ANÁLISIS FACTORIAL DE LAS VARIABLES SOCIOECONÓMICAS SOBRE LAS COMUNIDADES AUTÓNOMAS DE ESPAÑA

MINERÍA DE DATOS EN NEGOCIOS

Ágatha del Olmo Tirado | 2ºBIA | 13/12/2023

VNIVERSITAT DE VALÈNCIA

INTELIGENCIA Y ANALÍTICA DE NEGOCIOS

ÍNDICE

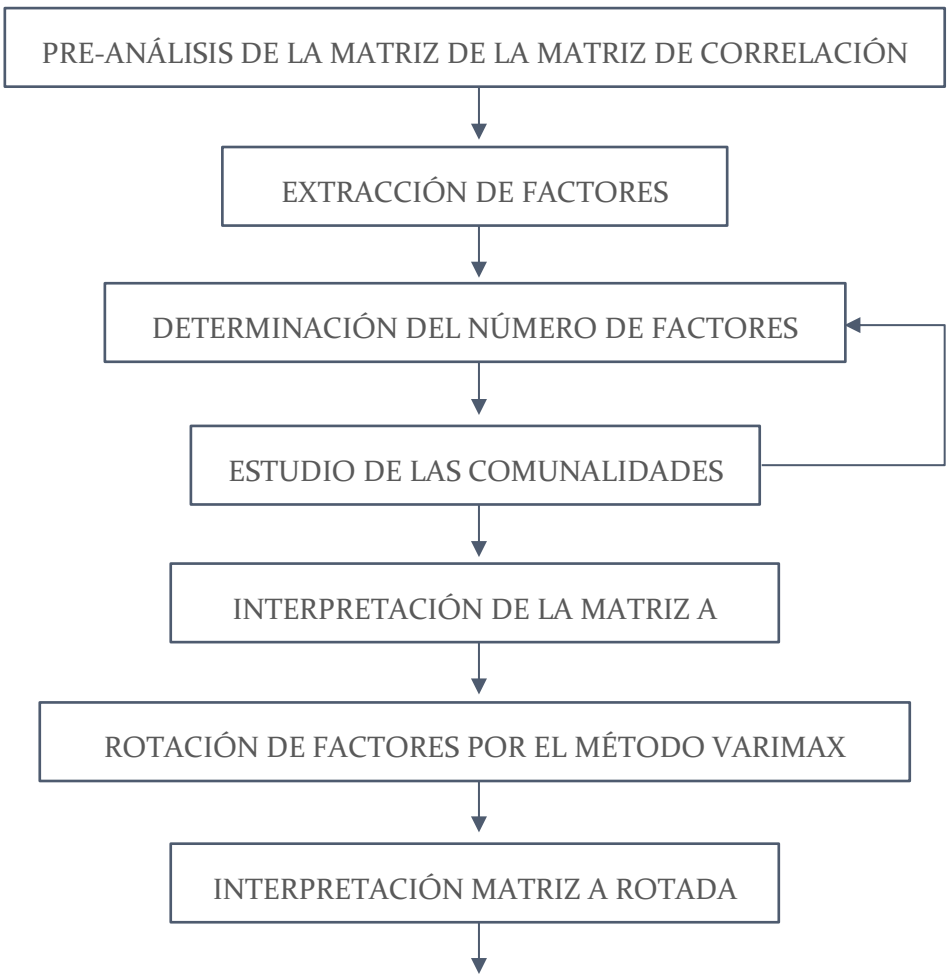
1. INTRODUCCIÓN

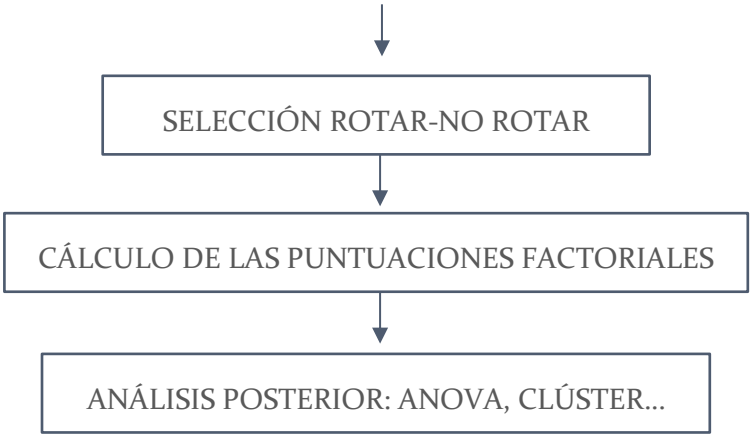
Este informe presenta un análisis factorial detallado basado en variables socioeconómicas de las Comunidades Autónomas españolas recopiladas en el archivo "comaut.sav". El objetivo de este estudio es emplear técnicas de factorización para tratar de explicar la mayor parte de las variables originales con un número menor de factores. Durante todo el estudio hemos usado el entorno RStudio de R. Los métodos usados han sido: Componentes Principales, rotación Varimax, Máxima verosimilitud y ejes principales.

Adicionalmente, hemos querido aprovechar el análisis factorial escogido para llevar a cabo una agrupación jerárquica Ward. De esta forma, crearemos clusters de Comunidades Autónomas homogeneos intragrupalmente y heterogeneos intergrupalmente explicados por los factores.

El criterio general de selección del método no se centrará únicamente en reducir el número de variables, sino también, por un lado, en que los factores expliquen la mayor parte de las variables originales, y por otro, en que los factores sean fácilmente interpretables.

El esquema que vamos a seguir es el siguiente:





2. ENSAYO DE OBTENCIÓN DE COMPONENTES PRINCIPALES

Uno de los métodos que se usan comúnmente para extraer factores explicativos de las variables es el análisis de componentes principales que se basa en el Teorema Espectral. Este enfoque implica la rotación de los ejes originales en un nuevo sistema de referencia para las variables: las componentes principales. Generalmente se obtienen tantas componentes como variables, explicando la totalidad de las originales. Pero si el objeto del estudio es el de factorizar, podemos seleccionar solo algunas componentes, aunque esto implique perder un pequeño porcentaje de explicación.

2.1. PREPROCESO DE COMPONENTES PRINCIPALES

Para iniciar nuestro análisis, hemos cargado las librerías “haven” y “corrplot” para su posterior uso, y hemos importado el archivo “comaut.sav” como “datos”.

Es crucial para el estudio que todas las variables del dataframe sean numéricas. Por lo tanto, hemos guardado la variable “Cautonoma” (que contiene el nombre de las CCAA, y por ende es tipo carácter) como un objeto en el environment y la hemos eliminado de “datos” para añadirla como nombres de las filas con *row.names(datos)<-Cautonoma*. Después, hemos usado la instrucción *as.data.frame(lapply(datos, as.numeric))* para asegurarnos de que el resto de las variables son, efectivamente, numéricas.

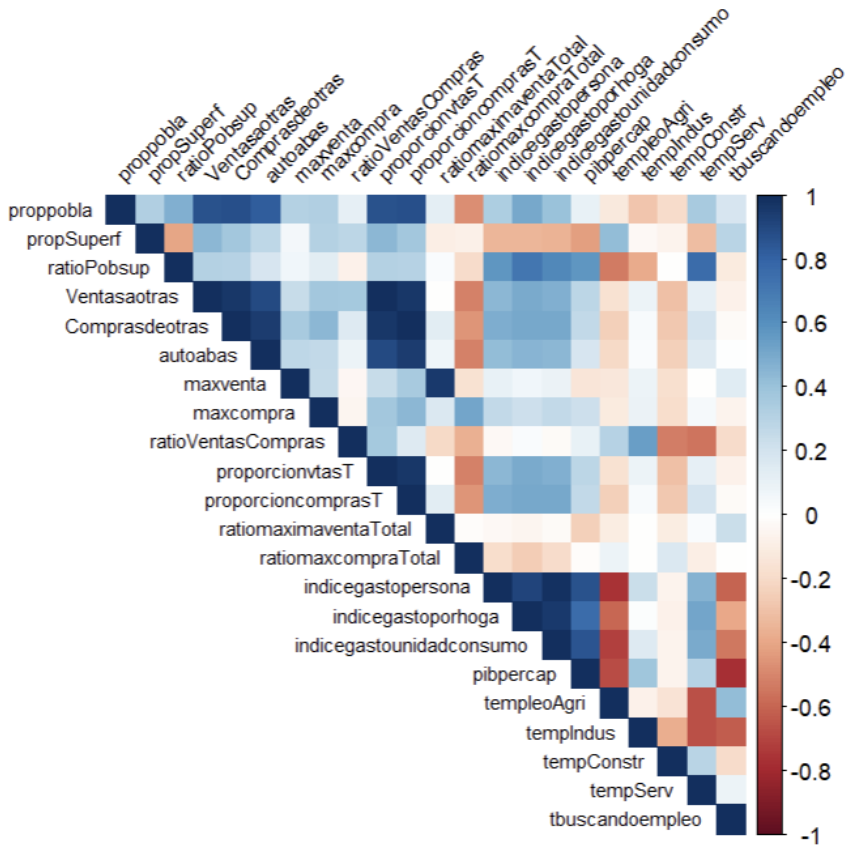
Revisando el significado de las variables, identificamos que “densidad”, “Gastoporunidadconsumo”, “Gastoporhogar” y “Gastoporpersona” explican lo mismo que “ratiopobSup”, “indicegastounidadconsumo”, “indicegastoporhoga” y “indicegastopersona”, respectivamente, a pesar de tener matices distintos en cuanto a tamaño. Por lo tanto, eliminamos las cuatro primeras ya que preferimos tener índices y ratios.

Posteriormente, hemos decidido relativizar las variables absolutas (excepto aquellas que tienen un significado importante para el estudio) para evitar el efecto del tamaño de las Comunidades Autónomas en la correlación. Para hacerlo, hemos dividido cada

valor entre la suma de todos sus valores. Las variables afectadas han sido “autoabas”, “Ventasaotras”, “Comprasdeotras”, “maxventa” y “maxcompra”.

Ya acabado el preproceso, podemos empezar a hacernos una idea de cómo se van a comportar los factores. Hemos creado una matriz de correlaciones con el código *cor(datos, use = "pairwise.complete.obs")*, y después lo hemos visualizado con *corrplot(matriz_correlaciones)*.

Buscamos “constelaciones”, grupos de correlaciones muy altas positiva o negativamente, que después se convertirán en aquello que cada uno de los factores explique. En nuestro caso la matriz se ve de la siguiente forma:



En nuestra matriz podemos detectar sobre todo correlaciones fuertes en un sentido económico con el gasto, compras y ventas. Sin embargo, no es muy sencillo ver más constelaciones con estas variables.

Nos interesan altas correlaciones para que se formen factores interesantes. Para comprobar fácilmente si las hay podemos calcular el determinante de la matriz de correlaciones, que nos ha de dar cercano a 0, y en nuestro caso efectivamente da -1.703072e-107.

Finalmente, usando la instrucción eigen(cor(datos)) obtenemos tanto los valores propios como los vectores propios de la matriz de correlación, lo que nos permite avanzar hacia la fase de extracción de factores.

2.2. NÚMERO DE FACTORES A CONSIDERAR

Para obtener todas las Componentes Principales empleamos la función *prcomp(datos, scale.=TRUE)*. La información principal la obtenemos mediante un simple summary(), que nos proporciona la desviación típica, la proporción de varianza y la proporción acumulada de cada Componente principal, ordenadas de más explicativas a menos explicativas. Estas las veremos a fondo cuando hayamos elegido el número de factores.

A partir de la desviación típica proporcionada por *prcomp()*, podemos extraer los valores propios (deltas) de cada Componente principal. El output, traducido desde la notación científica, es el siguiente:

[1] 8.161822 4.478076 3.96889 2.0311111.610429 1.09400 0.6125145 0.4236008
[9] 0.2816226 0.136107 0.07642417 0.05936747 0.04349551 0.01420612
[15] 0.005645522 0.002677848 2.393729e-31

De aquí podemos extraer la matriz T o matriz patrón (variables x factores) con el atributo *rotation*. Los coeficientes de esta matriz son los pesos factoriales (aij), que miden la relación funcional lineal entre variables y factores. Lo que nos interesa es que cada variable cargue alto en un factor y bajo en el resto, es decir, queremos que cada variable vaya mayoritariamente solo a un factor para poder formar factores interesantes. Es importante destacar que aún no hemos tipificado, y podemos verificarlo calculando su desviación típica, que todavía es diferente a 1. Podríamos tipificar multiplicando por $D^{1/2}$, pero es más sencillo tipificar directamente las puntuaciones.

Ahora que ya tenemos las Componentes principales y los valores propios, podemos escoger el número de factores. Los criterios comúnmente utilizados para seleccionar factores son varios, entre ellos:

- $\boldsymbol{\lambda_{ij} > 1}$: Dado que las variables originales ya explican un 1, esperamos que los factores, como mínimo, expliquen lo mismo que las originales.
- $\frac{\sum_1^p \lambda_{ij}}{n} \geq \boldsymbol{x}$: Buscamos que se explique un porcentaje mínimo x de las variables originales.

Como afirma el No Free Lunch Theorem, no hay criterio perfecto para cualquier situación. Con un total de 22 variables, exigir que los factores expliquen al menos lo que ya explicaban las originales puede resultar poco restrictivo, llevándonos a seleccionar muchos factores, lo cual complicaría su interpretación. En nuestro caso, escogeríamos 5 factores.

El segundo criterio puede ser más interesante, ya que nos permite ajustar el nivel de restricción según nuestras necesidades, pero a menudo se usa una explicación conjunta mínima del %. Realizamos los cálculos dividiendo la suma de los valores propios de los factores que nos interesen entre el número de variables originales, y descubrimos que con 4 factores ya superamos el 80% (se explica un 80,18137%), así que escogemos usar 4 factores.

Ya escogido el número de factores podemos crear el modelo con *prcomp(datos, scale. = TRUE, rank=4)*, que ahora no contiene tantos factores como variables sino los 4 que hemos seleccionado. Con el modelo ya podemos obtener la matriz A o matriz de correlaciones entre variables y factores con *cor(datos,mod2$x)*.

2.3. DISCUSIÓN DE LA CONVENIENCIA DE ELIMINAR O GENERAR VARIABLES

Para saber si podemos eliminar variables calculamos las comunalidades (h_j^2) como suma de los pesos factoriales (cada fila de la matriz A -variables-) al cuadrado. De esta forma vemos la proporción de la varianza explicada por los factores comunes en una variable. Si hay alguna variable que no resulta muy explicada por la totalidad de los factores (la comunalidad no es cercana a 1), lo podremos ver de esta forma. El output es el siguiente:

[1] 0.9173263 0.6379197 0.6990732 0.9824372 0.9664517 0.8483361
[7] 0.8777775 0.3339168 0.6871367 0.9824372 0.9664517 0.8943656
[13] 0.3781153 0.9506716 0.8646729 0.9476765 0.9461459

Como vemos, las número 8 ("maxcompra") y 13 ("ratiomaxcompratotal") son muy bajas (<0.6), así que las analizamos. Vemos que sí tienen importancia en el estudio, pues nos generan nuevas perspectivas para el estudio (es un error pensar que son las contrarias a "maxventa" y "ratiomaxcompratotal" y que por lo tanto generan multicolinealidad). Por lo tanto, la solución es aumentar el número de factores a 5.

Ajustamos los códigos necesarios, en este caso el modelo *prcomp(datos, scale. = TRUE, rank=5)* y la matriz A *cor(datos,mod3$x)*. Y volvemos a calcular las comunalidades, que esta vez nos dan lo siguiente:

[1] 0.9173367 0.7810601 0.7150379 0.9853488 0.9750246 0.8485447
[7] 0.9567742 0.9227151 0.7526849 0.9853488 0.9750246 0.9800231
[13] 0.9265753 0.9508031 0.8669986 0.9476814 0.9550261

Ahora todas las comunalidades están cercanas a 1, es decir, todas las variables se ven altamente explicadas por el conjunto de factores elegidos. Por lo tanto, el número de factores que vamos a utilizar en este estudio serán finalmente 5.

2.4. INTERPRETACIÓN DE LA SOLUCIÓN FACTORIAL

Habiendo determinado el número óptimo de factores, procedemos a estudiar la solución factorial. La solución factorial es el conjunto de factores que hemos escogido y su significado. El output del *summary()* del modelo con 5 factores es el siguiente:

Importance of first k=5 (out of 17) components:

	PC1	PC2	PC3	PC4	PC5
Standard deviation	2.857	2.1161	1.7230	1.42517	1.2690
Proportion of Variance	0.371	0.2036	0.1349	0.09232	0.07083
Cumulative Proportion	0.371	0.5745	0.7095	0.80181	0.8750

La primera fila indica la desviación estándar de cada componente principal, es decir, qué tan dispersos están los datos alrededor de la media. La segunda fila muestra qué proporción de varianza explica cada componente principal con respecto a las variables originales. La tercera fila es la proporción acumulativa de varianza, y en este caso el conjunto escogido de factores explica el 87,035% de la variabilidad en los datos originales.

Es interesante comentar que la tendencia decreciente en las desviaciones estándar y las proporciones de varianza se debe al criterio para ordenar las componentes: la primera captura la mayor cantidad posible de variabilidad en los datos, la segunda la mayor cantidad posible restante, y así sucesivamente.

Con la matriz A que antes hemos obtenido y modificado para 5 factores, podemos estudiar el significado de cada factor con la matriz de correlaciones (redondeada a 3 decimales):

VARIABLE	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4	FACTOR 5
Proppobla	0.806	-0.391	0.308	0.138	0.03
Propsuperf	0.056	-0.779	-0.032	0.164	-0.378
ratioPobsup	0.593	0.462	0.343	0.126	0.126
Ventasaotras	0.876	-0.439	-0.074	0.129	-0.054
Comprasdeotras	0.896	-0.401	0.057	-0.012	-0.093
Autoabas	0.823	-0.403	0.064	0.068	0.014
Maxventa	0.281	-0.283	0.25	-0.81	0.281
Maxcompra	0.386	-0.128	-0.003	-0.41	-0.767
Ratioventascompras	0.142	-0.389	-0.701	0.155	0.256
proporcionvtasT	0.876	-0.439	-0.736	0.13	-0.054
porporcioncomprasT	0.896	-0.401	0.057	-0.012	-0.093
ratiomaxventaTotal	0.069	-0.176	0.354	-0.857	0.293
ratiomaxcompraTotal	-0.444	0.275	0.019	-0.324	-0.741
indicegastopersona	0.788	0.539	-0.172	-0.096	-0.011
indicegastoporhoga	0.811	0.451	-0.024	0.051	0.048
indicegastudconsumo	0.812	0.521	-0.112	-0.057	0.002
pibpercap	0.601	0.65	-0.404	0.003	-0.094
templeoAgri	-0.533	-0.645	-0.112	0.17	-0.043
tempIndus	-0.067	-0.01	-0.906	-0.354	0.058
tempConstr	-0.229	0.365	0.362	0.196	-0.197
tempServ	0.415	0.508	0.691	0.18	-0.005
tbuscandoempleo	0.273	-0.535	0.639	0.058	0.074

Las correlaciones que hemos considerado altas han sido las mayores a 0.7. Hemos pintado las altas positivas de verde y las altas negativas de rojo. Las interpretaciones son las siguientes:

FACTOR 1: CCAA con mucha población y una potencia económica tanto interna como externa, lo que permite que su población tenga una capacidad económica más alta (o que el precio de vida es más alto).

FACTOR 2: CCAA de tamaño pequeño.

FACTOR 3: CCAA con poca potencia de compra-venta e industria.

FACTOR 4: CCAA con poco volumen de venta máximo a otra CCAA.

FACTOR 5: CCAA con poco volumen de compra máximo a otra CCAA.

2.5. ENSAYO DE LA SOLUCIÓN ROTADA VARIMAX

Hay dos tipos de rotaciones: ortogonales y oblicuas. La ventaja principal de las rotaciones ortogonales es la incorrelación, pues en las oblicuas puede darse que dos o más factores expliquen a la vez lo mismo, mientras que en las ortogonales se consigue lo contrario. Entre las rotaciones ortogonales se encuentra la rotación Varimax (Varianza Máxima), que trata de añadir varianza a los factores, de forma que haya unas pocas saturaciones altas y muchas casi nulas en las variables, de forma que su interpretación de las correlaciones suele ser más clara.

Para calcular la solución rotada usamos el código *varimax(mod2$rotation)*, y necesitamos las puntuaciones, que tipificamos con *scale()*. Para interpretarla necesitamos la matriz A, que es la matriz de correlaciones entre los factores y las variables originales. El output redondeado a 3 decimales es el siguiente:

VARIABLE	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4	FACTOR 5
Proppobla	0.901	0.103	0.246	-0.103	0.109
Propsuperf	0.563	-0.575	-0.173	0.182	-0.266
ratioPobsup	0.232	0.611	0.501	-0.039	0.189
Ventasaotras	0.964	0.202	-0.109	0.02	0.054
Comprasdeotras	0.952	0.225	-0.008	-0.131	-0.035
Autoabas	0.895	0.173	0.001	-0.095	0.091
Maxventa	0.25	0.001	-0.068	-0.943	-0.033
Maxcompra	0.41	0.16	-0.035	-0.136	-0.842
Ratioventascompras	0.271	-0.045	-0.738	0.198	0.304
proporcionvtasT	0.964	0.202	-0.109	0.016	0.054
porporcioncomprasT	0.952	0.225	0.008	-0.131	-0.035
ratiomaxventaTotal	0.021	-0.066	0.034	-0.986	-0.052
ratiomaxcompraTotal	-0.466	-0.055	0.075	0.052	-0.836
indicegastopersona	0.281	0.933	0.024	-0.032	-0.02
indicegastoporhoga	0.378	0.829	0.164	0.016	0.096

indicegastudconsumo	0.32	0.915	0.082	-0.026	0.01
pibpercap	0.075	0.944	-0.133	0.187	-0.072
templeoAgri	-0.025	-0.799	-0.276	0.166	0.012
tempIndus	-0.077	0.296	-0.922	-0.055	-0.089
tempConstr	-0.309	0.018	0.495	0.195	-0.121
tempServ	0.117	0.443	0.85	-0.03	0.083
tbuscandoempleo	0.147	-0.725	0.418	-0.216	0.094

Las correlaciones que hemos considerado altas han sido las mayores a 0.7. Hemos pintado las altas positivas de verde y las altas negativas de rojo. Las interpretaciones de cada factor son las siguientes:

FACTOR 1: CCAA con mucha población y potencia económica tanto interna como externa.

FACTOR 2: CCAA con gastos altos de la población y poca potencia agrícola.

FACTOR 3: CCAA con poca potencia de compra-venta y poca industria pero mucha tasa de empleo de servicios.

FACTOR 4: CCAA con poco volumen de venta máximo a otra CCAA.

FACTOR 5: CCAA con poco volumen de compra máximo a otra CCAA.

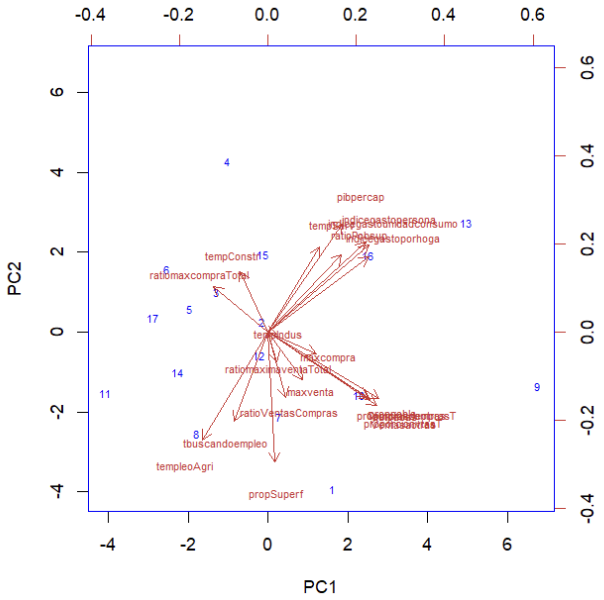
2.6. COMPARACIÓN DE LOS ENSAYOS PREVIOS

Comparamos la solución factorial rotada y sin rotar. A pesar de que la solución rotada sea más compleja, y que se podría decir que no cumple el principio de Ockham ya que las dos soluciones factoriales no dejan de ser las mismas, consideramos que la solución rotada es la más fácilmente interpretable. Esto lo vemos especialmente en el segundo factor, ya que, con el objeto socioeconómico del estudio, que un factor explique únicamente el tamaño de las CCAA no nos interesa, mientras que el gasto de las población y su potencia en un sector concreto es mucho más trascendental para el trabajo.

Podemos ver gráficamente las variables originales y los individuos en el plano de las dos primeras componentes principales (que explican aproximadamente la mitad de la variabilidad) con la instrucción biplot(x=modelo).

Cada vector rojo representa una variable original de la base de datos. Su dirección indica cómo se correlaciona con los PC1 y PC2 (los que apuntan a la misma dirección están altamente correlacionados, los que apuntan a la dirección contraria están inversamente correlacionados, pero si son perpendiculares están incorrelacionados), y su longitud indica la fuerza de esa correlación (cuanto más larga más intensa). Cada punto azul es un individuo o CCAA de la base de datos, y su ubicación en el plano se determina por sus puntuaciones en la solución factorial.

Con las puntuaciones de la solución factorial podremos, posteriormente, llevar a cabo la clusterización de las CCAA a partir de los factores.



3. ENSAYO ALTERNATIVO POR MÁXIMA VEROSIMILITUD

El análisis factorial por Máxima Verosimilitud o Maximum Likelihood Factor Analysis en inglés, es un enfoque estadístico usado para estimar los parámetros de un modelo de análisis factorial. La principal diferencia con el análisis de Componentes Principales es que a través de máxima verosimilitud la interpretación es mucho más sencilla porque trata de encontrar factores no observados o "latentes" que explican las relaciones entre las variables originales.

Ahora, la función es factanal(), y al usarse Máxima Verosimilitud, en los cálculos que realiza R en el background al ejecutar la función se divide entre el determinante de matriz de correlaciones A, de forma que si hay multicolinealidad se divide entre o y no converge. Es por esto que debemos seleccionar un número menor de variables para que la correlación no sea tan alta. En nuestro caso hemos optado por las siguientes: 'proppobla', 'ratioPobsup', 'Ventasaotras', 'Comprasdeotras', 'autoabas', 'maxventa', 'maxcompra', 'ratioVentasCompras', 'proporcionvtasT', 'ratiomaxcompraTotal', 'indicegastoporhoga', 'pibpercap', 'tempIndus', 'tempServ' y 'tbuscandoempleo'. Un total de 15 variables en comparación con las 22 con las que nos quedamos en el anterior análisis factorial.

3.1. NÚMERO DE FACTORES A CONSIDERAR

La diferencia con el método de PCA, en el método de máxima verosimilitud no puedes saber cuál es el número de factores que te interesa usar de forma previa, así que puedes incluir todos los factores posibles con la función *factanal()* (en nuestro caso nos permite hasta 9) para ver cuál es la explicación acumulada según el número de factores y seguir el criterio que seguimos anteriormente:

- $\frac{\sum_1^p \lambda_{ij}}{n} \geq \boldsymbol{x}$: Buscamos que se explique un porcentaje mínimo x de las variables originales.

Siendo este porcentaje mínimo el mismo que el anteriormente usado, un 80%, debemos usar un total de 5 factores, que explican un 87,5% de la varianza, mientras que con 4 se explica un 78,6%.

3.2. DISCUSIÓN DE LA CONVENIENCIA DE ELIMINAR O GENERAR VARIABLES

Podemos extraer las comunalidades restando 1 a las unicidades ya que comunalidades+unicidades=1.

Las comunalidades, *1-fa$uniquenesses*, redondeadas a tres decimales son las siguientes:

0.862, 0.858, 0.995, 0.99, 0.957, 0.424, 0.995, 0.992, 0.995, 0.865, 0.724, 0.995, 0.948, 0.907, 0.724

Y atendiendo de nuevo a los criterios anteriores, las menores a 0.6 se consideran bajas. Por lo tanto vemos que la variable número 4, que es "maxventa", es muy baja (0.424), así que eliminamos del *factanal()* "maxventa".

Ahora, con los "starting values" de la función por defecto no converge. Podríamos hacer un "start" desde las cargas factoriales de la anterior solución factorial, pero habría que crear de cero una matriz con las cargas concretas de las variables que hemos usado, que como no son el total no podemos usar tal cual como start=mod3$loadings, pues esa matriz tiene 22 filas y no 14. Para ahorrarnos este proceso podemos crear una matriz con valores aleatorios hasta que funcione de nuevo la función. Para que haya reproducibilidad de los datos tenemos que usar una seed en la que converjan los datos, que en este caso es la 3. De esta forma conseguimos que funcione, y volvemos a calcular las comunalidades restando 1 a las unicidades. Esta vez son las siguientes:

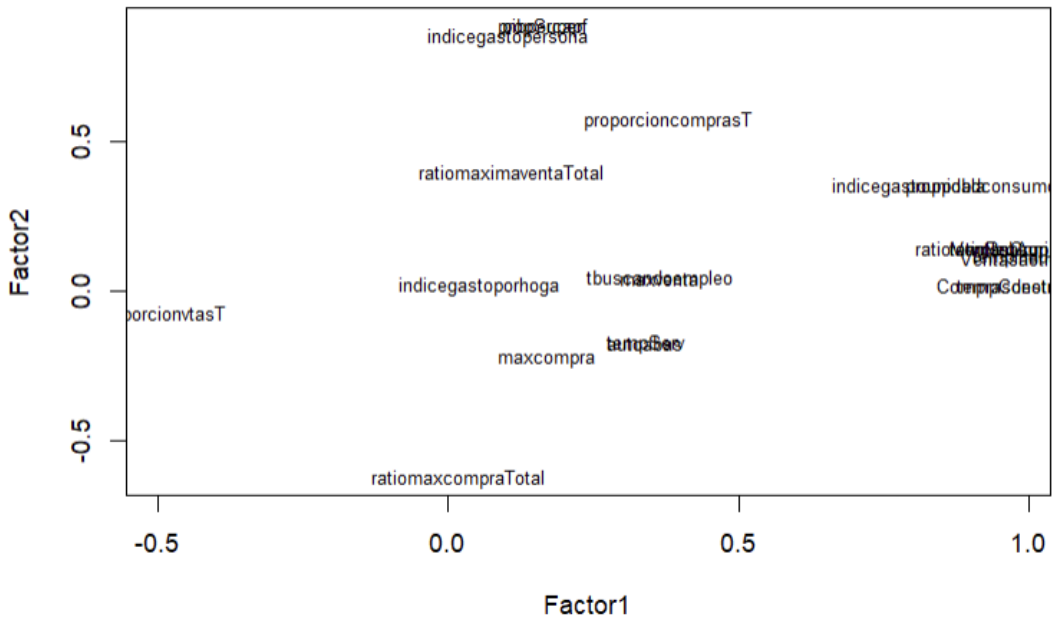
0.143, 0.178, 0.995, 0.312, 0.114, 0.737, 0.778, 0.995, 0.89, 0.907, 0.778, 0.218, 0.411, 0.97

Como vemos, muchas comunalidades bajan y a valores mucho más bajos que 0.424, que era la de "maxventa", de forma que volvemos a la solución factorial anterior, pues

empeora mucho la que obtenemos sin esta variable ya que el punto de inicio aleatorio es mucho peor al "default".

3.3. INTERPRETACIÓN DE LA SOLUCIÓN FACTORIAL

Para interpretar la solución factorial por máxima verosimilitud, tenemos que extraer las cargas factoriales que se corresponden con *fa$loadings()*. El gráfico que representa la relación entre las variables seleccionadas y los dos primeros factores que explican algo más de la mitad de la variabilidad total es el siguiente:



Y los valores concretos de las cargas factoriales redondeados a 3 decimales son los siguientes:

VARIABLE	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4	FACTOR 5
Proppobla	0.857	0.352	-0.145	0.007	0.074
ratioPobsup	0.162	0.888	0.215	0.014	0.061
Ventasaotras	0.952	0.141	0.117	0.012	0.239
Comprasdeotras	0.979	0.116	0.112	0.08	0.004
Autoabas	0.961	0.014	0.092	-0.104	-0.119
Maxventa	0.339	-0.17	-0.126	0.158	-0.117
Maxcompra	0.364	0.046	0.096	0.922	0.006
Ratioventascompras	0.17	-0.226	0.164	-0.138	0.93
proporcionvtasT	0.952	0.141	0.117	0.012	0.239
ratiomaxcompraTotal	-0.496	-0.074	0.014	0.761	-0.208
indicegastoporhoga	0.381	0.572	0.54	0.008	0
pibpercap	0.111	0.402	0.9	0.089	0.049
tempIndus	0.018	-0.623	0.679	0.055	0.315
tempServ	0.104	0.854	-0.024	-0.032	-0.409
tbuscandoempleo	0.052	0.019	-0.868	-0.001	-0.066

Las cargas factoriales que hemos considerado altas han sido las mayores a 0.7. Hemos pintado las altas positivas de verde y las altas negativas de rojo. Las interpretaciones de cada factor son las siguientes:

FACTOR 1: CCAA grandes con una gran importancia económica en compra-venta y en autoabastecimiento.

FACTOR 2: CCAA con mucha población y basadas en el sector servicios.

FACTOR 3: CCAA con alto PIB y baja tasa de búsqueda de empleo.

FACTOR 4: CCAA con altos índices de compra en cantidad a otras CCAA.

FACTOR 5: CCAA con altos índices de transacciones compra-venta a otras CCAA.

3.4. ENSAYO DE LA SOLUCIÓN ROTADA VARIMAX

La solución que obtenemos con la rotación Varimax es exactamente la misma que obtenemos sin rotar. Basándonos en el principio de la navaja de Ockham, en condiciones iguales el método sencillo siempre es el mejor, nos quedaríamos con la solución sin rotar.

3.5. COMPARACIÓN DE LOS ENSAYOS PREVIOS

Ahora elegimos entre la solución factorial rotada por Componentes principales y la solución sin rotar por Máxima Verosimilitud. Comparamos primero la capacidad explicativa de ambos, y los dos explican un 87,5% de la variabilidad con 5 factores. Así pues, comparamos la sencillez de interpretación: el hecho de que en Máxima Verosimilitud el factor 5 explique las transacciones compra-venta mientras que el 4 solo la cantidad de compra me parece poco claro comparado con los de Componentes principales que son compra por un lado y venta por el otro, así que escogemos de nuevo las Componentes principales rotadas.

4. ENSAYO DEL MÉTODO DE EJES PRINCIPALES

El análisis factorial de ejes principales es un método iterativo basado en la extracción sucesiva de los factores que explican la mayor parte de la varianza común. Esto se consigue a través de la estimación de la matriz A por el método de componentes principales. Su principal ventaja, según Winter y Dodou (2012), es su capacidad de recuperar factores débiles, y es recomendable sobre todo en análisis factoriales para pequeñas muestras (aunque son más propensas a no converger) en las que hay poca

cantidad de variables y con correlaciones moderadas, y se suele utilizar cuando el método de máxima verosimilitud falla al converger.

Es interesante el apunte de que las iteraciones no intentan buscar la mejor solución, sino conbvberger de ka forma más rápida descomponiendo sucesivamente los valores eigen.

Esta vez, la función es *fa()* del paquete "pysch", y debemos añadir el atributo *fm = "pa"*, porque puede usar distintos métodos de análisis factorial ("factoring methods"), y el "pa" se refiere al "principal factor solution". Además, como hemos comentado, es un método iterativo y por eso mismo se ha de especificar el número de iteraciones máximas (si no se da antes un criterio de parada), que hemos establecido en 50. Por último debemos especificar el tipo de rotación, ya que por defecto es "oblimin", y por ahora establecemos "none" para que no use ninguna.

4.1. NÚMERO DE FACTORES A CONSIDERAR

De nuevo, debemos decidir cuántos factores usamos con este método, y seguimos el criterio anterior:

- $\frac{\sum_1^p \lambda_{ij}}{n} \geq x$: Buscamos que se explique un porcentaje mínimo x de las variables originales.

Una vez más, la varianza mínima explicada que nos interesa es el 80%, y vemos que de nuevo con 5 factores superamos el 80% con un 85%, mientras que con 4 factores se queda corto, en un 78%.

4.2. DISCUSIÓN DE LA CONVENIENCIA DE ELIMINAR O CREAR VARIABLES

La función fa() del paquete "psych" facilita información muy interesante como por ejemplo las comunalidades. En este caso son las siguientes:

0.9, 0.65, 0.64, 1, 0.98, 0.82, 0.92, 0.9, 0.58, 1, 0.98, 0.95, 0.96, 0.96, 0.85, 0.96, 0.97, 0.68, 0.97, 0.24, 0.98, 0.67

La variable con la comunalidad más baja es la de "tempConstr" con 0.24, un valor muy lejano a 1. Viendo la importancia que tiene en el estudio aumentamos el número de factores a 6. Ahora, las comunalidades han cambiado:

0.91, 0.68, 0.85, 0.99, 1, 0.88, 0.93, 0.98, 0.82, 0.99, 1, 0.98, 0.93, 0.96, 0.87, 0.95, 0.97, 0.71, 1, 0.49, 0.95, 0.8

Esta vez la comunalidad más baja es de nuevo "tempConstr" pero ha duplicado su varianza explicada por todos los factores a 0.49, casi un 50%, por lo que dejamos la solución con 6 factores.

4.3. INTERPRETACIÓN DE LA SOLUCIÓN FACTORIAL

Para interpretar los 6 factores usamos de nuevo la matriz A, esta vez redondeada a 2 decimales:

VARIABLE	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4	FACTOR 5	FACTOR 6
Proppobla	0.8	0.39	0.3	-0.14	0.01	0.09
Propsuperf	0.06	0.72	-0.05	-0.14	0.31	-0.14
ratioPobsup	0.58	-0.45	0.34	-0.12	-0.09	0.4
Ventasaotras	0.88	0.44	-0.09	-0.13	0.05	-0.02
Comprasdeotras	0.9	0.4	0.05	0.01	0.09	-0.12
Autoabas	0.82	0.39	0.05	-0.07	-0.02	-0.22
Maxventa	0.28	0.29	0.25	0.79	-0.29	-0.01
Maxcompra	0.39	0.12	-0.01	0.42	0.78	0.18
Ratioventascompras	0.14	0.37	-0.67	-0.14	-0.22	0.38
proporcionvtasT	0.88	0.44	-0.09	-0.13	0.05	-0.02
porporcioncomprasT	0.9	0.4	0.05	0.01	0.09	-0.12
ratiomaxventaTotal	0.07	0.19	0.37	0.86	-0.31	0.02
ratiomaxcompraTotal	-0.44	-0.27	0.02	0.32	0.73	0.11
indicegastopersona	0.79	-0.55	-0.16	0.1	0	-0.06
indicegastoporhoga	0.8	-0.45	-0.01	-0.04	-0.04	0.15
indicegastudconsumo	0.81	-0.53	-0.1	0.06	-0.01	0.01
pibpercap	0.6	-0.66	-0.39	0.01	0.08	0.07
templeoAgri	-0.51	0.61	-0.11	-0.15	0.06	0.18
tempIndus	0.07	0	-0.92	0.36	-0.09	-0.11
tempConstr	-0.22	-0.32	0.31	-0.14	0.11	-0.46
tempServ	0.41	-0.51	0.7	-0.18	0.02	0.02
tbuscandoempleo	-0.27	0.53	0.6	-0.07	-0.03	0.28

Las cargas factoriales que hemos considerado elevadas han sido las mayores a 0.7. Hemos coloreado las altas positivas de verde y las altas negativas de rojo. Las interpretaciones de cada factor son las siguientes:

FACTOR 1: CCAA con una gran población, una gran capacidad económica interna y extrena y mucho gasto por parte de la población, familias y empresas.

FACTOR 2: CCAA con poca superficie de terreno.

FACTOR 3: CCAA con poca industria y mucho sector servicios.

FACTOR 4: CCAA con gran volumen máximo de ventas a otra CCAA.

FACTOR 5: CCAA con gran volumen máximo de compras a otra CCAA.

FACTOR 6: Este factor explica demasiado poco cada variable original, necesitamos ver la solución rotada por si mejora la capacidad explicativa de este factor.

4.4. ENSAYO DE LA SOLUCIÓN ROTADA VARIMAX

Para llevar a cabo la solución rotada Varimax simplemente hace falta especificar en la función *fa()* el atributo *rotate="varimax"*.

VARIABLE	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4	FACTOR 5	FACTOR 6
Proppobla	0.86	-0.13	0.36	0.1	-0.02	0.1
Propsuperf	0.53	-0.58	-0.12	-0.13	0.18	0.05
ratioPobsup	0.15	0.68	0.6	0.02	-0.02	0.11
Ventasaotras	0.96	0.18	0	-0.01	0	0.21
Comprasdeotras	0.96	0.2	0.02	0.13	0.06	0.06
Autoabas	0.91	0.15	-0.02	0.1	-0.09	-0.02
Maxventa	0.24	0	-0.03	0.93	0.03	0.05
Maxcompra	0.36	0.12	-0.02	0.13	0.9	0.07
Ratioventascompras	0.19	-0.07	-0.32	-0.18	-0.17	0.78
proporcionvtasT	0.96	0.18	0	-0.01	0	0.21
porporcioncomprasT	0.96	0.2	0.02	0.13	0.06	0.06
ratiomaxventaTotal	0.01	-0.05	0.05	1	0.04	-0.02
ratiomaxcompraTotal	-0.49	-0.07	-0.04	-0.06	0.81	-0.18
indicegastopersona	0.31	0.92	-0.09	0.02	0.04	-0.04
indicegastoporhoga	0.36	0.84	0.17	-0.03	-0.01	0.09
indicegastudconsumo	0.33	0.92	0.01	0.02	0.03	0
pibpercap	0.08	0.93	-0.19	-0.20	0.11	0.11
templeoAgri	-0.09	-0.77	-0.04	-0.14	0.02	0.3
tempIndus	-0.04	0.21	-0.89	0.07	0.04	0.39
tempConstr	-0.20	0.02	0.06	-0.14	-0.05	-0.65
tempServ	0.11	0.52	0.69	0.01	-0.04	-0.42
tbuscandoempleo	0.06	-0.64	0.59	0.19	0	0.05

Las cargas factoriales que hemos considerado elevadas han sido de nuevo las mayores a 0.7. Hemos coloreado las altas positivas de verde y las altas negativas de rojo. Las interpretaciones de cada factor son las siguientes:

FACTOR 1: CCAA con una gran población y capacidad económica interna y externa.

FACTOR 2: CCAA con un gran gasto por parte de familias y empresas, y poca agricultura.

FACTOR 3: CCAA con poca industria.

FACTOR 4: CCAA con grande volumen de ventas máximo a otra CCAA.

FACTOR 5: CCAA con gran volumen de compras máximo a otra CCAA.

FACTOR 6: CCAA con gran capacidad económica en sentido de transacciones compraventa.

4.5. COMPARACIÓN DE LOS ENSAYOS PREVIOS

Comparamos primero la solución rotada con la solución sin rotar con el método de ejes principales. Queda claro que al haber seleccionado un factor más es más complicada su interpretación, pero al rotar la solución podemos interpretarlo más fácilmente ya que se extrapolan los valores. Por lo tanto nos quedaríamos con la solución rotada de entre estas dos.

Ahora, comparamos la "solución ganadora" hasta ahora que es la de componentes principales rotada con la solución también rotada del método de ejes principales. Mientras que es interesante que la segunda nos de un factor explicativo de la industria, el sexto factor no parece tener una interpretación muy clara teniendo ya los factores 4 y 5, así que seguimos seleccionando la primera, por el método de componentes principales, ya que el resto de factores son idénticos.

5. CLUSTERIZACIÓN POR EL MÉTODO JERÁRQUICO WARD

Antes de iniciar el proceso de clusterización, cabe mencionar la relación entre la distancia de Mahalanobis y las componentes principales con distancia euclidiana para el clustering. La distancia de Mahalanobis es una medida de distancia que tiene en cuenta la covarianza entre las variables. Cuando la matriz de covarianza es diagonal, la distancia de Mahalanobis se reduce a la distancia Euclídea escalada. En el Análisis de Componentes Principales (PCA), la matriz de covarianza entre las componentes principales es diagonal, lo que significa que las componentes principales son incorrelacionadas. Sin embargo, cuando se realiza clustering, para cada CP la distancia depende de su propia varianza explicada, así que con la distancia euclidiana se pueden ver las CP como no anómalas, mientras que con la distancia de Mahalanobis podríamos concluir que lo son.

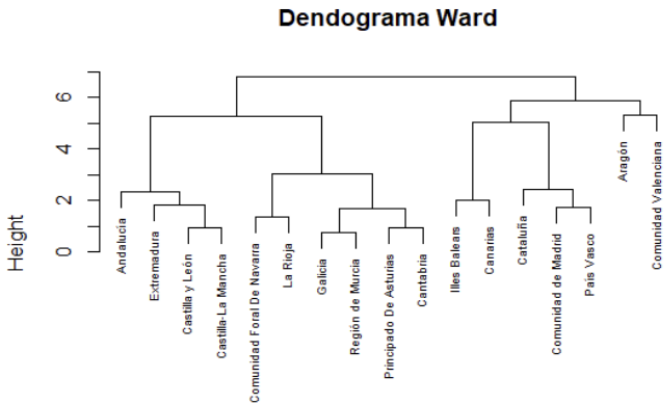
Guardamos la solución factorial rotada de componentes principales (puntuaciones de las CCAA para cada factor), que ha sido la elegida, y cambiamos los nombres de las columnas (factores) por una pequeña interpretación con *colnames(solucion_factorial)*. Ya guardada, la usamos para la clusterización por el método jerárquico Ward con el código *as.array(puntuaciones%*%mat.rotacion,dimnames=C(FR1,FR2,FR3,FR4,FR5))*, que son las puntuaciones rotadas.

El método jerárquico Ward se centra en minimizar la varianza dentro de los grupos, es decir, hacerlos lo más homogéneos en sentido intragrupal. Comienza considerando

cada punto de datos como un grupo individual y luego fusiona gradualmente los grupos más cercanos entre sí, de manera que se minimice la variabilidad en los grupos.

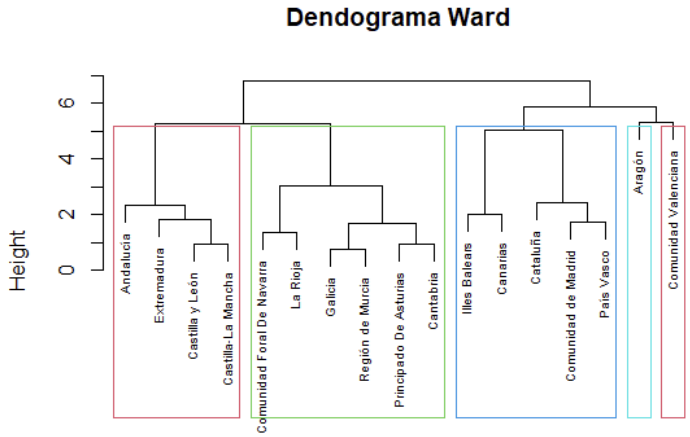
Para realizar la agrupación tenemos que calcular las distancias usando el código *dist(sol_factorial)*, y después realizar los pasos habituales de clusterización: primero se usa la instrucción *hclust()*, especificando el método que en nuestro caso es "ward.D". Observamos el dendograma y según la cantidad de homogeneidad interna que pierdan los clusters en el proceso de agruparse elegimos el número de grupos que usar.

El dendograma lo creamos con *plot(cluster)*, y se ve de la siguiente forma:

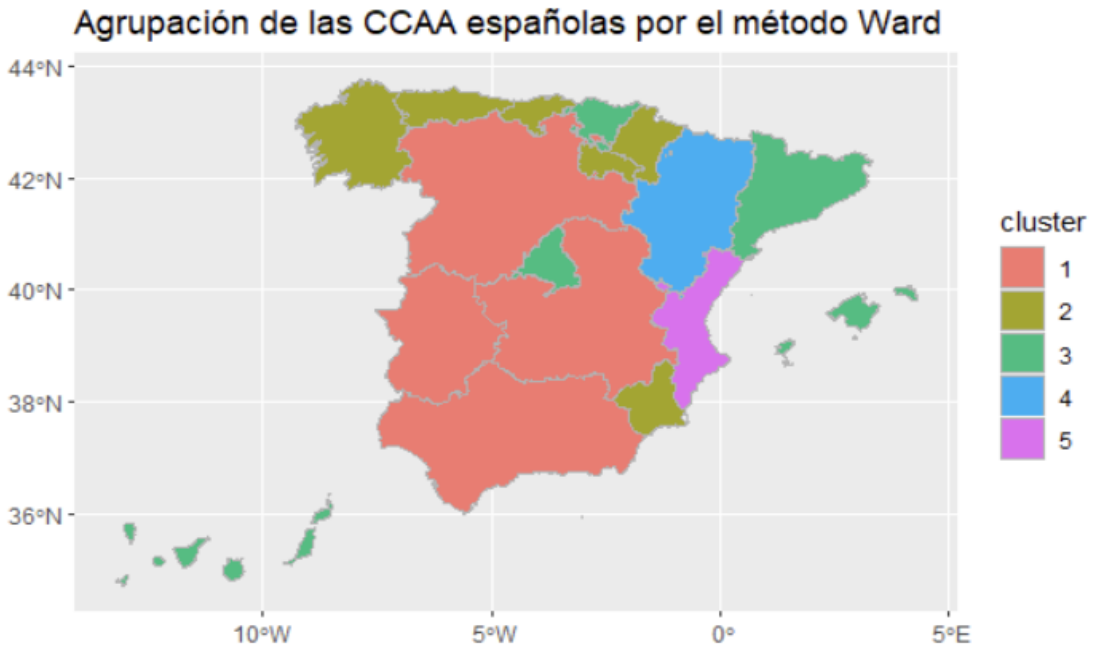


Observando el dendograma podemos ver que con 5 grupos no se pierde demasiada homogeneidad interna mientras que se reduce mucho el número de clusters.

Por lo tanto, la solución de la agrupación se ve tal como se muestra en el segundo dendograma.



Es muy interesante el aislamiento de las dos CCAA que componen los últimos clusters: Aragón y la Comunidad Valenciana. Así que tendremos en cuenta que al analizarlos estaremos hablando de unos grupos concretos y lo suficientemente diferentes o extraños como para que se hayan exiliado en diferentes grupos.



Para estudiar cómo se caracterizan debemos realizar un análisis de varianza que nos de qué factores y variables originales son los más diferenciadores de cada grupo.

6. ANÁLISIS DE VARIANZA SOBRE FACTORES Y VARIABLES ORIGINALES

Para poder estudiar qué diferencia a cada uno de los 5 grupos, miramos las medias y detectamos las que tengan una diferencia mayor entre grupos dentro de una misma agrupación. Para esto usamos la función *aggregate()* con el modelo previo al que llamamos Z, la variable con la asignación a grupos que hemos añadido a la solución factorial previamente y usando la función *FUN=mean*. Las medias son las siguientes:

CLUSTER	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4	FACTOR 5
1	0.56	-1.226	-0.04	0.438	-0.126
2	-0.015	0.32	-0.745	0.12	-3.562
3	-0.695	0.034	-0.598	0.148	0.569
4	0.287	0.861	0.92	0.198	0.13
5	0.51	0.075	-0.112	-3.754	0.003

Los grupos más diferentes al analizar las medias de cada uno en cada factor son precisamente uno diferente para cada cluster, es decir, cada factor caracteriza a un cluster de forma más significativa.

Para estudiar cómo de significativas son estas diferencias y su dirección realizamos el análisis de varianza. Usamos primero ANOVA para encontrar los factores más significativos en la agrupación con el *summary()* de la función *aov(factor~cluster)*. Encontramos diferencias significativas en los factores 2, 4 y 5 (aunque el 3 estuvo cerca con un p-valor de 0.118).

Habiendo identificado los factores con diferencias significativas buscamos entre qué grupos y su dirección con un test Scheffe, que es el más duro al encontrar diferencias, con la función *ScheffeTest(anova)* del paquete "DescTools". Este análisis es útil cuando tienes más de dos grupos y deseas determinar cuáles de ellos son significativamente diferentes entre sí, cosa que con el ANOVA no podemos conseguir. Los resultados son los siguientes:

FACTOR 2 – ALTOS GASTOS DE LA POBLACIÓN Y POCA AGRICULTURA
Con unos gastos de la población y poca agricultura mayores para el 4 que para el 1 a un nivel de significación de 0.05.

FACTOR 4 – POCO VOLUMEN DE VENTAS MÁXIMO A OTRA CCAA
Con un volumen máximo de ventas menor a cualquier nivel de significación para el grupo 5 en comparación con todos los demás grupos.

FACTOR 5 – POCO VOLUMEN DE COMPRAS MÁXIMO A OTRA CCAA
Con un volumen máximo de compras menor a cualquier nivel de significación para todos los grupos en comparación con el grupo 2, menos para el 1, que se da al contrario. Y con un volumen máximo de compras menor a un nivel de significación del 0.1 en el grupo 3 en comparación con el grupo 1.

De esta forma podemos hacernos una idea de cómo se caracteriza cada grupo:

GRUPO 1: se caracteriza por tener un volumen máximo de compras menor, sugiriendo una posible menor dependencia de deuda externa o necesidad de importaciones significativas. Además, muestra un volumen máximo de ventas mayor al Grupo 5, indicando una fuerte capacidad para vender productos a otras regiones. Este grupo exhibe gastos de la población más bajos, sugiriendo una capacidad económica menor, y una mayor presencia de agricultura, lo que podría asociarse a áreas conocidas como "la España vaciada", zonas más rurales.

GRUPO 2: muestra un volumen máximo de compras menor que todos los grupos excepto el Grupo 1, indicando una menor dependencia de deuda externa o necesidad de importaciones significativas. Además, presenta un volumen máximo de ventas mayor que el Grupo 5, sugiriendo una capacidad significativa para vender productos a otras CCAA. Este grupo exhibe un mayor gasto de la población en comparación con el

Grupo 1 y una menor presencia de agricultura, lo que lo coloca en zonas más urbanas y desarrolladas económicamente.

GRUPO 3: tiene un volumen máximo de compras menor que el Grupo 1, lo que sugiere una menor dependencia de deuda externa o necesidad de importaciones significativas. Muestra un volumen máximo de ventas mayor que el Grupo 1, indicando una capacidad para vender productos a otras CCAA. Por último, este grupo tiene una menor presencia de agricultura en comparación con el Grupo 1.

GRUPO 4: presenta un volumen máximo de compras mayor que el Grupo 5, lo que podría indicar una mayor dependencia de importaciones o una actividad económica más intensa que involucra compras a otras CCAA. Los gastos de la población son más altos que en el Grupo 1, sugiriendo una capacidad económica superior. Además, este grupo tiene una menor presencia de agricultura en comparación con el Grupo 1.

GRUPO 5: se caracteriza por un volumen máximo de ventas menor que todos los grupos, sugiriendo una capacidad limitada para vender productos a otras regiones, posiblemente debido a restricciones económicas o estructurales. No se proporcionan detalles sobre gastos de la población y presencia de agricultura.

Ahora realizamos el análisis varianza sobre las variables originales. En resumen, las medias nos indican las diferencias a primera vista siguientes:

Propsuperf: más alta para los grupos 1 y 2.
RatioPobsup: mucho más alta para los grupos 4 y 5.
Comprasdeotras: levemente menor para el grupo 3.
maxventa: más alta para el grupo 5.
maxcompra: más alta para el grupo 2.
ratiomaximaventatotal: mucho más alta para el grupo 5.
ratiomaxcompraTotal: levemente más alta para el grupo 2.
indicegastopersona: menor para el grupo 1.
indicegastounidadconsumo: mayor para el grupo 4.
pibpercap: menor para el grupo 1.
templeoAgri: mayor para el grupo 1.
tempServ: mayor para el grupo 4.

Ahora, realizamos el test ANOVA con estas variables en las que ya hemos detectado diferencias. Las variables con diferencias más o menos significativas fueron “PropSuperf”, “ratioPobsup”, “ratiomaximaventaTotal”, “indicegastounidadconsumo”, “pibpercap”, “templeoAgri” y “tempServ”.

Para conocer la dirección de la diferencia y en qué grupos se encuentra, realizamos el test Scheffe, cuyos resultados son los sig:

PROPSUPERF:
Con una proporción de superficie mayor a cualquier nivel de significación para el

grupo 1 en comparación con los grupos 3 y 4, y para el 5 a un nivel de significación del 0.1.

RATIOPOBSUP:
Con cantidad de población según la superficie mayor a un nivel de significación del 0.1 para el 4 en comparación con el 1 y 3.

RATIOMAXIMAVENTATOTAL:
Con una importancia de su mayor cliente mayor a cualquier nivel de significación para el grupo 5 en comparación con el resto de los grupos.

INDICEGASTOPERSONA, INDICEGASTOUNIDADCONSUMO Y PIBPERCAP:
Con un gasto medio por persona/unidad de consumo en relación al gasto nacional/ PIB per cápita mayor a un nivel de significación del 0.1 para el 4 en comparación con el grupo 1.

TEMPLEOAGRI:
Con una tasa de empleo en el sector agrario menor a un nivel de significación del 0.1 para el grupo 4 en comparación con el 1.

TEMPSERV:
Con una tasa de empleo en el sector servicios mayor a cualquier nivel de significación para el grupo 4 en comparación con los grupos 1 y 3.

Ahora podemos interpretar los cinco grupos según los resultados que acabamos de obtener:

GRUPO 1: se caracteriza por tener una proporción de superficie mayor en comparación con los grupos 3 y 4, y 5. Además, exhibe una cantidad de población según la superficie menor que en el Grupo 4. En cuanto a la importancia de su mayor cliente, esta es menor en comparación con el Grupo 5. Respecto al gasto medio por persona/unidad de consumo y al PIB per cápita, los tres son menores en comparación con el Grupo 4, de forma que tiene menor importancia económica. Además, la tasa de empleo en el sector agrario es mayor para el Grupo 1 en comparación con el Grupo 4, así que debe encontrarse en zonas más bien rurales.

GRUPO 2: presenta una importancia de su mayor cliente menor en comparación con el Grupo 5, lo cual podría significar una menor dependencia económica de su cliente principal (CCAA). En el resto de las variables no presenta diferencias interesantes.

GRUPO 3: se caracteriza por tener una proporción de superficie menor en comparación con el grupo 1. Además, una importancia de su mayor cliente menor en comparación con el Grupo 5. Por último, muestra una tasa de empleo en el sector servicios menor en comparación con el Grupo 4, indicando menor dependencia económica en actividades del sector servicios.

GRUPO 4: se caracteriza por tener una proporción de superficie menor en comparación con el grupo 1. Además de una cantidad de población según la superficie

mayor en comparación con el 1 y 3. Respecto al gasto medio por persona/unidad de consumo y al PIB per cápita, los tres son mayores en comparación con el Grupo 1, de forma que tiene mayor importancia económica. Una tasa de empleo en el sector agrario menor en comparación con el grupo 1, pero mayor tasa de empleo en el sector servicios en comparación con el Grupo 3, indicando mayor dependencia económica en actividades de servicios.

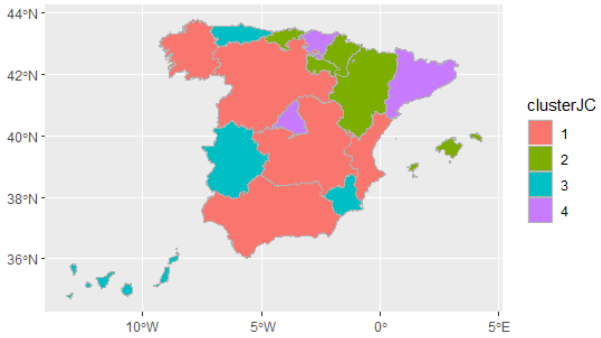
GRUPO 5: se caracteriza por tener una proporción de superficie menor en comparación con el grupo 1. Además, una importancia de su mayor cliente mayor en comparación con el resto de los grupos, lo cual podría significar una mayor dependencia económica de su cliente principal.

Ya interpretados los grupos de nuestro clustering, podemos ver cómo se diferencian con los que realizamos anteriormente en la práctica 3.

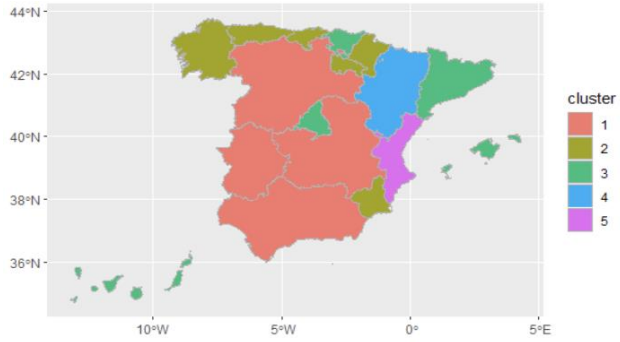
7. COMPARACIÓN CON RESULTADOS DE LA PRÁCTICA 3

En la práctica 3 acabamos eligiendo otro método jerárquico, el complete, que nos dio como resultado esta agrupación (primer gráfico):

Agrupación método complete con variables originales



Agrupación método Ward con factores por CP rotados



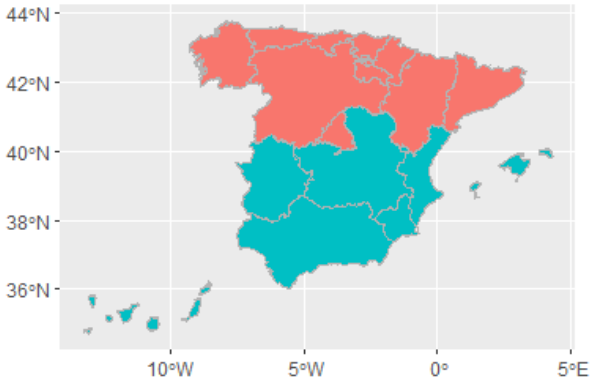
Como vemos, en la previa práctica elegimos realizar solo 4 grupos, mientras que ahora hemos elegido 5. Algunas características se mantienen mientras que otras difieren mucho.

Para empezar, el grupo 4 del primero es casi idéntico al grupo 3 del segundo, que tienen diferencias significativas que indican que son las potencias económicas. Además, ambos grupos 1 son muy parecidos, caracterizados por ser agrícolas, con una superficie mayor, pero con una menor concentración de población.

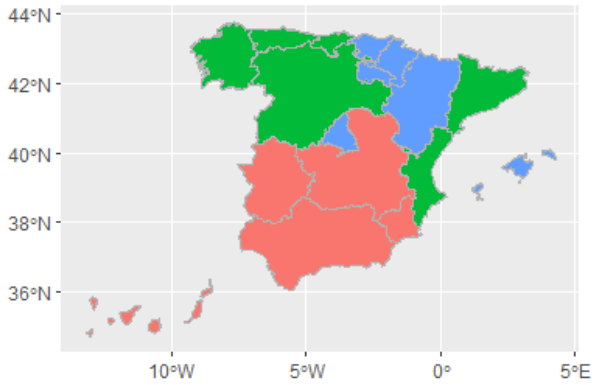
En cambio, el resto de los grupos son muy dispares, y mientras que en el primero no había ninguna CCAA aislada, en el segundo nos han surgido dos CCAA, Aragón por un lado, y por el otro la Comunidad Valenciana.

Algunas agrupaciones realizadas con criterios distintos no metodológicos, por una sola variable, son: por periferia o interior (ROJO=periferia, AZUL=interior), por lengua propia o no (ROJO=no lengua propia, AZUL=lengua propia), por norte o sur (ROJO=Norte, AZUL=Sur), y según el nivel de paro (ROJO= más del 15%, AZUL= menos del 10%, VERDE= entre el 10% y el 15%).

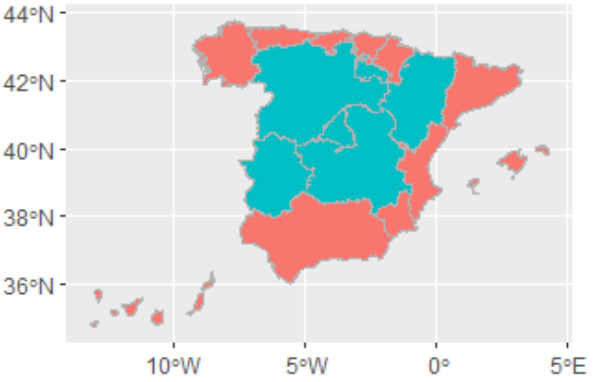
Agrupación según su ubicación en Norte-Sur



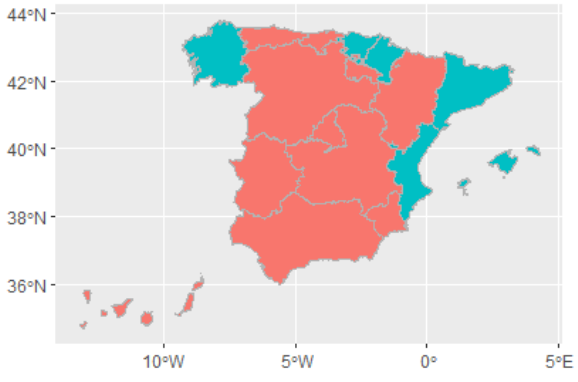
Agrupación según su nivel del paro



Agrupación según si son periferia o interior



Agrupación según si tienen lengua propia o no



Estas agrupaciones difieren mucho de las obtenidas en este estudio por Componentes Principales, tal vez podemos ver pequeñas semejanzas en las CCAA de interior y el grupo 1 sin contar Aragón, Madrid y las Islas Canarias, de forma que puede que el hecho de que sean interiores o periféricas influya, pero habría que realizar un estudio más a fondo para saberlo.

Lo que podemos concluir del hecho de que haya tantas diferencias según el método o criterio de agrupación es que España es un país muy variopinto, y que no se puede definir en un solo sentido, ya que tiene una riqueza cultural y social muy diversa según las zonas en las que nos encontremos.

8. CONCLUSIONES

Tras un extenso análisis factorial a través de distintos métodos hemos concluido que el mejor ha sido el ensayo sobre la obtención de componentes principales A través de un riguroso preproceso y criterios de selección de factores, hemos logrado una interpretación detallada de qué explican los factores socioeconómicos estudiados. La aplicación de la rotación Varimax ha servido como estrategia para mejorar la claridad interpretativa de las correlaciones. Los factores, por orden de mayor explicación a menor sobre el total de las variables originales, han sido los siguientes:

FACTOR 1: Cuando es alto, la CCAA es una gran potencia económica**.**
FACTOR 2: Cuando es alto, la población tiene pocos gastos y la CCAA poca agricultura.
FACTOR 3: Cuando es alto, hay mucha compraventa, poca industria y mucho servicio.
FACTOR 4: Cuando es alto, el volumen de la mayor venta en la CCAA es bajo.
FACTOR 5: Cuando es alto, el volumen de la mayor compra en la CCAA es bajo.

Este enfoque nos ha facilitado la realización de análisis posteriores, como la clusterización de las Comunidades Autónomas basada en los factores indicados, por el método jerárquico de Ward. La agrupación ha quedado de la siguiente forma:

GRUPO 1: Andalucía, Extremadura, Castilla y León y Castilla-La Mancha.
GRUPO 2: Comunidad Foral De Navarra, La Rioja, Galicia, Región de Murcia, Principado de Asturias y Cantabria.
GRUPO 3: Islas Baleares, Canarias, Cataluña, Comunidad de Madrid y País Vasco.
GRUPO 4: Aragón.
GRUPO 5: Comunidad Valenciana.

Siendo las CCAA con un comportamiento singular Aragón y la Comunidad Valenciana, que se han agrupado de forma aislada del resto.

Las variables originales y factores más diferenciadores según el ANOVA tras haber analizado las medias de los grupos fueron “propsuperf”, “ratioPobsup”, “ratiomaximaventaTotal”, “indicegastopersona”, “indicegastounidadconsumo”, “pibpercap”, “templeoAgri”, “tempServ”, y el segundo, cuarto y quinto factor. A través del estudio del test Scheffe hemos podido detectar la dirección y la significatividad de las diferencias en cada variable y factor. De forma que la interpretación de cada grupo queda de esta forma:

GRUPO 1: Se destaca por una proporción de superficie mayor, menor dependencia de deuda externa, capacidad para vender a otras regiones y una presencia significativa de agricultura. Indica una posible ubicación en áreas rurales con menor capacidad económica.

GRUPO 2: Muestra menor dependencia de deuda externa, capacidad para vender productos a otras CCAA y mayor gasto de la población. Sugiere una ubicación en zonas urbanas y económicamente desarrolladas.

GRUPO 3: Caracterizado por menor proporción de superficie, capacidad para vender a otras CCAA y menor presencia de agricultura. Indica una ubicación con menor dependencia económica en actividades del sector servicios.

GRUPO 4: Destaca por mayor importancia económica, menor presencia de agricultura y mayor tasa de empleo en el sector servicios. Sugiere una ubicación con actividad económica intensa y mayor desarrollo.

GRUPO 5: Se caracteriza por una proporción de superficie menor, mayor dependencia de su cliente principal y menor capacidad para vender a otras regiones. Indica una posible mayor dependencia económica de su cliente principal.

BIBLIOGRAFÍA

Is Mahalanobis distance equivalent to the Euclidean one on the PCA-rotated data? (n.d.). Cross Validated. https://stats.stackexchange.com/questions/166525/is-mahalanobis-distance-equivalent-to-the-euclidean-one-on-the-pca-rotated-data

Revelle, W. (2023, December 20). *Procedures for Psychological, Psychometric, and Personality Research [R package psych version 2.3.12]*. https://cran.r-project.org/web/packages/psych/index.html

Imle. (n.d.). *A.Factorial sobre coches.sav*. https://www.uv.es/mlejarza/datamine/faccoches2a.html

colaboradores de Wikipedia. (2023, May 14). *Máxima verosimilitud*. Wikipedia, La Enciclopedia Libre. https://es.wikipedia.org/wiki/M%C3%A1xima_verosimilitud

Interpretar todos los estadísticos y gráficas para Análisis factorial - Minitab. (n.d.). (C) Minitab, LLC. All Rights Reserved. 2023. https://support.minitab.com/es-mx/minitab/21/help-and-how-to/statistical-modeling/multivariate/how-to/factor-analysis/interpret-the-results/all-statistics-and-graphs/#factor-score-coefficients

IBM documentation. (2023, August 4). https://www.ibm.com/docs/es/spss-statistics/saas?topic=analysis-factor-scores