FACTOR ANALYSIS OF THE SOCIOECONOMIC VARIABLES ON THE AUTONOMOUS COMMUNITIES OF SPAIN

DATA MINING IN BUSINESS

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

VNIVERSITAT DE VALÈNCIA

BUSINESS INTELLIGENCE & ANALYTICS

INDEX

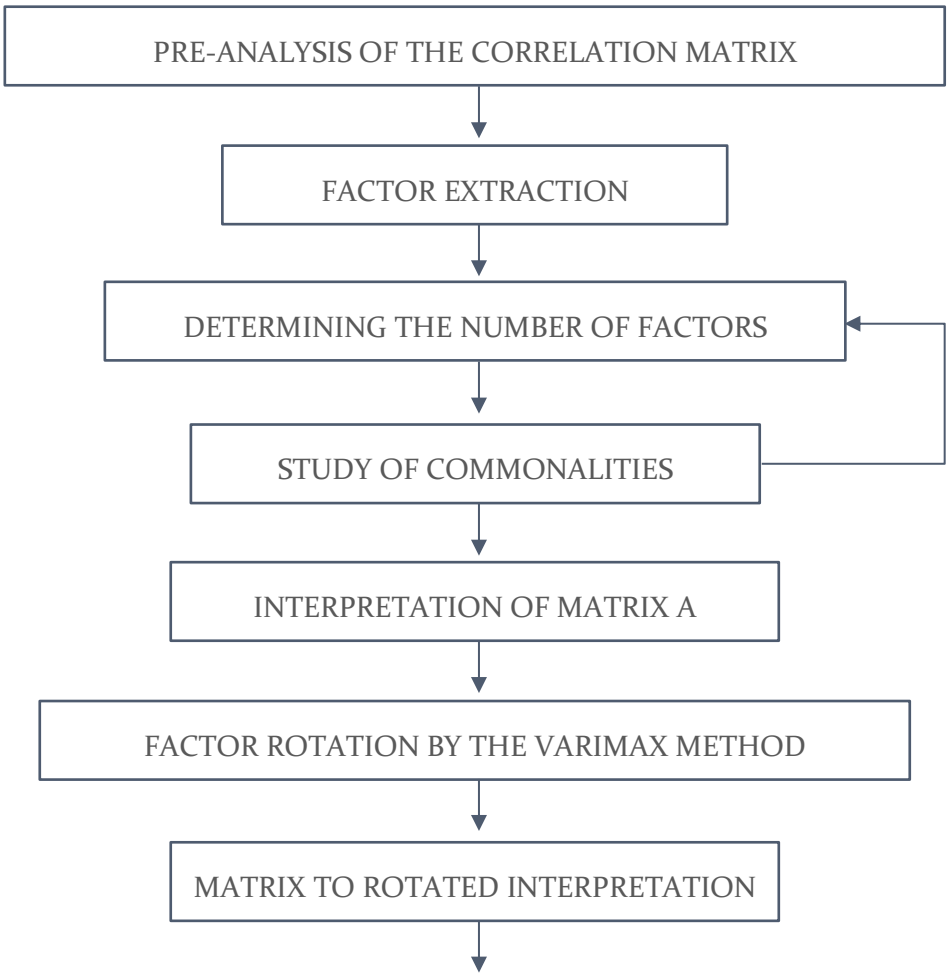
1. INTRODUCTION

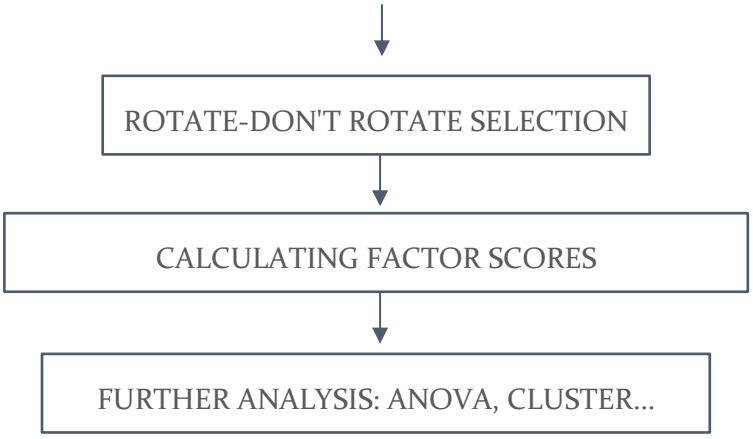
This report presents a detailed factor analysis based on socioeconomic variables of the Spanish Autonomous Communities collected in the "comaut.sav" file. The aim of this study is to use factorization techniques to try to explain most of the original variables with a smaller number of factors. Throughout the study we used R's RStudio environment. The methods used were: Principal Components, Varimax Rotation, Maximum Likelihood and Principal Axes.

In addition, we wanted to take advantage of the chosen factor analysis to carry out a hierarchical Ward grouping. In this way, we will create clusters of Autonomous Communities that are homogeneous within groups and heterogeneous between groups explained by factors.

The general criterion for selecting the method will focus not only on reducing the number of variables, but also, on the one hand, on the fact that the factors explain most of the original variables, and on the other hand, that the factors are easily interpretable.

The scheme we are going to follow is as follows:





2. MAIN COMPONENT TESTING

One of the methods that are commonly used to extract explanatory factors from variables is principal component analysis which is based on the Spectral Theorem. This approach involves the rotation of the original axes in a new reference system for the variables: the principal components. Generally, as many components are obtained as there are variables, explaining the totality of the originals. But if the object of the study is to factor, we can select only a few components, even if this implies losing a small percentage of explanation.

2.1. PRE-PROCESSING OF PRINCIPAL COMPONENTS

To start our analysis, we have loaded the "haven" and "corrplot" libraries for later use, and we have imported the "comaut.sav" file as "data".

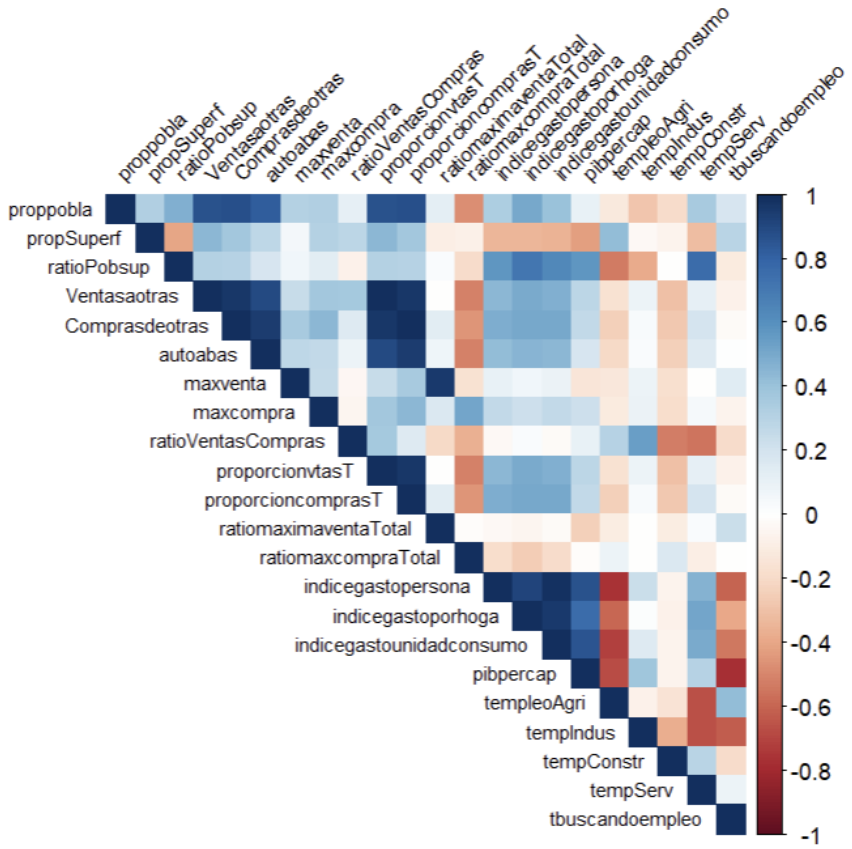
It is crucial for the study that all variables in the dataframe are numerical. Therefore, we have saved the variable "Cautonoma" (which contains the name of the Autonomous Communities, and therefore is a character type) as an object in the environment and we have removed it from "data" to add it as row names with *row.names(data)<-Cautonoma*. Then, we used the *as.data.frame(lapply(data, as.numeric))* statement to make sure that the rest of the variables are indeed numeric.

Reviewing the meaning of the variables, we identified that "density", "Expenditure per unitconsumption", "Expenditure per household" and "Expenditure per person" explain the same as "ratiopobSup", "indexexpenditureunitconsumption", "indexexpenditures/hoga" and "indexexpenditureperson", respectively, despite having different nuances in terms of size. Therefore, we eliminate the first four as we prefer to have indices and ratios.

Subsequently, we have decided to relativize the absolute variables (except those that have an important significance for the study) in order to avoid the effect of the size of the Autonomous Communities on the correlation. To do this, we've divided each value by the sum of all its values. The variables affected were "self-sufficiency", "Other-selling", "Other's purchases", "max-selling" and "max-buying".

Once the pre-process is finished, we can start to get an idea of how the factors are going to behave. We created a correlation matrix with the cor*(data, use = "pairwise.complete.obs")*code, and then visualized it with *corrplot(matriz_correlaciones).*

We look for "constellations", groups of very high correlations positively or negatively, which will then become what each of the factors explains. In our case, the matrix looks like this:



In our matrix we can detect strong correlations in an economic sense with spending, purchases and sales. However, it is not very easy to see more constellations with these variables.

We are interested in high correlations for interesting factors to be formed. To easily check if there are, we can calculate the determinant of the correlation matrix, which should give us close to 0, and in our case it does indeed give -1.703072e-107.

Finally, using the eigen(cor(data)) instruction, we obtain both the eigenvalues and the eigenvectors of the correlation matrix, which allows us to move on to the factor extraction phase.

2.2. NUMBER OF FACTORS TO CONSIDER

To get all the Principal Components we use the *prcomp(data, scale.=TRUE) function*. The main information is obtained by means of a simple summary(), which provides us with the standard deviation, the proportion of variance and the cumulative proportion of each Principal Component, ordered from most explanatory to least explanatory. We will look at these in depth when we have chosen the number of factors.

From the standard deviation provided by *prcomp(),* we can extract the eigenvalues (deltas) of each Principal Component. The output, translated from scientific notation, is as follows:

[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

From here we can extract the T matrix or standard matrix (variables x factors) with the rotation attribute. The coefficients of this matrix are the factor weights (aij), which measure the linear functional relationship between variables and factors. What we are interested in is that each variable loads high in one factor and low in the rest, that is, we want each variable to go mostly to only one factor in order to form interesting factors. Importantly, we haven't typed yet, and we can verify this by calculating its standard deviation, which is still different from 1. We could typify by multiplying by $D^{1/2}$, but it's simpler to type the scores directly.

Now that we have the Principal Components and the eigenvalues, we can choose the number of factors. The criteria commonly used to select factors are several, including:

- $\boldsymbol{\lambda_{ij} > 1}$: Since the original variables already explain a 1, we expect the factors to at least explain the same as the original variables.
- $\frac{\sum_1^p \lambda_{ij}}{n} \geq \boldsymbol{x}$: We are looking for a minimum percentage x of the original variables to be explained.

As the No Free Lunch Theorem states, there is no perfect criterion for every situation. With a total of 22 variables, requiring the factors to explain at least what the originals already explained may be unrestrictive, leading us to select many factors, which would complicate their interpretation. In our case, we would choose 5 factors.

The second criterion may be more interesting, as it allows us to adjust the level of constraint according to our needs, but a minimal joint explanation of % is often used. We make the calculations by dividing the sum of the eigenvalues of the factors we are

interested in by the number of original variables, and we discover that with 4 factors we already exceed 80% (80.18137% is explained), so we choose to use 4 factors.

Once the number of factors has been chosen, we can create the model with *prcomp(data, scale. = TRUE, rank=4),* which now does not contain as many factors as variables but the 4 that we have selected. With the model we can already obtain matrix A or matrix of correlations between variables and factors with *cor(data,mod2$x)*.

2.3. DISCUSSION OF THE ADVISABILITY OF ELIMINATING OR GENERATING VARIABLES

To find out if we can eliminate variables, we calculate the commonalities () as the sum of the factor weights (each row of the matrix A -variables-) squared. In this way we see the proportion of variance explained by the common factors in a variable. If there is a variable that is not well explained by all the factors (commonality is not close to 1), we can see it this way. The output is as follows: h_j^2

[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

As we can see, numbers 8 ("maxbuy") and 13 ("maxmaxtotalbuyratio") are very low (<0.6), so we analyze them. We see that they do have importance in the study, because they generate new perspectives for the study (it is a mistake to think that they are the opposite of "maxventa" and "ratiomaxcompratotal" and that therefore they generate multicollinearity). Therefore, the solution is to increase the number of factors to 5.

Let's adjust the necessary codes, in this case the *prcomp(data, scale. = TRUE, rank=5)* model and the matrix A *cor(data,mod3$x)*. And we recalculate the commonalities, which this time give us the following:

[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

Now all commonalities are close to 1, i.e., all variables are highly explained by the set of factors chosen. Therefore, the number of factors that we are going to use in this study will ultimately be 5.

2.4. INTERPRETATION OF THE FACTORIAL SOLUTION

Having determined the optimal number of factors, we proceed to study the factorial solution. The factorial solution is the set of factors we have chosen and their meaning. The output of the *summary()* of the 5-factor model is as follows:

Importance of first k=5 (out of 17) components: PC1 PC2 PC3 PC4 PC5Standard deviation 2.857 2.1161 1.7230 1.42517 1.2690Proportion of Variance 0.371 0.2036 0.1349 0.09232 0.07083Cumulative Proportion 0.371 0.5745 0.7095 0.80181 0.8750

The first row indicates the standard deviation of each major component, i.e., how spread out the data is around the mean. The second row shows what proportion of variance explains each principal component with respect to the original variables. The third row is the cumulative proportion of variance, and in this case the chosen set of factors explains 87.035% of the variability in the original data.

It is interesting to note that the decreasing trend in standard deviations and variance ratios is due to the criterion for ordering the components: the former captures as much variability as possible in the data, the latter the remaining as much as possible, and so on.

With the matrix A that we have previously obtained and modified for 5 factors, we can study the meaning of each factor with the correlation matrix (rounded to 3 decimal places):

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
Salesto others	0.876	-0.439	-0.074	0.129	-0.054
ShoppingOther	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
MaxPurchase	0.386	-0.128	-0.003	-0.41	-0.767
SalesRatioPurchases	0.142	-0.389	-0.701	0.155	0.256
proportionvtasT	0.876	-0.439	-0.736	0.13	-0.054
portionpurchasesT	0.896	-0.401	0.057	-0.012	-0.093
ratiomaxsaleTotal	0.069	-0.176	0.354	-0.857	0.293
ratiomaxbuyTotal	-0.444	0.275	0.019	-0.324	-0.741
IndexExpenditurePerson	0.788	0.539	-0.172	-0.096	-0.011
indexexpoporhoga	0.811	0.451	-0.024	0.051	0.048
IndexExpenditure Consumption	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
Looking for a job	0.273	-0.535	0.639	0.058	0.074

The correlations that we have considered high have been those greater than 0.7. We've painted the positive highs green and the negative highs red. The interpretations are as follows:

FACTOR 1: Autonomous Communities with a large population and an economic power both internal and external, which allows its population to have a higher economic capacity (or that the price of living is higher).

FACTOR 2: Small Autonomous Communities.

FACTOR 3: Autonomous Communities with little buying and selling power and industry.

FACTOR 4: Autonomous Communities with a low maximum sales volume to another Autonomous Community.

FACTOR 5: Autonomous Communities with a low maximum purchase volume from another Autonomous Community.

2.5. VARIMAX ROTATED SOLUTION TEST

There are two types of rotations: orthogonal and oblique. The main advantage of orthogonal rotations is the incorrelation, since in oblique rotations two or more factors can explain the same thing at the same time, while in orthogonal rotations the opposite is achieved. Among the orthogonal rotations is the Varimax rotation (Maximum Variance), which tries to add variance to the factors, so that there are a few high saturations and many almost zero saturations in the variables, so that their interpretation of the correlations is usually clearer.

To calculate the rotated solution we use the *varimax(mod2$rotation)* code, and we need the scores, which we type with *scale()*. To interpret it, we need matrix A, which is the matrix of correlations between the factors and the original variables. The output rounded to 3 decimal places is as follows:

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
Salesto others	0.964	0.202	-0.109	0.02	0.054
ShoppingOther	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
MaxPurchase	0.41	0.16	-0.035	-0.136	-0.842
SalesRatioPurchases	0.271	-0.045	-0.738	0.198	0.304
proportionvtasT	0.964	0.202	-0.109	0.016	0.054
portionpurchasesT	0.952	0.225	0.008	-0.131	-0.035
ratiomaxsaleTotal	0.021	-0.066	0.034	-0.986	-0.052
ratiomaxbuyTotal	-0.466	-0.055	0.075	0.052	-0.836
IndexExpenditurePerson	0.281	0.933	0.024	-0.032	-0.02
indexexpoporhoga	0.378	0.829	0.164	0.016	0.096
IndexExpenditure Consumption	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
Looking for a job	0.147	-0.725	0.418	-0.216	0.094

The correlations that we have considered high have been those greater than 0.7. We've painted the positive highs green and the negative highs red. The interpretations of each factor are as follows:

FACTOR 1: Autonomous Communities with a large population and economic power, both internal and external.

FACTOR 2: Autonomous Communities with high population expenditures and little agricultural power.

FACTOR 3: Autonomous Communities with little buying and selling power and little industry but a high rate of employment in services.

FACTOR 4: Autonomous Communities with a low maximum sales volume to another Autonomous Community.

FACTOR 5: Autonomous Communities with a low maximum purchase volume from another Autonomous Community.

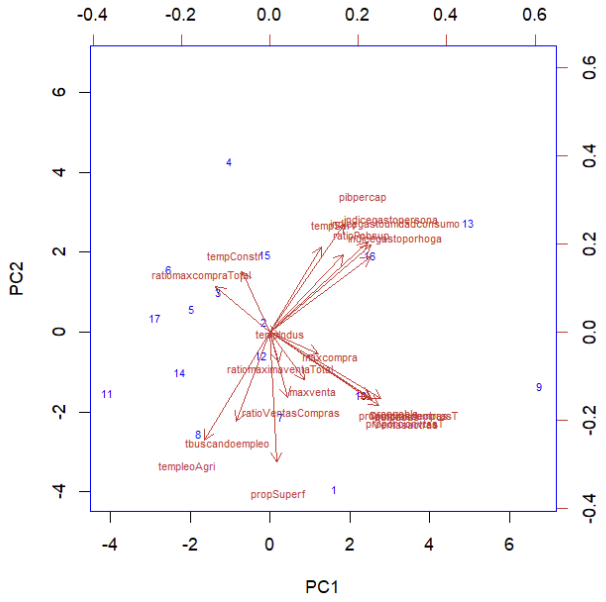
2.6. COMPARISON OF PREVIOUS TESTS

We compare the rotated and unrotated factorial solution. Despite the fact that the rotated solution is more complex, and that it could be said that it does not comply with Ockham's principle since the two factorial solutions are still the same, we consider that the rotated solution is the most easily interpretable. We see this especially in the second factor, since, for the socioeconomic purpose of the study, we are not interested in a factor explaining only the size of the Autonomous Communities, while the expenditure of the population and its power in a specific sector is much more transcendental for work.

We can graphically see the original variables and individuals in the plane of the first two principal components (which explain about half of the variability) with the biplot(x=model) instruction.

Each red vector represents an original variable in the database. Its direction indicates how it correlates with PC1 and PC2 (those pointing in the same direction are highly correlated, those pointing in the opposite direction are inversely correlated, but if they are perpendicular they are uncorrelated), and its length indicates the strength of that correlation (the longer it is, the more intense). Each blue dot is an individual or Autonomous Community in the database, and its location on the plane is determined by its scores in the factorial solution.

With the scores of the factorial solution, we will then be able to carry out the clustering of the Autonomous Communities based on the factors.



3. ALTERNATIVE TEST FOR MAXIMUM PLAUSIBILITY

Maximum Likelihood Factor Analysis is a statistical approach used to estimate the parameters of a factor analysis model. The main difference with Principal Component analysis is that through maximum plausibility the interpretation is much simpler because it tries to find unobserved or "latent" factors that explain the relationships between the original variables.

Ahora, the function is factanal(), and when using Maximum Likelihood, in the calculations that R performs in the background when executing the function it is divided by the determinant of correlation matrix A, so that if there is multicollinearity it is divided by 0 and does not converge. This is why we must select a smaller number of variables so that the correlation is not so high. In our case, we have opted for the following: 'proppobla', 'ratioPobsup', 'Ventasaotras', 'Comprasdeotras', 'autoabas', 'maxventa', 'maxcompra', 'ratioVentasCompras', 'proportionvtasT', 'ratiomaxcompraTotal', 'indexexpoporhoga', 'pibpercap', 'tempIndus', 'tempServ' and 'tbuscando empleo'. A total of 15 variables compared to the 22 we were left with in the previous factor analysis.

3.1. NUMBER OF FACTORS TO CONSIDER

The difference with the PCA method, in the maximum likelihood method you cannot know what is the number of factors that you are interested in using previously, so you can include all the possible factors with the *factanal function()* (in our case it allows us up to 9) to see what is the cumulative explanation according to the number of factors and follow the criterion we followed above:

- $\frac{\sum_1^p \lambda_{ij}}{n} \geq \boldsymbol{x}$: We are looking for a minimum percentage x of the original variables to be explained.

This minimum percentage being the same as the previous one, 80%, we must use a total of 5 factors, which explain 87.5% of the variance, while 4 explains 78.6%.

3.2. DISCUSSION OF THE ADVISABILITY OF ELIMINATING OR GENERATING VARIABLES

We can extract the commonalities by subtracting 1 from the unities, since communalities+unicities=1.

The commonalities, *1-fa$uniquenesses*, rounded to three decimal places are as follows:

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

And again according to the previous criteria, those below 0.6 are considered low. Therefore we see that variable number 4, which is "maxventa", is very low (0.424), so we remove " *maxventa" from* factanal().

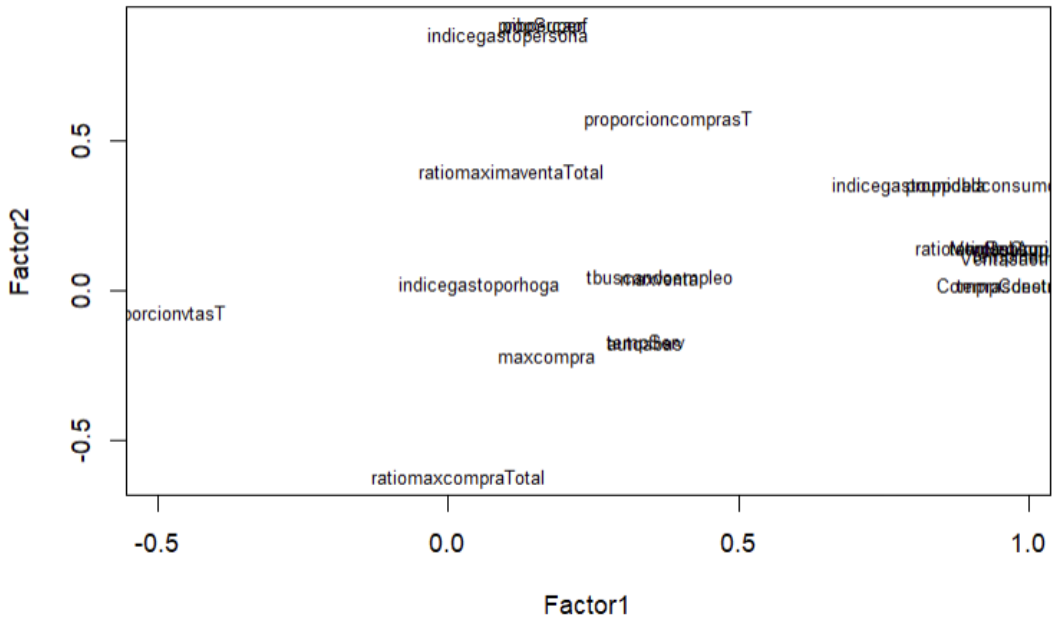
Now, with the "starting values" of the default function it does not converge. We could make a "start" from the factorial loads of the previous factorial solution, but we would have to create from scratch a matrix with the specific loads of the variables we have used, which since they are not the total we cannot use as such as start=mod3$loadings, because that matrix has 22 rows and not 14. To save this process, we can create an array with random values until the function works again. In order for there to be reproducibility of the data, we have to use a seed on which the data converges, which in this case is 3. In this way we get it to work, and we recalculate the commonalities by subtracting 1 from the unities. This time they are as follows:

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

As we can see, many commonalities go down and to values much lower than 0.424, which was the "maxventa", so we return to the previous factorial solution, since the one we obtain without this variable is much worse since the random starting point is much worse than the "default".

3.3. INTERPRETATION OF THE FACTORIAL SOLUTION

To interpret the factorial solution by maximum likelihood, we need to extract the factorial loads that correspond to *fa$loadings()*. The graph that represents the relationship between the selected variables and the first two factors that explain just over half of the total variability is as follows:



And the specific values of factor loads rounded to 3 decimal places are as follows:

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
Salesto others	0.952	0.141	0.117	0.012	0.239
ShoppingOther	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
MaxPurchase	0.364	0.046	0.096	0.922	0.006
SalesRatioPurchases	0.17	-0.226	0.164	-0.138	0.93
proportionvtasT	0.952	0.141	0.117	0.012	0.239
ratiomaxbuyTotal	-0.496	-0.074	0.014	0.761	-0.208
indexexpoporhoga	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
Looking for a job	0.052	0.019	-0.868	-0.001	-0.066

The factor loads that we have considered high have been those greater than 0.7. We've painted the positive highs green and the negative highs red. The interpretations of each factor are as follows:

FACTOR 1: Large Autonomous Communities with great economic importance in buying and selling and self-sufficiency.

FACTOR 2: Autonomous Communities with a large population and based in the service sector.

FACTOR 3: Autonomous communities with high GDP and low job search rate.

FACTOR 4: Autonomous Communities with high rates of purchase in quantity from other Autonomous Communities.

FACTOR 5: Autonomous Communities with high rates of purchase-sale transactions to other Autonomous Communities.

3.4. VARIMAX ROTATED SOLUTION TEST

The solution we get with the Varimax rotation is exactly the same as the one we get without rotating. Based on the principle of Ockham's razor, all things being equal, the simple method is always the best, we would keep the solution without rotating.

3.5. COMPARISON OF PREVIOUS TESTS

Now we choose between the factorial solution rotated by Principal Components and the solution not rotated by Maximum Likelihood. We first compared the explanatory capacity of both, and both explain 87.5% of the variability with 5 factors. So, we compare the simplicity of interpretation: the fact that in Maximum Likelihood the factor 5 explains the buy-sell transactions while the 4 only the purchase quantity seems to me unclear compared to those of Principal Components which are buy on the one hand and sell on the other, so we choose again the rotated Principal Components.

4. TESTING THE MAIN AXIS METHOD

Principal axis factor analysis is an iterative method based on the successive extraction of the factors that explain most of the common variance. This is achieved through the estimation of matrix A by the principal component method. Its main advantage, according to Winter and Dodou (2012), is its ability to recover weak factors, and it is recommended especially in factor analyses for small samples (although they are more likely not to converge) in which there are few variables and with moderate correlations, and it is usually used when the maximum likelihood method fails to converge.

Interestingly, iterations do not try to find the best solution, but rather to conbvberger the fastest way by successively decomposing the eigen values.

This time, the function is *fa()* from the "pysch" package, and we need to add the attribute *fm = "pa",* because it can use different factoring methods, and the "pa" refers to the "main factor solution". In addition, as we have mentioned, it is an iterative method and for this reason the maximum number of iterations must be specified (if a stop criterion is not given first), which we have set at 50. Finally we need to specify the type of rotation, since by default it is "oblimin", and for now we set "none" so that it doesn't use any.

4.1. NUMBER OF FACTORS TO CONSIDER

Again, we need to decide how many factors we use with this method, and we follow the criteria above:

- $\frac{\sum_1^p \lambda_{ij}}{n} \geq \boldsymbol{x}$: We are looking for a minimum percentage x of the original variables to be explained.

Once again, the minimum explained variance that we are interested in is 80%, and we see that again with 5 factors we exceed 80% with 85%, while with 4 factors it falls short, by 78%.

4.2. DISCUSSION OF THE DESIRABILITY OF DELETING OR CREATING VARIABLES

The fa() function of the "psych" package provides very interesting information such as commonalities. In this case they are as follows:

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

The variable with the lowest commonality is that of "tempConstr" with 0.24, a value far from 1. Seeing the importance it has in the study, we increased the number of factors to 6. Now, the commonalities have changed:

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

This time the lowest commonality is again "tempConstr" but it has doubled its variance explained by all factors to 0.49, almost 50%, so we leave the solution with 6 factors.

4.3. INTERPRETATION OF THE FACTORIAL SOLUTION

To interpret the 6 factors we use the matrix A again, this time rounded to 2 decimal places:

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
Salesto others	0.88	0.44	-0.09	-0.13	0.05	-0.02
ShoppingOther	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
MaxPurchase	0.39	0.12	-0.01	0.42	0.78	0.18
SalesRatioPurchases	0.14	0.37	-0.67	-0.14	-0.22	0.38
proportionvtasT	0.88	0.44	-0.09	-0.13	0.05	-0.02
portionpurchasesT	0.9	0.4	0.05	0.01	0.09	-0.12
ratiomaxsaleTotal	0.07	0.19	0.37	0.86	-0.31	0.02
ratiomaxbuyTotal	-0.44	-0.27	0.02	0.32	0.73	0.11
IndexExpenditurePerson	0.79	-0.55	-0.16	0.1	0	-0.06
indexexpoporhoga	0.8	-0.45	-0.01	-0.04	-0.04	0.15
IndexExpenditure Consumption	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
Looking for a job	-0.27	0.53	0.6	-0.07	-0.03	0.28

The factor loads that we have considered high have been greater than 0.7. We've colored the positive highs green and the negative highs red. The interpretations of each factor are as follows:

FACTOR 1: Autonomous Communities with a large population, a large internal and external economic capacity and a lot of expenditure by the population, families and companies.

FACTOR 2: Autonomous Communities with a small area of land.

FACTOR 3: Autonomous Communities with little industry and a large service sector.

FACTOR 4: Autonomous Communities with a high maximum volume of sales to another Autonomous Community.

FACTOR 5: Autonomous Communities with a large maximum volume of purchases from another Autonomous Community.

FACTOR 6: This factor explains each original variable too little, we need to see the solution rotated in case it improves the explanatory capacity of this factor.

4.4. VARIMAX ROTATED SOLUTION TEST

To perform the rotated Varimax solution, simply specify the *rotate="varimax"* attribute in the fa() *function*.

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
Salesto others	0.96	0.18	0	-0.01	0	0.21
ShoppingOther	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
MaxPurchase	0.36	0.12	-0.02	0.13	0.9	0.07
SalesRatioPurchases	0.19	-0.07	-0.32	-0.18	-0.17	0.78
proportionvtasT	0.96	0.18	0	-0.01	0	0.21
portionpurchasesT	0.96	0.2	0.02	0.13	0.06	0.06
ratiomaxsaleTotal	0.01	-0.05	0.05	1	0.04	-0.02
ratiomaxbuyTotal	-0.49	-0.07	-0.04	-0.06	0.81	-0.18
IndexExpenditurePerson	0.31	0.92	-0.09	0.02	0.04	-0.04
indexexpoporhoga	0.36	0.84	0.17	-0.03	-0.01	0.09
IndexExpenditure Consumption	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
Looking for a job	0.06	-0.64	0.59	0.19	0	0.05

The factor loads that we have considered high have again been greater than 0.7. We've colored the positive highs green and the negative highs red. The interpretations of each factor are as follows:

FACTOR 1: Autonomous Communities with a large population and internal and external economic capacity.

FACTOR 2: Autonomous Communities with a large expenditure by families and companies, and little agriculture.

FACTOR 3: Autonomous Communities with little industry.

FACTOR 4: Autonomous Communities with a large maximum sales volume to another Autonomous Community.

FACTOR 5: Autonomous Communities with a large maximum volume of purchases from another Autonomous Community.

FACTOR 6: Autonomous Communities with great economic capacity in terms of purchase and sale transactions.

4.5. COMPARISON OF PREVIOUS TRIALS

We first compared the rotated solution with the unrotated solution with the main axis method. It is clear that having selected one more factor makes it more complicated to interpret, but by rotating the solution we can interpret it more easily since the values are extrapolated. Therefore, we would be left with the rotated solution between these two.

Now, let's compare the "winning solution" so far which is the rotated main component solution with the rotated solution of the main axis method. While it is interesting that the second one gives us an explanatory factor of the industry, the sixth factor does not seem to have a very clear interpretation already having factors 4 and 5, so we continue to select the first, by the principal components method, since the rest of the factors are identical.

5. CLUSTERING BY THE HIERARCHICAL WARD METHOD

Before starting the clustering process, it is worth mentioning the relationship between the Mahalanobis distance and the principal components with Euclidean distance for clustering. Mahalanobis distance is a measure of distance that takes into account the covariance between variables. When the covariance matrix is diagonal, the Mahalanobis distance is reduced to the scaled Euclidean distance. In Principal Component Analysis (PCA), the matrix of covariance between the principal components is diagonal, meaning that the principal components are uncorrelated. However, when clustering, for each CP the distance depends on its own explained variance, so with the Euclidean distance we can see the CPs as non-anomalous, while with the Mahalanobis distance we could conclude that they are.

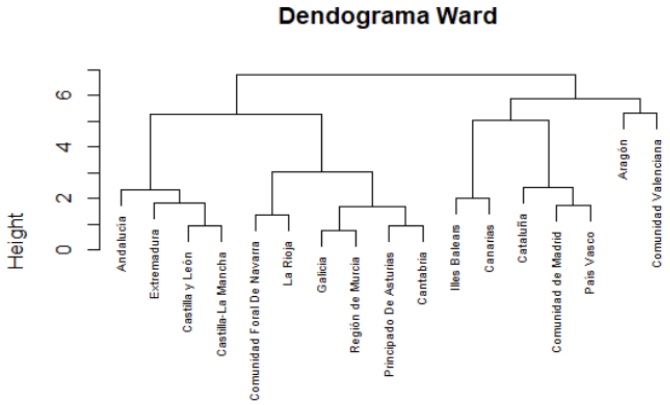
We save the rotated factorial solution of principal components (scores of the Autonomous Communities for each factor), which has been chosen, and change the names of the columns (factors) to a small interpretation with *colnames(solucion_factorial*). Once saved, we use it for clustering by the hierarchical Ward method with the *as.array*

code(scores%%mat.rotation,dimnames=C(FR1,FR2,FR3,FR4,FR5)))*, which are the rotated scores.

The Ward hierarchical method focuses on minimizing variance within groups, i.e., making them as homogeneous as possible within groups. It starts by considering each data point as an individual group and then gradually merges the groups closest to each other in a way that minimizes variability in the groups.

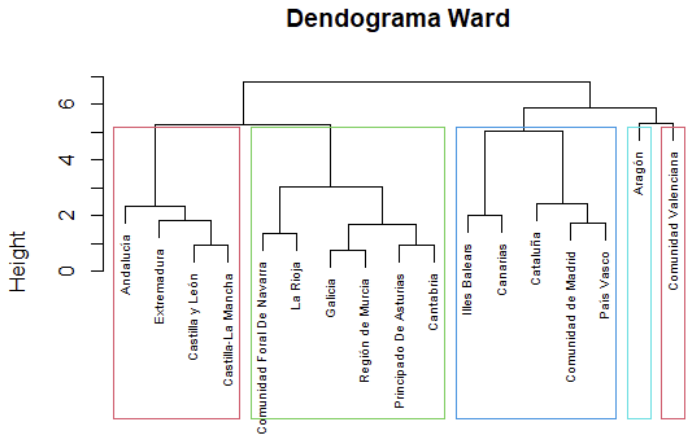
To perform the clustering we have to calculate the distances using the *dist(sol_factorial)* code, and then perform the usual clustering steps: first we use the *hclust() statement*, specifying the method which in our case is "ward. D". We look at the dendogram and according to the amount of internal homogeneity that the clusters lose in the process of grouping, we choose the number of groups to use.

We create the dendogram with *plot(cluster),* and it looks like this:

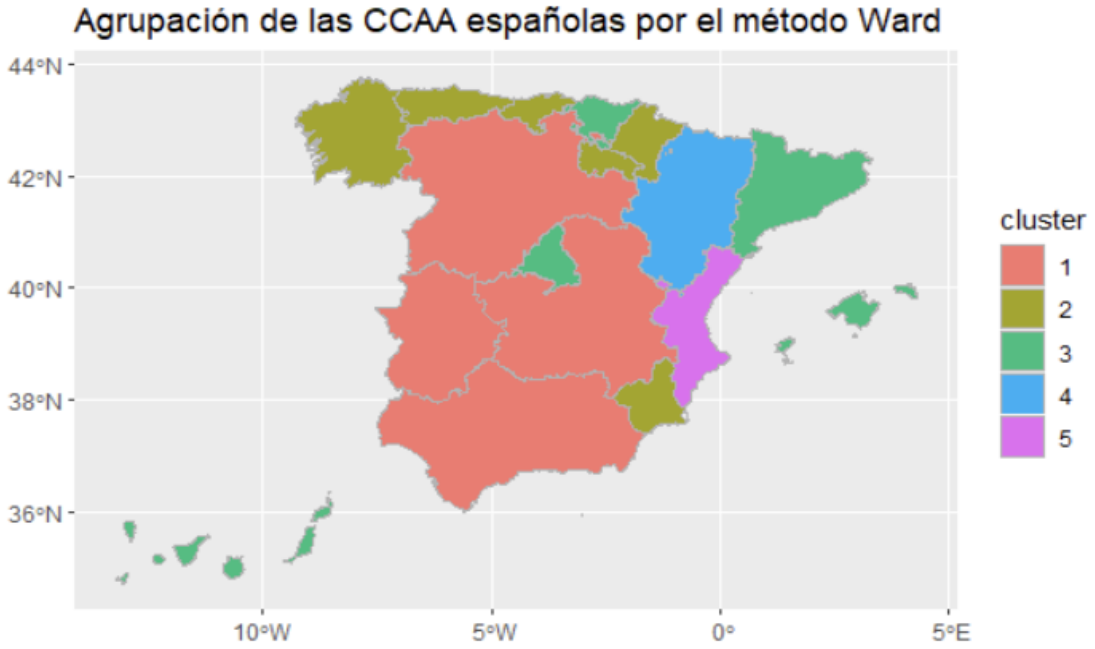


Looking at the dendogram we can see that with 5 groups you don't lose too much internal homogeneity while the number of clusters is greatly reduced.

Therefore, the solution of the clustering looks as shown in the second dendogram.



The isolation of the two Autonomous Communities that make up the last clusters is very interesting: Aragon and the Valencian Community. So we will keep in mind that when we analyze them we will be talking about specific groups that are different or strange enough that they have been exiled in different groups.



To study how they are characterized, we must perform an analysis of variance to find out which original factors and variables are the most differentiating in each group.

6. ANALYSIS OF VARIANCE ON ORIGINAL FACTORS AND VARIABLES

In order to study what differentiates each of the 5 groups, we look at the means and detect those that have a greater difference between groups within the same grouping. To do this, we use the *aggregate() function* with the previous model that we call Z, the variable with the assignment to groups that we have previously added to the factorial solution and using the *FUN=mean* function. The averages are as follows:

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

The most different groups when analyzing the means of each in each factor are precisely a different one for each cluster, that is, each factor characterizes a cluster in a more significant way.

To study how significant these differences are and their direction, we performed the analysis of variance. We first use ANOVA to find the most significant factors in the clustering with the *summary()* of the aov*(factor~cluster)* function. We found significant differences in factors 2, 4 and 5 (although 3 was close with a p-value of 0.118).

Having identified the factors with significant differences, we looked for which groups and their direction with a Scheffe test, which is the hardest to find differences, with the *ScheffeTest(anova)* function of the "DescTools" package. This analysis is useful when you have more than two groups and want to determine which of them are significantly different from each other, which we can't do with ANOVA. The results are as follows:

FACTOR 2 – HIGH POPULATION EXPENDITURE AND LOW AGRICULTURE
With population expenditure and little agriculture higher for 4 than for 1 at a significance level of 0.05.

FACTOR 4 – LOW MAXIMUM SALES VOLUME TO ANOTHER AUTONOMOUS COMMUNITY
With a maximum sales volume lower than any level of significance for group 5 compared to all other groups.

FACTOR 5 – LOW VOLUME OF PURCHASES MAXIMUM TO ANOTHER AUTONOMOUS COMMUNITY
With a maximum volume of purchases lower than any level of significance for all groups compared to group 2, except for group 1, which is the opposite. And with a maximum volume of purchases below a significance level of 0.1 in group 3 compared to group 1.

In this way we can get an idea of how each group is characterized:

GROUP 1: is characterized by having a lower maximum volume of purchases, suggesting a possible lower dependence on external debt or the need for significant imports. In addition, it shows a higher peak sales volume than Group 5, indicating a strong ability to sell products to other regions. This group exhibits lower population expenditures, suggesting a lower economic capacity, and a greater presence of agriculture, which could be associated with areas known as "the emptied Spain", more rural areas.

GROUP 2: shows a lower maximum volume of purchases than all groups except Group 1, indicating a lower dependence on external debt or the need for significant imports. In addition, it has a higher maximum sales volume than Group 5, suggesting a significant capacity to sell products to other Autonomous Communities. This group exhibits a higher expenditure of the population compared to Group 1 and a lower presence of agriculture, which places it in more urban and economically developed areas.

GROUP 3: has a lower maximum volume of purchases than Group 1, suggesting less dependence on external debt or need for significant imports. It shows a maximum sales volume higher than Group 1, indicating a capacity to sell products to other Autonomous Communities. Finally, this group has a lower presence of agriculture compared to Group 1.

GROUP 4: has a higher maximum volume of purchases than Group 5, which could indicate a greater dependence on imports or a more intense economic activity that involves purchases from other Autonomous Communities. The population's expenditures are higher than in Group 1, suggesting a higher economic capacity. In addition, this group has a lower presence of agriculture compared to Group 1.

GROUP 5: Characterized by a lower peak sales volume than all groups, suggesting a limited ability to sell products to other regions, possibly due to economic or structural constraints. No details are provided on population expenditures and presence of agriculture.

Now we perform the analysis of variance on the original variables. In summary, the averages indicate the following differences at first glance:

Propsuperf: highest for groups 1 and 2.
Pobsup ratio: much higher for groups 4 and 5.
Purchases of others: slightly lower for group 3.
MaxVenta: Highest for group 5.
MaxPurchase: Highest for group 2.
MaximumTotalSaleRatio: Much higher for group 5.
ratiomaxpurchaseTotal: slightly higher for group 2.
PersonExpenditure Index: Lower for group 1.
ExpenditureUnitConsumption Index: higher for group 4.
PIBPERCAP: Lower for group 1.
templeoAgri: higher for group 1.
tempServ: Highest for group 4.

Now, we perform the ANOVA test with these variables in which we have already detected differences. The variables with more or less significant differences were "PropSuperf", "Pobsup ratio", "maximumTotalSales ratio", "Consumption-unit expenditure index", "GDPpercap", "Agri employment" and "tempServ".

To find out the direction of the difference and in which groups it is found, we performed the Scheffe test, the results of which are the following:

PROPSUPERF
: With a surface area ratio greater than any significance level for group 1 compared to groups 3 and 4, and for group 5 at a significance level of 0.1.

RATIOPOBSUP: With population quantity according to area greater than a significance level of 0.1 for 4 compared to 1 and 3.

MAXIMUMTOTALSALE RATIO: With an importance of your biggest customer greater than any level of significance for group 5 compared to the rest of the groups.

PERSON-EXPENDITURE INDEX, CONSUMPTION-UNIT INDEX AND GDPPERCAP:
With an average expenditure per person/unit of consumption in relation to national expenditure/GDP per capita greater than a significance level of 0.1 for 4 compared to group 1.

TEMPLEOAGRI: With an employment rate in the agricultural sector below a significance level of 0.1 for group 4 compared to group 1.

TEMPSERV: With an employment rate in the service sector higher than any level of significance for group 4 compared to groups 1 and 3.

We can now interpret the five groups according to the results we have just obtained:

GROUP 1: It is characterized by having a higher proportion of surface area compared to groups 3 and 4, and 5. In addition, it exhibits a smaller amount of population according to the area than in Group 4. As for the importance of its largest customer, this is lower compared to Group 5. With regard to average expenditure per person/unit of consumption and GDP per capita, all three are lower compared to Group 4, so they are of less economic importance. In addition, the employment rate in the agricultural sector is higher for Group 1 compared to Group 4, so it must be in rather rural areas.

GROUP 2: it has an importance of its largest customer compared to Group 5, which could mean a lower economic dependence on its main customer (Autonomous Community). In the rest of the variables, there are no interesting differences.

GROUP 3: It is characterized by having a smaller proportion of surface area compared to group 1. In addition, an importance of its largest customer is lower compared to Group 5. Finally, it shows a lower employment rate in the services sector compared to Group 4, indicating less economic dependence on activities in the service sector.

GROUP 4: It is characterized by having a smaller proportion of surface area compared to group 1. In addition to a larger population amount according to the area compared to 1 and 3. In terms of average expenditure per person/unit of consumption and GDP per capita, all three are higher compared to Group 1, so they are of greater economic importance. A lower employment rate in the agricultural sector compared to group 1,

but higher employment rate in the services sector compared to group 3, indicating greater economic dependence on service activities.

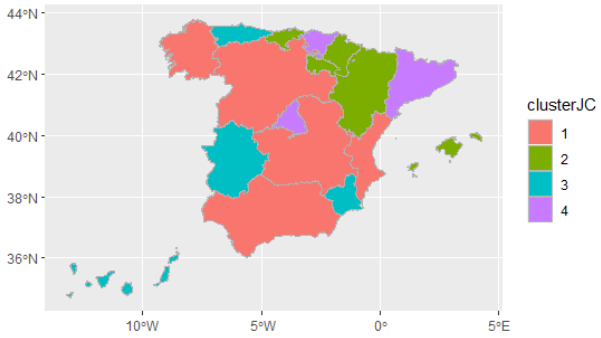
GROUP 5: It is characterized by having a smaller proportion of surface area compared to group 1. In addition, an importance of their larger customer compared to the rest of the groups, which could mean a greater economic dependence on their main customer.

Once we have interpreted the groups of our clustering, we can see how they differ from those we carried out previously in practice 3.

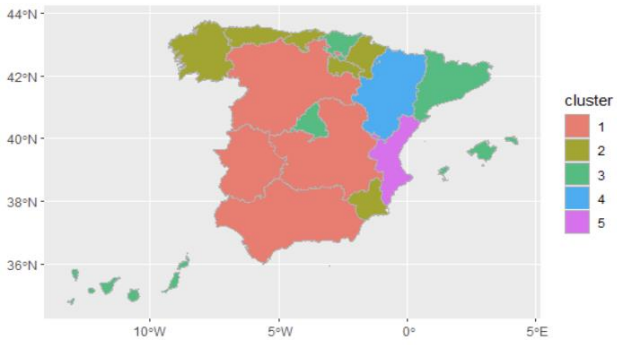
7. COMPARISON WITH RESULTS FROM PRACTICE 3

In practice 3 we ended up choosing another hierarchical method, the complete method, which resulted in this grouping (first graph):

Complete Method Grouping with Original Variables



Ward Method Grouping with Rotated CP Factors



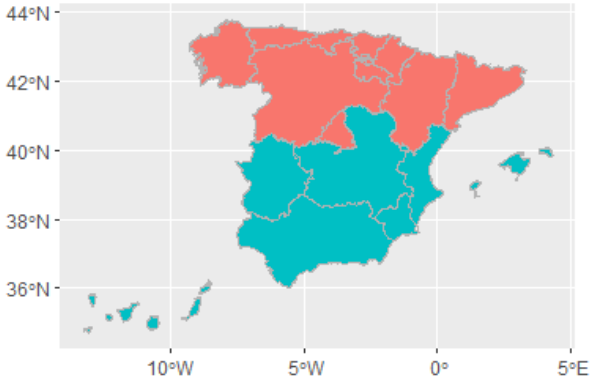
As we can see, in the previous practice we chose to do only 4 groups, while now we have chosen 5. Some features remain while others differ greatly.

To begin with, group 4 of the former is almost identical to group 3 of the latter, which have significant differences indicating that they are the economic powerhouses. In addition, both groups 1 are very similar, characterized by being agricultural, with a larger area, but with a lower concentration of population.

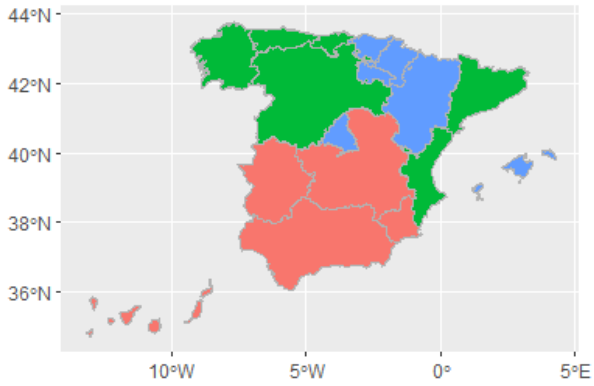
On the other hand, the rest of the groups are very disparate, and while in the first there were no isolated Autonomous Communities, in the second there have been two Autonomous Communities, Aragon on the one hand, and the Valencian Community on the other.

Some groupings made with different non-methodological criteria, by a single variable, are: by periphery or interior (RED=periphery, BLUE=interior), by own language or not (RED=no own language, BLUE=own language), by north or south (RED=North, BLUE=South), and according to the level of unemployment (RED= more than 15%, BLUE= less than 10%, GREEN= between 10% and 15%).

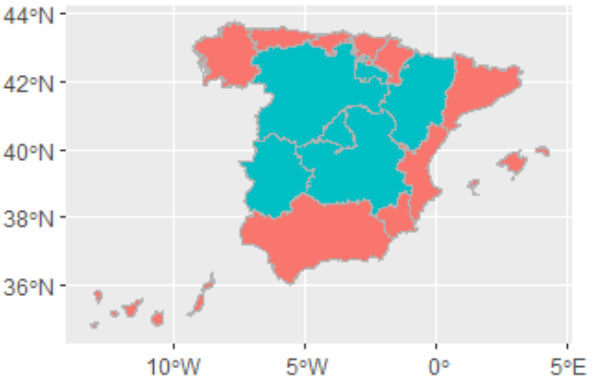
Grouping according to your location in North-South unemployment level



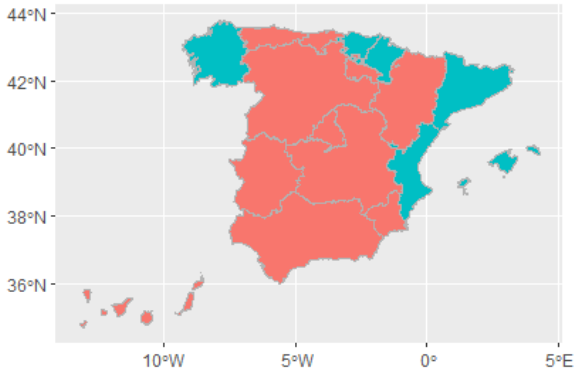
Grouping according to your



Grouping according to whether they are periphery or interior have their own language or not



Grouping according to whether they



These groupings differ greatly from those obtained in this study by Principal Components, perhaps we can see small similarities in the inland Autonomous Communities and group 1 without counting Aragon, Madrid and the Canary Islands, so that the fact that they are internal or peripheral may have an influence, but a more in-depth study would have to be carried out to find out.

What we can conclude from the fact that there are so many differences according to the method or criterion of grouping is that Spain is a very diverse country, and that it cannot be defined in a single sense, since it has a very diverse cultural and social wealth depending on the areas in which we are.

8. CONCLUSIONS

After an extensive factor analysis through different methods, we have concluded that the best one has been the test on obtaining principal components. Through a rigorous

pre-process and factor selection criteria, we have achieved a detailed interpretation of what explains the socioeconomic factors studied. The application of the Varimax rotation has served as a strategy to improve the interpretative clarity of the correlations. The factors, in order of the greatest to the lowest explanation of the total of the original variables, were the following:

FACTOR 1: When it is high, the Autonomous Community is a great economic power**.**
FACTOR 2: When it is high, the population has little expenditure and the Autonomous Community has little agriculture.
FACTOR 3: When it's high, there's a lot of buying and selling, little industry and a lot of service.
FACTOR 4: When it is high, the volume of the largest sale in the Autonomous Community is low.
FACTOR 5: When it is high, the volume of the largest purchase in the Autonomous Community is low.

This approach has made it easier for us to carry out subsequent analyses, such as the clustering of the Autonomous Communities based on the factors indicated, by Ward's hierarchical method. The grouping is as follows:

GROUP 1: Andalusia, Extremadura, Castilla y León and Castilla-La Mancha.
GROUP 2: Chartered Community of Navarre, La Rioja, Galicia, Region of Murcia, Principality of Asturias and Cantabria.
GROUP 3: Balearic Islands, Canary Islands, Catalonia, Community of Madrid and Basque Country.
GROUP 4: Aragon.
GROUP 5: Valencian Community.

The Autonomous Communities with a unique behaviour are Aragon and the Valencian Community, which have been grouped in isolation from the rest.

The original variables and the most differentiating factors according to the ANOVA after analyzing the means of the groups were "propsuperf", "Pobsup ratio", "maximumsaleTotal" ratio, "person-expenditure index", "consumption-unit expenditure index", "pibpercap", "templeoAgri", "tempServ", and the second, fourth and fifth factors. Through the study of the Scheffe test, we have been able to detect the direction and significance of the differences in each variable and factor. So the interpretation of each group is as follows:

GROUP 1: It stands out for its larger proportion of surface area, less dependence on external debt, the ability to sell to other regions and a significant presence of agriculture. It indicates a possible location in rural areas with less economic capacity.

GROUP 2: Shows less dependence on external debt, ability to sell products to other Autonomous Communities and higher expenditure by the population. It suggests a location in urban and economically developed areas.

GROUP 3: Characterized by a smaller proportion of surface area, the ability to sell to other Autonomous Communities and a lower presence of agriculture. It indicates a location with less economic dependence on service sector activities.

GROUP 4: Stands out for its greater economic importance, lower presence of agriculture and higher employment rate in the service sector. It suggests a location with intense economic activity and further development.

GROUP 5: It is characterized by a smaller proportion of surface area, greater dependence on its main customer and less capacity to sell to other regions. It indicates a possible greater economic dependence on your main customer.

BIBLIOGRAPHY

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 on coches.sav*. https://www.uv.es/mlejarza/datamine/faccoches2a.html

Wikipedia contributors. (2023, May 14). *Maximum verisimilitude*. Wikipedia, the free encyclopedia. https://es.wikipedia.org/wiki/M%C3%A1xima_verosimilitud

Interpret all statistics and graphs for Factor Analysis - 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