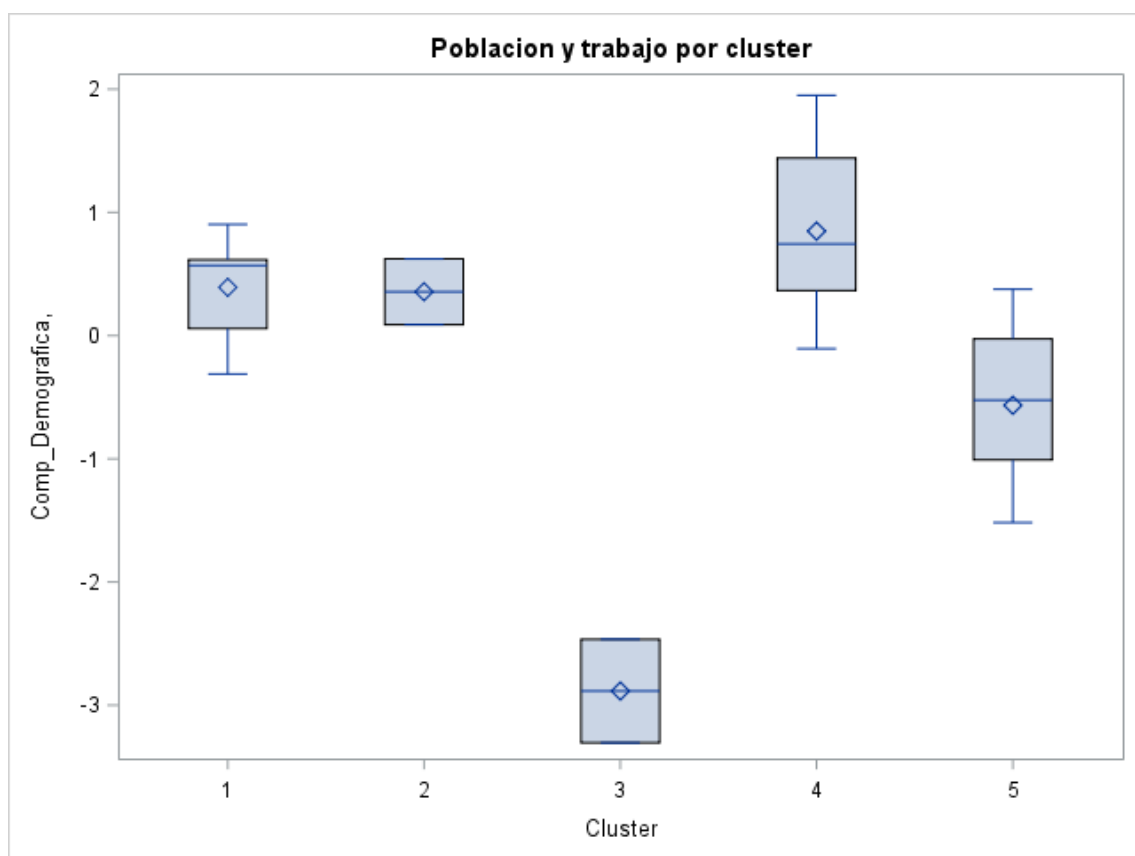
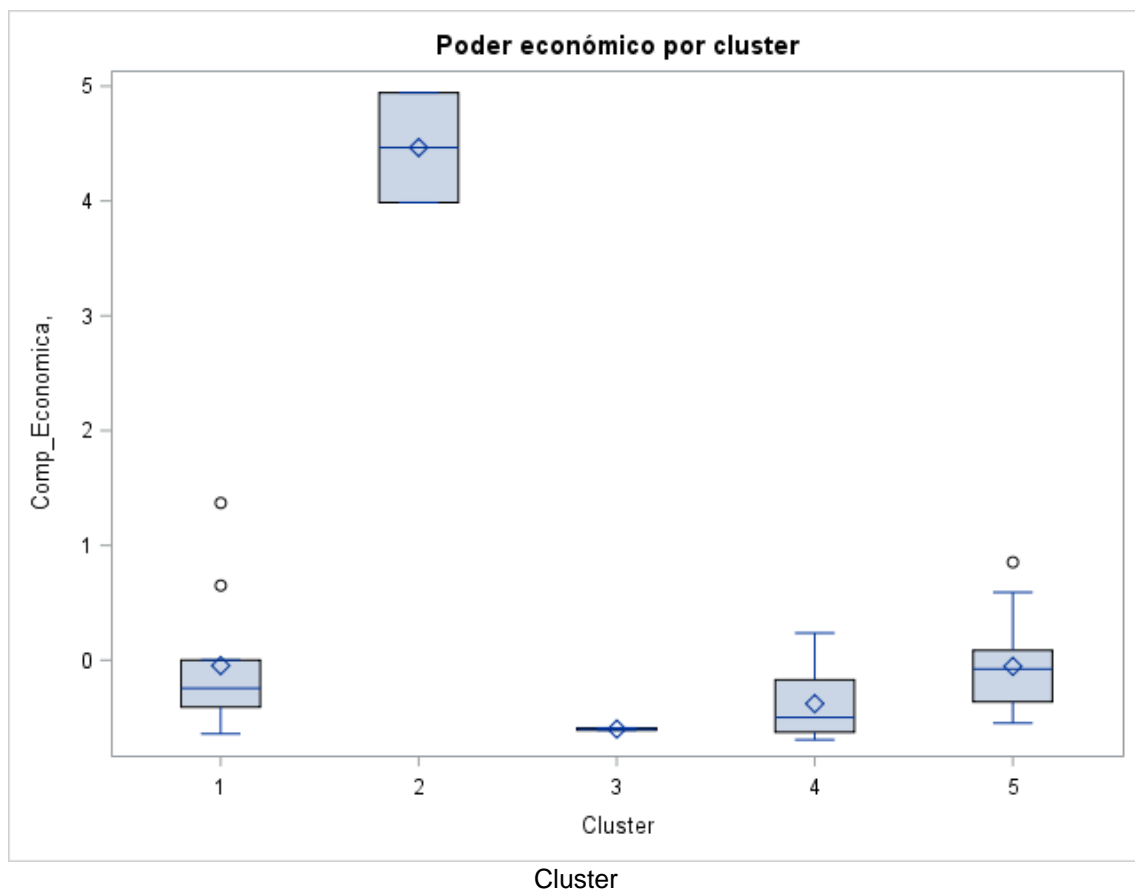
title 'Población agricola por cluster';
RUN;

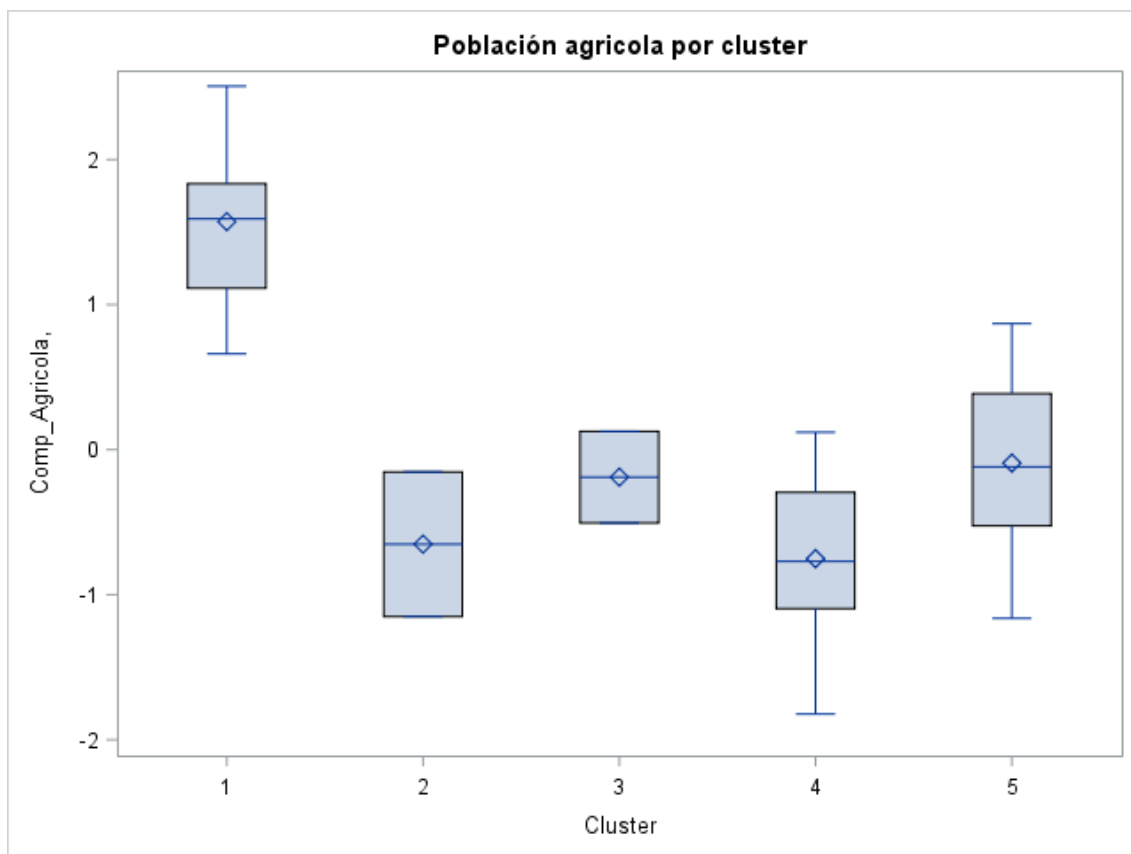
```

Antes de ver los resultados, daremos nombres y apellidos a los cluster, con el fin de identificarlos mejor, siendo esto:

1. **Cluster 1:** Badajoz, Ciudad Real, Cuenca, Cáceres, Córdoba, Granada, Jaén, Sevilla, Toledo y Valencia.
2. **Cluster 2:** Barcelona y Madrid.
3. **Cluster 3:** Ceuta y Melilla.
4. **Cluster 4:** Asturias, Bizkaia, Burgos, Cantabria, A Coruña, Gipuzkoa, León, Lugo, Ourense, Palencia, Salamanca, Segovia, Soria, Teruel, Zamora y Ávila.
5. **Cluster 5:** Albacete, Alicante, Almería, Álava, Islas Baleares, Castellón, Cádiz, Girona, Guadalajara, Huelva, Huesca, Lleida, Murcia, Málaga, Navarra, Las Palmas, Pontevedra, La Rioja, Santa Cruz de Tenerife, Tarragona, Valladolid y Zaragoza,

Dando como resultado los siguientes gráficos:





Para facilitar la comprensión de los datos, se ha optado por realizar las medias de cada cluster en función del factor. Con la ayuda de Excel (para generar el mapa de calor), se ha obtenido la siguiente tabla:

	Comp_economica	Comp_demográfica	Comp_agrícola
Cluster 1	-0,0479	0,3915	1,5722
Cluster 2	4,4651	0,3556	-0,6517
Cluster 3	-0,5998	-2,8847	-0,1895
Cluster 4	-0,3788	0,8485	-0,7512
Cluster 5	-0,0541	-0,5651	-0,0918

Cluster 1) Aquí se encuentran las provincias del sur de España, que no son costeras (salvo Valencia y Córdoba). Es obvio ver que en este cluster se encuentran concentradas las provincias que se centran más en la agricultura.

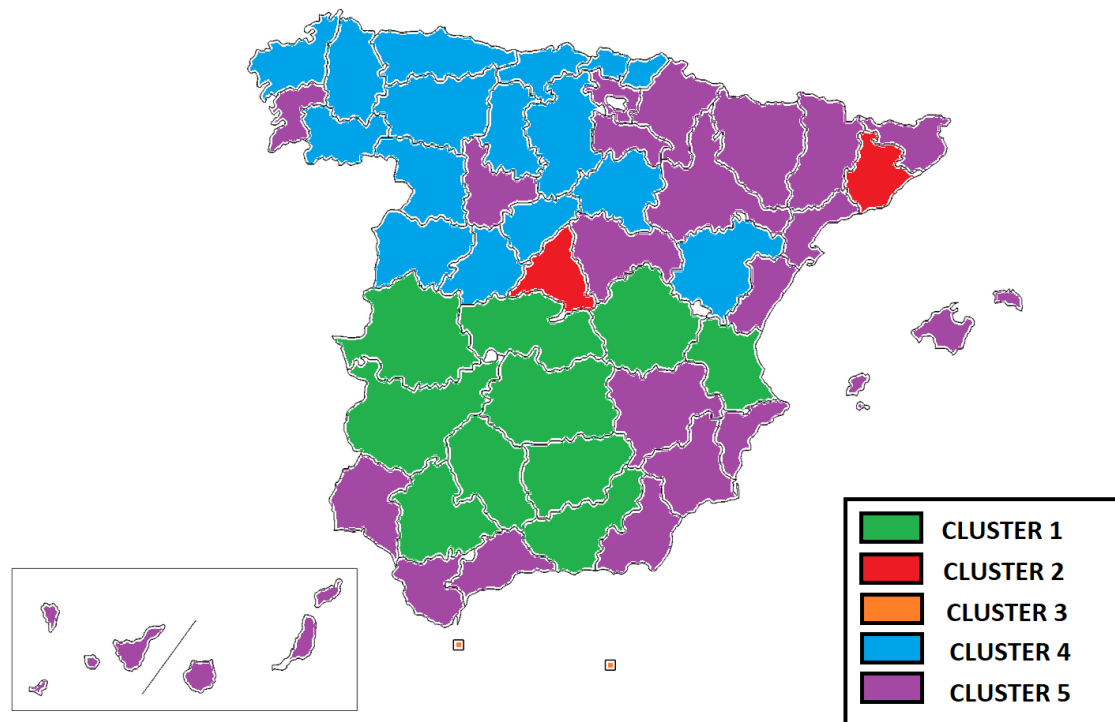
Clúster 2) Tiene un gran poder económico, ya que se trata de las 2 ciudades más importantes de España y donde se sitúan las centrales de muchas empresas. Sin embargo tiene una tasa de trabajo en relación a la población bastante normal en comparación con los demás clusters, probablemente por la cantidad de población joven que se encuentra en esas ciudades y que todavía no trabaja.

Clúster 3) En este clúster se encuentran Ceuta y Melilla, las cuales no tienen poder económico en comparación con los otros clusters ya que son 2 ciudades pequeñas que destacan principalmente por ser frontera con África.

Clúster 4) Este cluster se centra en las provincias del noroeste de España y destaca por su alta componente demográfica (buena tasa de trabajo en relación con las personas totales), debido probablemente a que es una zona de fábricas y tejidos industriales.


Clúster 5) En este cluster se agrupan principalmente las provincias con una economía basada en el turismo, el cual se caracteriza por empleo temporal y depender de la época del año (posible causa para un mal ratio de población/empleo) . Además de todo lo anterior se junta que no tiene una economía fuerte y que casi no tienen bases agrícolas.


Para tener una visión más amplia de los clústeres, se ha optado por emplear un mapa con las provincias coloreadas en función del cluster al que pertenece, siendo el resultado:





5. Inspeccionar hasta qué punto son capaces de discriminar tales clusters las puntuaciones factoriales.


Realizando un análisis discriminante en spss con el fin de contrastar que los clusters sean agrupados de forma correcta y nos fijaremos en la matriz resultante de los valores clasificados según este método. Para ello, seleccionaremos las siguientes opciones:


 **Análisis discriminante** ✕


 Población Total [...]


 Población Total (...)


 Mortalidad [T_M]


 Tasa Bruta de Na...


 IPC [IPC]


 Número de empr...


 Construcción (nº ...)

 Comercio, transp...

 Información y co...

 Actividades finan...

 Actividades profe...

 Educación, sani...

➡

Variable de agrupación:

CLUSTER(1 5)

[Definir rango...](#)

[Estadísticos...](#)


[Método...](#)


[Clasificar...](#)


[Guardar...](#)

➡

Independientes:

 Comp_Economica, [Factor1]

 Comp_Demografica, [Factor2]

 Comp_Agricola, [Factor3]

☐ Introducir independientes juntos

☒ Usar método de inclusión por pasos

➡

Variable de selección:

[Valor...](#)


[Aceptar](#)

[Pegar](#)

[Restablecer](#)

[Cancelar](#)

[Ayuda](#)

 **Análisis discriminante: Clasificación** ✕

Probabilidades previas

☐ Todos los grupos iguales

☒ Calcular según tamaños de grupos

Usar matriz de covarianzas

☒ Intra-grupos

☐ Grupos separados

Visualización

☐ Resultados para cada caso

☐ Limitar los casos a los primeros:

☒ Tabla de resumen

☐ Clasificación dejando uno fuera

Gráficos

☐ Grupos combinados

☐ Grupos separados

☐ Mapa territorial

☐ Reemplazar los valores perdidos con la media

[Continuar](#)

[Cancelar](#)

[Ayuda](#)

Obteniendo de esta forma la siguiente tabla:

Resultados de clasificación^a

		Pertenencia a grupos pronosticada						
	Cluster	1	2	3	4	5	Total	
Original	Recuento	1	10	0	0	0	0	10
		2	0	2	0	0	0	2
		3	0	0	2	0	0	2
		4	0	0	0	16	0	16
		5	1	0	0	1	20	22
	%	1	100,0	,0	,0	,0	,0	100,0
		2	,0	100,0	,0	,0	,0	100,0
		3	,0	,0	100,0	,0	,0	100,0
		4	,0	,0	,0	100,0	,0	100,0
		5	4,5	,0	,0	4,5	90,9	100,0

a. 96,2% de casos agrupados originales clasificados correctamente.

Como podemos observar, el análisis discriminante clasifica de forma correcta a todos los clusters, salva la excepción de una provincia, la cual fue categorizada mal.