STUDY OF THE SOCIO-ECONOMIC GROUPING OF THE AUTONOMOUS COMMUNITIES OF SPAIN

DATA MINING IN BUSINESS

Ágatha del Olmo Tirado | 2nd BIA | 28/10/2023

VNIVERSITAT DE VALÈNCIA

BUSINESS INTELLIGENCE & ANALYTICS

INDEX

INTRODUCTION

This report presents a detailed analysis of the grouping of Spanish Autonomous Communities, based on socioeconomic data collected in the "comaut.csv" file. The aim of this study is to use both direct and hierarchical clustering techniques to group communities into homogeneous but heterogender groups, which will allow us to better understand the similarities and differences between the various Spanish regions.

We have used the direct techniques "farthest first" and "canopy" through the Weka program, and the hierarchical techniques "simple chaining", "full chaining", "Ward" and "centroid" in RStudio, as well as a "SOM" (Self-Organised Map) network grouping.

We have conducted a detailed analysis of the socioeconomic variables that could be influencing group formation, and we have explored the differences in the means of the

variables between the groups. After identifying the most differentiating means at first glance, we performed an analysis of variance (ANOVA) to identify significant differences between the groups in the most relevant variables, and a multiple analysis of these in order to interpret the results more effectively.

The results and conclusions derived from this analysis offer a deeper understanding of the diversity within the country and can be useful for decision-making and strategic planning in the area.

To track the work, we recommend seeing the link to the Rstudio script.

VARIABLE SELECTION AND MODIFICATION

An essential preliminary step is the correct selection of interesting variables for the project and the modification of these through the Rstudio program.

It is important to note that the autonomous cities of Ceuta and Melilla are not found in the database, since we do not consider that even if they are relativized, which would eliminate the effect of being so small in relation to the rest, they are interesting enough for the study taking into account their extraordinary situation.

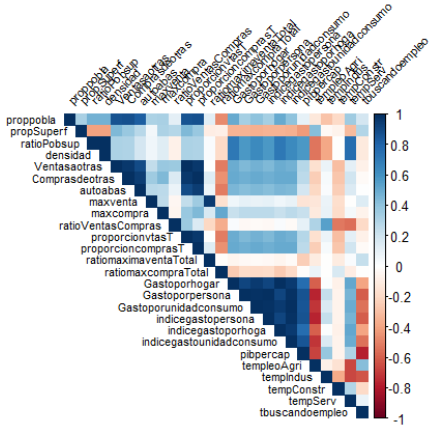
First, we set a seed -***set.seed(16)***- (to ensure the reproducibility of the data (especially considering that some clusters have random traits that we will discuss later).

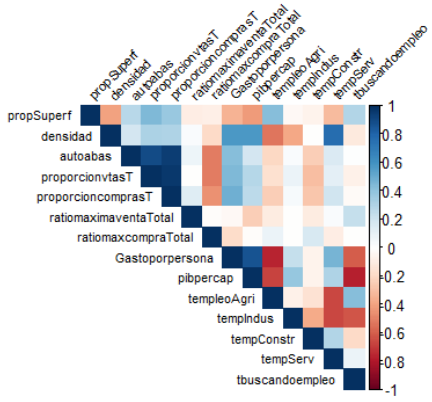
Later we load the libraries that we will need for the study, which in our case are ***haven*** for the import and export of files, ***kohonen*** for the subsequent creation of the SOM, ***dplyr*** for the manipulation of the data, ***corrplot*** for the visualization of the correlations, ***DescTools*** for the analysis of means, ***gridExtra*** for the visualization of plots and Ggplot2 for the creation of complex plots.

Once the libraries we are interested in have been loaded, we load the data in R with the ***< - read.csv2("./comaut.csv", dec = ",")*** statement (it is important to indicate the decimals in read.csv2, because in Excel it was opened with decimals by commas and R only understands them by periods). The dataset was then sorted in alphabetical order according to the variable Cautonoma, which contains the name of each autonomous community ***data <- data %>% arrange(Cautonoma)***.

With all the data loaded into R, we can now perform a study of the variables of the "data" dataframe. Two steps that we consider essential are the elimination of redundant variables and the selection or modification so that they have their range of values between 0 and 1 and not between -1 and 1 because their behavior is "better". It should be noted that we could use the Euclidean distance applied to the principal components or the Mahalanobis distance to "decorrelate" and type from the beginning, and even after eliminating redundancies we could directly use the normalized distance, but we will use the Euclidean distance, so we must eliminate high correlations and typify.

To see the correlation between variables, we first configure the data so that the variable "Cautonoma" becomes the name of the rows in the dataframe with the statement ***row.names(data)=data$Cautonoma*** since we need all variables to be numeric, we also save the variable as a ***cautonoma object <- data$Cautonoma*** to be able to delete it ***data$Cautonoma <- NULL*** but to be able to use it in the future. Now that we have the necessary database format we can create a correlation matrix with ***correlation_matrix <- cor(data)*** and visualize it through ***corrplot(correlation_matrix, method = "color", type = "upper", tl.col = "black", tl.srt = 45)***. So we can clearly see the values in the correlation matrix represented by a heat map, as we see in the first image.





The variables with high correlation and that we consider less interesting for the study are ratioPobsup, proppobla, Salestootras, Comprasdeotras, maxventa, maxcompra, ratioSalesPurchases, Expenditure per household, Expenditure per unitconsumption, indexexpenditureperson, indexexpenditure, indexexpenditureunitconsumption, and the matrix of correlations is as we see in the second image.

After eliminating redundancies avoiding multicollinearity, we must relativize the variables since depending on their scale some have more or less weight, (we do it for all variables that are not proportions, rates or ratios, i.e., self-sufficiency, expenditure per person, pibpercap and density) by assigning the following instruction ***datos_rel$variable<-data$variable/sum(data$variable)***.

Now, since we are using Euclidean distance, we type the variables to mean=0 and variance=1, with the datos_tipificados <- ***as.data.frame(scale(datos_rel))) statement***, and we can check that it has worked correctly with the mean and variance ***mean(datos_tipificados$variable)*** and ***var(datos_tipificados$variable),*** which should give respectively 0 and 1.

In this way we have already prepared the dataset so that the clustering analysis can be carried out correctly, so we save the new database in a csv file for use in Weka. To do this, we first add back the autonomous variable that we have previously saved as an object in the environment ***datos_save$Cautonoma <- cautonoma*** and save it to our directory using the ***write.csv(datos_save, file ="comaut_mod.csv", row.names = FALSE) statement***.

THE CLUSTERING PROCESS IN RSTUDIO

PREVIOUS TASKS

First, we make the SOM network, and then, we create the hierarchical groupings (Complete, Simple, Ward and Centroid), since we can then choose the most appropriate number of clusters, and then we perform the SOM clustering.

With the dataset already modified, we create the model matrix that we call Z without including either "Cautonoma" or the intercept (for this we use -1 at the beginning). ***Z=model.matrix(~-1+density+propSuperf+autoabas+proportionsT+proportionpurchasesT+maximumsaleratioTotal+ratiomaxpurchaseTotal+expenditureperperson+pibpercap+templeoAgri+tempIndus+tempConstr+tempServ+tseekingemployment, datos_tipificados).***

As in the database datos_tipificados the rows were the Autonomous Communities, in the matrix they are also the Autonomous Communities, so there is no need to tell you.

CREATION OF THE SOM

To create the self-organizing map, we create the training set, ***data_train*** <- ***datos_tipificados[, c(1:14)]***, taking all the data we have, and converting it to matrix ***data_train_matrix*** <- ***as.matrix(data_train***). Then, we indicate the number of profiles we want to obtain, which in our case is 9, with a 3x3 in hexagonal shape so that there are more neighbors per profile ***som_grid<- somgrid(xdim = 3, ydim=3, topo="hexagonal")***. In this way, we can already carry out the SOM itself with the instruction

```
som_model <- som(data_train_matrix,
        grid=som_grid,
        rlen=1000,
        alpha=c(0.05,0.01),
        keep.data = TRUE )
```

We can analyze different network metrics. To begin with, we can see which profile each individual fell into, using "unit.classif", which tells us in order the number of the profile, we can make a for that gives us the information in a more compact way:

```
for (i in 1:length(unique(som_model$unit.classif))) {
 profile <- unique(som_model$unit.classif)[i]
 cat("Profile:", profile, "\n")
```

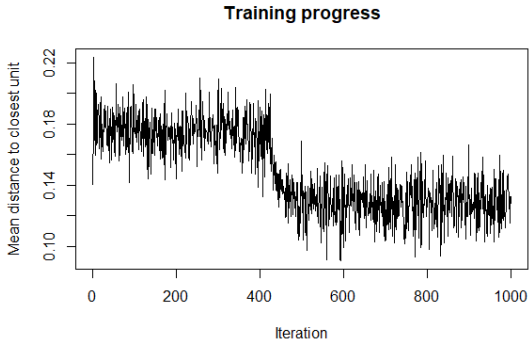
cat("Autonomous Communities:", cautonoma[som_model$unit.classif == profile], "\n\n")

}

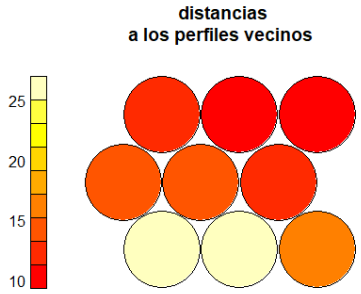
The output is as follows:

Profile: 3
Autonomous Communities: Andalusia Castilla y León
Profile: 7
Autonomous Communities: Aragon, Cantabria, Balearic Islands
Profile: 9
Autonomous Communities: Canary Islands Extremadura Region of Murcia
Profile: 6
Autonomous Communities: Castilla-La Mancha
Profile: 1
Autonomous Communities: Catalonia Community of Madrid
Profile: 4
Autonomous Communities: Chartered Community of Navarre Basque Country
Profile: 2
Autonomous Communities: Valencian Community
Profile: 5
Autonomous Communities: Galicia
Profile: 8
Autonomous Communities: La Rioja Principality of Asturias

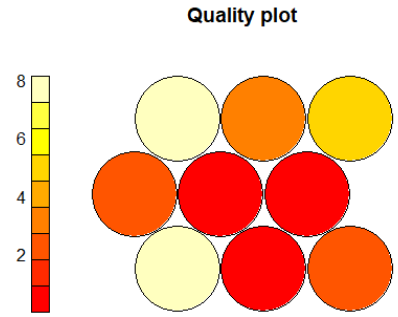
We can also see the progress of network training with the ***plot(som_model, type="changes")*** statement. The X-axis represents iterations during training, while the Y-axis shows the average distance to the nearest codebook vector. This plot represents the average distance between each input data and its corresponding weight vector. In the network, each weight vector is adjusted during the training process to get closer to the input data, so as the number of iterations increases, we would expect the distance as a whole to decrease. Here we see a very large drop from 400 iterations, but from there it reaches a minimum from which it will not be able to go down because the network does not manage to improve further, in this case it seems to be around the value 0.12.



We can also see the sum of distances between profiles to immediate neighbors with the statement ***plot(som_model, type="dist.neighbours", main = " distances to the***



neighboring profiles")tags. Not only is the network competitive in the sense that only "one neuron wins," but it is also competitive in the sense that it is not only competitive.
also cooperative, since every time a profile updates its weights, so do its profiles
neighbors, that's why it's interesting to use the form hexagonal, which maximizes the number of
Neighbors. In this plot we can see that profiles 6, 7 and 8 (from left bottom to right up) are
the most similar to each other, with 1 and 2 being the

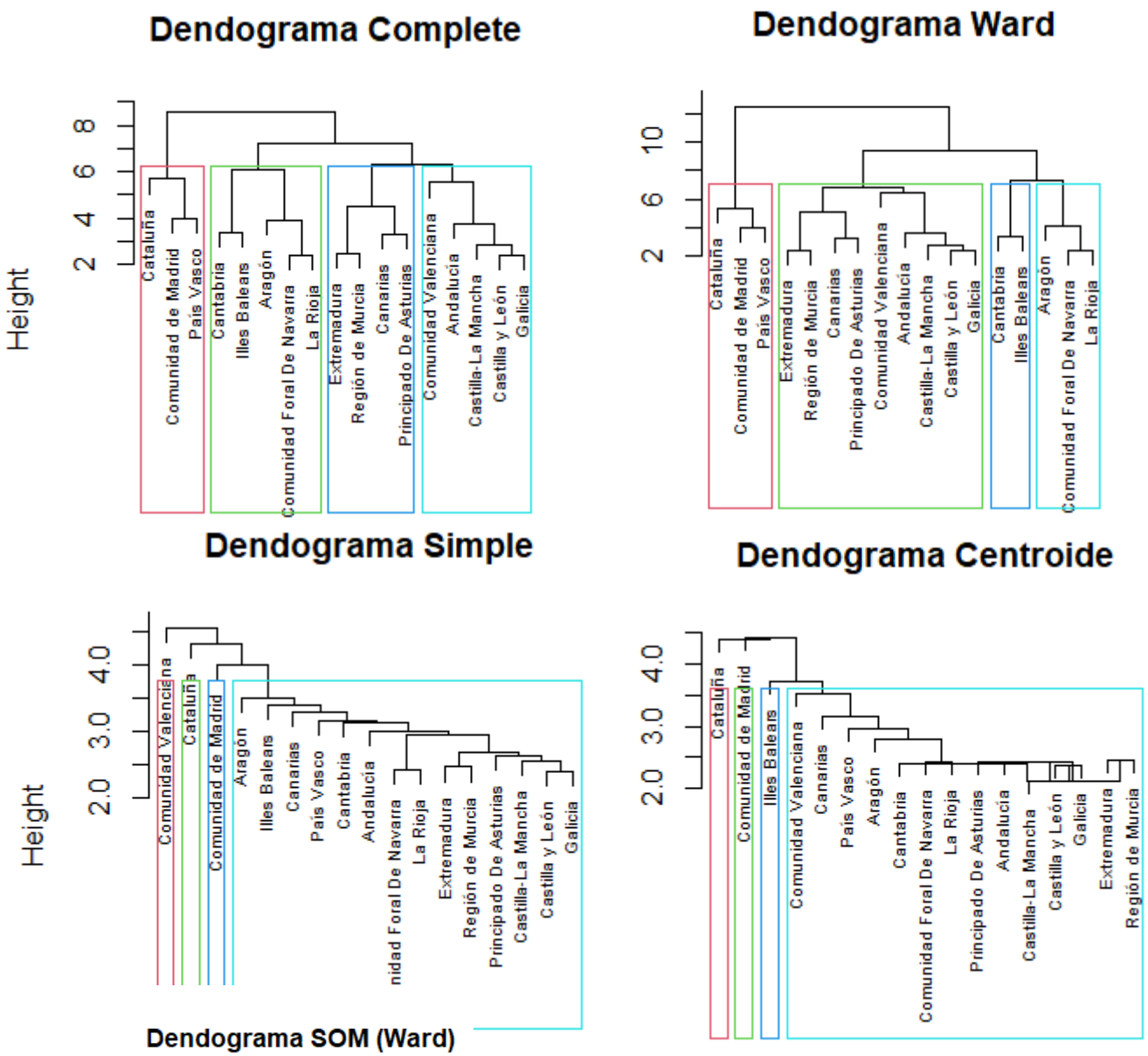


In addition, we can see "quality", which shows us the average distance of the instances assigned to the node to each profile. The smaller these distances, the better the instances will be represented by the profile to which they have been assigned.
As we can see, the nodes that most resemble their instances are 2, 5, and 6. (You can go back to the previous page to see which Autonomous Communities are in each of the nodes).

Once the network and its characteristics have been analyzed, we can group the hierarchical and SOM and choose a number of clusters.

SETTING UP THE GROUPING

The hierarchical clusterers that we are going to make are the following:



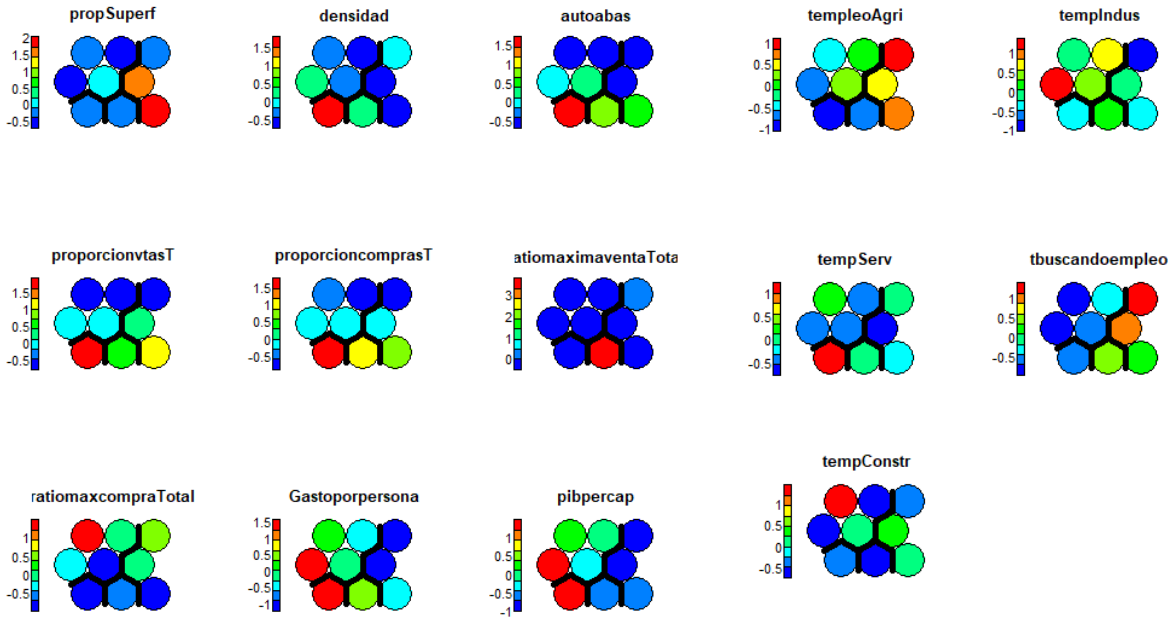
For all of us we use the ***clusX<- hclust(dist(Z), method= "X")***, and then ***plot(clusX, cex = 0.6)*** to see it and be able to choose a cluster cardinality at the end.

The Y-axis represents the loss of internal homogeneity, and we are not interested in seeing very large jumps, so we believe that separating the clusters into 4, since it does not lose too much homogeneity even if the groups are greatly

reduced. To clearly see how they are separated, add ***rect.hclust(clusJCe, k = 4, border = 2:5).***

Once the cardinality has been chosen, we can run clusterX <- cutree(clusX, k = 4) for everyone, and we can analyze what characterizes each group in the SOM with the instructions:

```
par(mfrow = c(3, 3))

plot_list <- list()

for (i in 1:14) {

  current_plot <- plot(som_model, type = "property", property =
getCodes(som_model)[, i],

                       main = colnames(getCodes(som_model))[i], palette.name =
coolBlueHotRed)

  add.cluster.boundaries(som_model, clusterSOM)

  plot_list[[length(plot_list) + 1]] <- current_plot

}
```



As we can see in the graphs, the most differentiating groups in the variables are:
-**propSuperf**: Group 4, with the exception of Profile 9, exhibits the highest values
-**density**: Profile 1 significantly outperforms the others in terms of density
-**autoabas**: Again, Profile 1 stands out as the highest
-**T ratios** and **T purchase ratios**: Profiles 1 and 2 have the highest values
-**maximumsaleTotalratio**: In this variable, Profile 2 is the most prominent
-**ratiomaxpurchaseTotal**: Clusters 1 and 2 show the lowest values
-**expenditure per person**: Cluster 3 registers the lowest value

-**pibpercap**: Clusters 1 and 4 have the highest values
-**templeoAgri, tempConstr and Tbuscaempleo**: Cluster 3 outperforms all others in t hese variables
-**tempServ**, the first profile is the highest by far

THE CLUSTERING PROCESS AT WEKA

We started using Weka, a graphical environment that allows us to create and analyze experiments on clustering tasks, and specifically its Explorer interface. Here we perform the groupings with the FarthestFirst and Canopy direct methods.

PREVIOUS TASKS

The starting point is the " comaut_mod.csv" file, which we have just saved and contains socio-economic information on the 17 autonomous communities in the Valencian metropolitan area. When working with this file in CSV format, certain initial adjustments need to be made. To do this, open the "Invoke Options Dialog" dialog and change the field separator from "," to ";" to make sure the data is imported correctly. Since we're looking to group communities together and not classify them, we've set the class to "No Class" so that a class attribute isn't included.

SETTING UP THE GROUPINGS

The groupings that we have selected are grouped as follows:

-FarthestFirst: randomly chooses an element from the data as the first centroid and calculates its distance from the rest. The furthest point becomes the new controller of the dataset and repeats the distance calculation, and makes assignments until it has the necessary groupings.

-Canopy: the first centroid is randomly fixed and provisional groups are generated that we call canopies according to thresholds, and the elements that are not in any canopy in common are candidates for centroids. Afterwards, a classification of the K-means algorithm is performed, but the distance between the elements of the same canopy is not calculated.

To add the groupings we go to Preprocess and select the filter within "unsupervised" since clustering is unsupervised (it does not have a response variable since there are no previous classes in which to classify), and within "attribute" we select "AddCluster", and right-click on it entering "Show properties". There, we choose the Canopy clusterer, select the attribute to ignore, which is Cautonoma, and right-click on the clusterer to set the number of clusters to 4. After applying it, we use the "RenameAttribute" filter and change it to clusterCanopy so that we can then identify it correctly and to allow the creation of the next grouping. We repeat the process for FarthestFirst, but this time we also ignore the cluster we just made, and rename it to clusterFF.

Once the groupings are done, we can save the database with save as comaut_directos.csv to be able to import it to R with the instruction below.

datos_dir <- read.delim("./comaut_directos.csv", sep=",")

CLUSTER ANALYSIS

ELEMENTS OF EACH GROUPING

We analyze the elements of the direct groupings since we have the hierarchical ones in the dendograms of the past.

FARTHEST FIRST:

GROUP 1: Valencian Community, Galicia, Aragon, Cantabria, Navarre, Balearic Islands, La Rioja, Canary Islands, Extremadura, Principality of Asturias and Region of Murcia.

GROUP 2: Catalonia

GROUP 3: Community of Madrid and Basque Country

GROUP 4: Andalusia, Castilla y León and Castilla-La Mancha

CANOPY

GROUP 1: Valencian Community, Galicia, Aragon, Balearic Islands, Principality of Asturias, Region of Murcia, Basque Country and Castilla y León

GROUP 2: Canary Islands, Extremadura, Andalusia and Castilla-La Mancha

GROUP 3: Catalonia and the Community of Madrid

GROUP 4: Autonomous Community of Navarre and La Rioja

As we can see, the most characteristic is the marginalization of Catalonia, Madrid, the Basque Country and the Valencian Community, so we will take into account when analyzing their averages that we will be talking about specific groups and sufficiently different or strange that they have been exiled in different groups.

MEAN ANALYSIS

In order to study what differentiates each group, 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 we call Z, the corresponding model, and with FUN=mean. When we analyzed them, we detected differences in the means of the following variables:

-Simple hierarchical: density, self-sufficiency, sales ratioT, purchasing ratioT, totalsalemax ratio, maximumpurchaseTotal, expenditure per person, pibpercap, tempServ

-Complete hierarchy: density, self-sufficiency, sales ratioT, purchasing ratioT, maximumpurchaseTotalRatio, expenditure per person, GDPpercap, job seeking.

-Hierarchical Ward: density, self-sufficiency, sales ratioT, purchasing ratioT, maximumpurchaseTotal, expenditure per person, pibpercap, employmentAgri, tempIndus, tempConstr, tempServ, job seeking.

-Hierarchical Centroid: density, self-sufficiency, sales ratioT and purchase ratioT, maximumpurchaseTotalRatio, pibpercap, templeoAgri, tempIndus, tempconstr and tempServ.

-Canopy: density, self-sufficiency, sales ratio, purchasing ratio, expenditure per person, pibpercap, templeoagri, tempindus, tbuscando empleo empleo

-FarthestFirst: density, self-sufficiency, sales ratioT, purchasing ratioT, maximumpurchaseTotalRatio, expenditure per person, pibpercap, employmentAgri, tempServ, jobseekingJob

-SOM: density, self-sufficiency, sales ratio, purchasing ratio, totalsale, expenditure per person, pibpercap, employmentAgri, tempServ, employment, employment, employment, employment, employment

ANOVA ANALYSIS

Analysis of variance (ANOVA) is a statistical technique used to assess whether there are significant differences between groups based on a variable or factor.

Specifically, in this case, we will detect significant differences between the groups of Spanish Autonomous Communities in the variables in which we have found apparent differences between the means of the groups (previous section) with the *aov function (variable-cluster)*. To do this, we have first converted the cluster variables to factor with *as.factor()* because in this case it refers to discrete categories that represent different groups, and does not have a continuous numerical meaning. Now ANOVA will treat them as levels and not as variables with a numerical interpretation. The results have been, in summary, as follows:

Variable \ Método	Jer. Centroide	Jer. Complete	Jer. Simple	Jer. Ward	Jer. Canopy	FarthestFirst	SOM
Densidad	Todos los niveles	0.01	Cualquier nivel	0.01	0.01	0.001	0.01
Autoabas	0.001	0.01	Cualquier nivel	0.05	0.05	Cualquier nivel	0.05
ProporcionventasT	0.001	Cualquier nivel	0.001	0.01	0.001	Cualquier nivel	0.01
ProporcioncomprasT	0.01	0.001	0.001	0.01	0.01	Cualquier nivel	0.01
RatiomaxcompraTotal	-	0.01	Cualquier nivel	0.05	-	-	Cualquier nivel
Gastoporpersona	-	0.001	-	0.001	0.001	0.01	0.001
Pibpercap	-	Cualquier nivel	-	Cualquier nivel	Cualquier nivel	0.001	0.001
TempleoAgri	-	0.001	-	0.05	-	-	0.01
TempIndus	-	0.01	-	0.01	0.01	-	No
TempConstr	0.01	Cualquier nivel	Cualquier nivel	Cualquier nivel	Cualquier nivel	Cualquier nivel	-
TempServ	-	0.05	-	0.05	No	No	No
Tbuscandoempleo	-	0.001	-	0.001	Cualquier nivel	-	0.001

MULTIPLE COMPARISONS

Finally, we performed the Scheffe analysis, which is the toughest as it finds significant differences, with the ScheffeTest(anova) function on the variables in which we have detected significant differences. 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. In this way we will be able to understand what really differentiates the groups in each grouping.

HIERARCHICAL CENTROID

Variable	Comparación Grupos	Nivel de Significación	Sentido de la Diferencia
Densidad	1 vs 3	6.3e-05	Grupo 3 > Grupo 1
Densidad	2 vs 3	0.0061	Grupo 3 > Grupo 2
Densidad	3 vs 4	0.0054	Grupo 3 < Grupo 4
Autoabas	1 vs 2	0.0022	Grupo 2 > Grupo 1
Autoabas	2 vs 3	0.0317	Grupo 3 < Grupo 2
Autoabas	2 vs 4	0.0084	Grupo 4 < Grupo 2
ProporcionvtasT	1 vs 2	0.0190	Grupo 2 > Grupo 1
ProporcionvtasT	2 vs 4	0.0181	Grupo 4 < Grupo 2
ProporcioncomprasT	1 vs 2	0.0264	Grupo 2 > Grupo 1
ProporcioncomprasT	2 vs 4	0.0426	Grupo 4 < Grupo 2
TempConstr	1 vs 4	0.0129	Grupo 4 > Grupo 1
TempConstr	2 vs 4	0.0447	Grupo 4 > Grupo 2
TempConstr	3 vs 4	0.0773	Grupo 4 > Grupo 3
TempServ	1 vs 3	0.0501	Grupo 3 > Grupo 1
TempServ	1 vs 4	0.0930	Grupo 4 > Grupo 1

Group 1: This group has a lower density and level of self-sufficiency (Autobaas) compared to other groups, which could indicate a lower population concentration and lower domestic production capacity. In addition, their buy and sell ratios (ProportionvtaasT and ProportioncomprasT) are lower compared to group 2, which could suggest a lower economic capacity.

Group 2: This group has a lower density than group 3, but a higher level of self-sufficiency than group 1, which could indicate higher domestic production capacity. Their buy and sell ratios are higher than those of group 1, which could indicate greater economic capacity. However, these indices are lower compared to group 4, suggesting that their economic capacity is lower than that of group 4.

Group 3: This group has the highest density compared to groups 1 and 2, which could indicate a high population concentration. Its level of self-sufficiency is higher than that of group 2, which could indicate a high domestic production capacity.

Group 4: This group has the highest density compared to group 3, which could indicate a very high concentration of population. Their level of self-sufficiency is lower than that

of group 2, but their buying and selling indices are higher, which could indicate high economic capacity.

HIERARCHICAL COMPLETE

Variable	Comparación Grupos	Nivel de Significación	Sentido de la Diferencia
Densidad	1 vs 4	0.0399	Grupo 4 > Grupo 1
Densidad	2 vs 4	0.0450	Grupo 4 > Grupo 2
Densidad	3 vs 4	0.0961	Grupo 4 > Grupo 3
Autoabas	4 vs 2	0.0426	Grupo 4 > Grupo 2
Proporción Ventas T	1 vs 2	0.0320	Grupo 1 > Grupo 2
Proporción Ventas T	1 vs 3	0.0381	Grupo 1 > Grupo 3
Proporción Ventas T	2 vs 4	0.0034	Grupo 4 > Grupo 2
Proporción Ventas T	3 vs 4	0.0042	Grupo 4 > Grupo 3
Proporción Compras T	1 vs 2	0.0731	Grupo 1 > Grupo 2
Proporción Compras T	1 vs 3	0.0826	Grupo 1 > Grupo 3
Proporción Compras T	2 vs 4	0.0068	Grupo 4 > Grupo 2
Proporción Compras T	3 vs 4	0.0082	Grupo 4 > Grupo 3
Ratio Max Compra Total	1 vs 2	0.0690	Grupo 1 > Grupo 2
Ratio Max Compra Total	2 vs 4	0.0982	Grupo 4 > Grupo 2
Gasto por Persona	1 vs 4	0.0192	Grupo 4 > Grupo 1
Gasto por Persona	2 vs 4	0.1187	Grupo 4 > Grupo 2
Gasto por Persona	3 vs 4	0.0082	Grupo 4 > Grupo 3
PIB per Capita	1 vs 4	0.00015	Grupo 4 > Grupo 1
PIB per Capita	3 vs 4	0.00011	Grupo 4 > Grupo 3
T Buscando Empleo	1 vs 2	0.0537	Grupo 1 > Grupo 2
T Buscando Empleo	2 vs 3	0.0034	Grupo 3 > Grupo 2
T Buscando Empleo	2 vs 4	0.9628	Grupo 4 > Grupo 2
T Buscando Empleo	3 vs 4	0.0212	Grupo 4 > Grupo 3

Group 1: This group is characterized by having a significantly lower density compared to Group 4, suggesting a different spatial distribution of variables. In addition, it has a significantly higher proportion of total sales than Group 2 and Group 3, indicating that sales play a more prominent role in its economic activity. It also exhibits significantly higher expenditure per person and GDP per capita compared to Group 4, suggesting greater economic capacity.

Group 2: In contrast, this group does not show significant differences in density or proportions of purchases or sales compared to other groups. However, it stands out for a significantly lower job-seeking rate than Group 3. Expenditure per person and GDP per capita do not differ significantly from other groups.

Group 3: Group 3 is characterized by a significantly higher density compared to Group 4, but lower proportion of total sales, expenditure per person, and GDP per capita. In addition, it has a significantly higher rate of job seekers than Group 2 and Group 4.

Group 4: This group stands out for having the highest density and significantly greater self-sufficiency than Group 2. Although it shows lower proportions of total sales and purchases compared to Group 1 and Group 2, it has significantly higher per capita expenditure and GDP per capita than Group 2 and lower than Group 1 and Group 3. The job search rate is significantly lower than Group 3 and higher than Group 2.

SIMPLE HIERARCHICAL

Variable	Comparación Grupos	Nivel de Significación	Sentido de la Diferencia
Densidad	1 vs 4	7.4e-05	Grupo 4 > Grupo 1
Densidad	2 vs 4	0.0068	Grupo 4 > Grupo 2
Densidad	3 vs 4	0.0050	Grupo 4 > Grupo 3
Autoabas	1 vs 2	0.0009	Grupo 2 > Grupo 1
Autoabas	1 vs 4	0.0199	Grupo 4 > Grupo 1
Autoabas	2 vs 4	0.0199	Grupo 4 > Grupo 2
Proporción Ventas	1 vs 2	0.0175	Grupo 2 > Grupo 1
Proporción Compras	1 vs 2	0.0117	Grupo 2 > Grupo 1
Ratio Máxima Venta Total	1 vs 3	1.2e-08	Grupo 3 > Grupo 1
Ratio Máxima Venta Total	2 vs 3	1.9e-07	Grupo 3 > Grupo 2
Ratio Máxima Venta Total	3 vs 4	2.4e-07	Grupo 3 > Grupo 4

Group 1: This group has a lower density compared to group 4, which could indicate a lower population concentration. Their level of self-sufficiency is also lower than groups

2 and 4, which could reflect lower domestic production capacity. In addition, their sales and purchase ratios are lower than group 2.

Group 2: Although this group has a lower density compared to groups 3 and 4, its level of self-sufficiency is higher than that of group 1, indicating a higher domestic production capacity. Their sales-to-purchase ratios are higher than those of Group 1 but do not compare with Groups 3 or 4.

Group 3: This group stands out for having the highest total maximum ratio, surpassing groups 2 and 4 in this aspect. Although its density is lower than that of Group 4, it is still higher than that of Group 2. This could indicate a balance between population concentration and economic efficiency.

Group 4: With the highest density among all groups, this indicates a high population concentration. Although its level of self-sufficiency is lower than that of Group2, its proportions in sales and purchases are not adversely affected, maintaining a solid economic balance.

HIERARCHICAL WARD

Variable	Comparación Grupos	Nivel de Significación	Sentido de la Diferencia
Densidad	1 vs 4	0.0288	Grupo 4 > Grupo 1
Densidad	2 vs 4	0.0373	Grupo 4 > Grupo 2
Densidad	3 vs 4	7.4e-05	Grupo 4 > Grupo 3
Autoabas	1 vs 4	0.2547	Grupo 4 > Grupo 1
Autoabas	2 vs 4	0.1425	Grupo 4 > Grupo 2
Autoabas	3 vs 4	0.1603	Grupo 4 > Grupo 3
Proporción Ventas	1 vs 4	0.1310	Grupo 4 > Grupo 1
Proporción Ventas	2 vs 4	0.0385	Grupo 4 > Grupo 2
Proporción Ventas	3 vs 4	0.0639	Grupo 4 > Grupo 3
Proporción Compras	1 vs 4	0.1351	Grupo 4 > Grupo 1
Proporción Compras	2 vs 4	0.0639	Grupo 4 > Grupo 2
Gasto por Persona	1 vs 4	0.0069	Grupo 4 > Grupo 1
PIB per Cápita	1 vs 2	0.0011	Grupo 2 > Grupo 1
PIB per Cápita	1 vs 4	1.6e-05	Grupo 4 > Grupo 1
PIB per Cápita	3 vs 4	0.0171	Grupo 4 > Grupo 3
Tasa de Empleo en Industria	1 vs 2	0.0393	Grupo 2 > Grupo 1

Group 1: This group has a lower density than group 4, which could indicate a lower population concentration. Its level of self-sufficiency is also lower than group 4, which could reflect lower domestic production capacity, but its GDP is higher than 2 and 4 and it has a higher rate of employment in industry than group 2, indicating high economic capacity.

Group 2: Although this group has a lower density compared to group 4, its level of self-sufficiency is higher than that of group 1, indicating a higher domestic production capacity.

Group 3: This group has a lower density compared to group 4. Although its level of self-sufficiency is lower than that of group 4, this does not seem to negatively affect its sales and purchase shares, especially considering that it has a higher GDP than group 4.

Group 4: This group has the highest density among all groups, indicating a high concentration of population. Although their level of self-sufficiency is lower than that of group 2, their sales and purchase shares are not adversely affected, indicating a solid economic balance.

CANOPY

Variable	Comparación Grupos	Nivel de Significación	Sentido de la Diferencia
Densidad	3 vs 1	0.0366	Grupo 3 > Grupo 1
Densidad	3 vs 2	0.0443	Grupo 3 > Grupo 2
Densidad	4 vs 3	0.0551	Grupo 4 > Grupo 3
Autoabas	3 vs 1	0.1394	Grupo 3 > Grupo 1
Autoabas	3 vs 2	0.2273	Grupo 3 > Grupo 2
Proporción Ventas	3 vs 1	0.0404	Grupo 3 > Grupo 1
Proporción Ventas	3 vs 4	0.0452	Grupo 3 > Grupo 4
Proporción Compras	3 vs 1	0.0954	Grupo 3 > Grupo 1
Proporción Compras	3 vs 4	0.0561	Grupo 3 > Grupo 4
Gasto por Persona	2 vs 1	0.0699	Grupo 2 > Grupo 1
Gasto por Persona	3 vs 2	0.0062	Grupo 3 > Grupo 2
PIB per Cápita	3 vs 1	0.0731	Grupo 3 > Grupo 1
PIB per Cápita	3 vs 2	0.0374	Grupo 3 > Grupo 2
PIB per Cápita	3 vs 4	0.7773	Grupo 3 > Grupo 4
Tasa de Empleo en Industria	4 vs 2	0.0148	Grupo 4 > Grupo 2
Tasa de Empleo en Industria	4 vs 3	0.1270	Grupo 4 > Grupo 3
Tasa de Empleo en Industria	4 vs 1	0.00088	Grupo 4 > Grupo 1
En búsqueda de empleo	2 vs 1	0.00104	Grupo 2 > Grupo 1
En búsqueda de empleo	3 vs 2	0.00822	Grupo 3 > Grupo 2
En búsqueda de empleo	4 vs 2	0.00088	Grupo 4 > Grupo 2
En búsqueda de empleo	4 vs 3	0.71930	Grupo 4 > Grupo 3

Group 1: This group has a higher density than groups 2 and 3, but lower than group 4. Although its level of self-sufficiency is lower than that of group 3, it outperforms group 4 in this respect. In terms of the proportion of sales and purchases, it is below group 3 but above group 4. Group 1 has a lower per capita expenditure than Group 2 but higher than Group 3. Its GDP per capita is lower than that of groups 2 and 3, and its employment rate in industry is the lowest among all groups.

Group 2: This group is characterized by having a lower density than groups 1 and 4, but higher than group 3. Although its level of self-sufficiency is lower than that of Group 1, it surpasses Groups 3 and 4 in this respect. In terms of the proportion of sales and purchases, it is below Group 1, but exceeds Groups 3 and 4. Group 2 has the highest expenditure per person among all groups. Its GDP per capita is higher than those of Groups 1 and 4, and its employment rate in industry is higher than that of Group 1, but lower than that of Group 4.

Group 3: This group has a lower density than group 1 but higher than group 2. Their level of self-sufficiency is higher than that of group 4 but lower than that of group 2. In terms of the ratio of sales to purchases, it outperforms all other groups. Group 3 has a lower per capita expenditure than Group 2 but higher than Group 1. Its GDP per capita is the highest of all groups, and its employment rate in industry is higher than that of Group 2.

Group 4: This group has a higher density than any other group, although their level of self-sufficiency is lower than that of group 2, their sales and purchase ratios are not adversely affected, indicating a strong economic equilibrium.

FARTHESTFIRST

Variable	Comparación Grupos	Nivel de Significación	Sentido de la Diferencia
Densidad	3 vs 1	0.0056	Grupo 3 > Grupo 1
Densidad	4 vs 3	0.0065	Grupo 4 > Grupo 3
Autoabas	2 vs 1	0.0004	Grupo 2 > Grupo 1
Autoabas	3 vs 2	0.0078	Grupo 3 < Grupo 2
Autoabas	4 vs 2	0.0061	Grupo 4 < Grupo 2
Autoabas	4 vs 3	0.9990	Grupo 4 = Grupo 3
Proporción Ventas	2 vs 1	0.00076	Grupo 2 > Grupo 1
Proporción Ventas	3 vs 1	0.05910	Grupo 3 > Grupo 1
Proporción Ventas	3 vs 4	0.01895	Grupo 3 > Grupo 4
Proporción Compras	2 vs 1	0.0026	Grupo 2 > Grupo 1
Proporción Compras	3 vs 2	0.1204	Grupo 3 < Grupo 2
Proporción Compras	3 vs 4	0.0663	Grupo 3 < Grupo 4
Gasto por Persona	4 vs 3	0.0486	Grupo 4 > Grupo 3
PIB per Cápita	3 vs 1	0.0231	Grupo 3 > Grupo 1
PIB per Cápita	3 vs 2	0.0115	Grupo 3 > Grupo 2
PIB per Cápita	4 vs 3	0.1401	Grupo 4 > Grupo 3

Group 1: This group has a higher density than groups 2 and 3, but lower than group 4. Their level of self-sufficiency is lower than that of group 3, but higher than that of group 4. In terms of the proportion of sales and purchases, it is below group 3 but above group 4.

Group 2: This group has a lower density than groups 1 and 4, but higher than group 3. Its level of self-sufficiency is lower than that of Group 1, but exceeds Groups 3 and 4. In terms of the proportion of sales and purchases, it is below Group 1 but exceeds Groups 3 and 4.

Group 3: This group has a lower density than group 1 but higher than group 2. Their level of self-sufficiency is higher than that of group 4 but lower than that of group 2. In terms of the ratio of sales to purchases, it outperforms all other groups.

Group 4: This group has a higher density than any other group. Although their level of self-sufficiency is lower than that of group 2, their sales and purchase shares are not adversely affected, indicating a solid economic balance.

SOM

Variable	Comparación Grupos	Nivel de Significación	Sentido de la Diferencia
Densidad	3 vs 1	0.0314	Grupo 3 < Grupo 1
Densidad	4 vs 1	0.0270	Grupo 4 < Grupo 1
Autoabas	2 vs 1	0.0749	Grupo 2 < Grupo 1
Autoabas	3 vs 1	0.1513	Grupo 3 < Grupo 1
Proporción Ventas	3 vs 1	0.0124	Grupo 3 < Grupo 1
Proporción Ventas	4 vs 1	0.0595	Grupo 4 < Grupo 1
Proporción Compras	2 vs 1	0.0307	Grupo 2 < Grupo 1
Proporción Compras	3 vs 1	0.0940	Grupo 3 < Grupo 1
Proporción Compras	4 vs 1	0.8626	Grupo 4 = Grupo 1
Ratio Máxima Venta Total	3 vs 2	8.2e-09	Grupo 3 < Grupo 2
Ratio Máxima Venta Total	4 vs 2	1.6e-08	Grupo 4 < Grupo 2
Ratio Máxima Venta Total	4 vs 3	0.7576	Grupo 4 = Grupo 3

Gasto por Persona	4 vs 1	0.0041	Grupo 4 < Grupo 1
PIB per Cápita	4 vs 1	0.0044	Grupo 4 < Grupo 1
PIB per Cápita	4 vs 3	0.0196	Grupo 4 < Grupo 3
Temp. Empleo Agrícola	4 vs 1	0.0439	Grupo 4 < Grupo 1
Temp. Empleo Agrícola	4 vs 3	0.0539	Grupo 4 < Grupo 3
T. Buscando Empleo	4 vs 3	0.0078	Grupo 4 > Grupo 3

Group 1: This group has a higher density and level of self-sufficiency compared to Groups 3 and 4. This could indicate that the Autonomous Communities in this group have a stronger economy and are more self-sufficient.

Group 2: The only comparison available for Group 2 in the table is with Group 1 in terms of self-sufficiency. Group 2 has a lower level of self-sufficiency than Group 1. This could suggest differences in the economic or employment structure between these two groups.

Group 3: Group 3 has a lower density than Group 1 but a higher level of self-sufficiency than Group 4. This could indicate that, although these Autonomous Communities may have a lower concentration of certain attributes or characteristics (represented by "Density"), they are able to be self-sufficient more efficiently than Group 4.

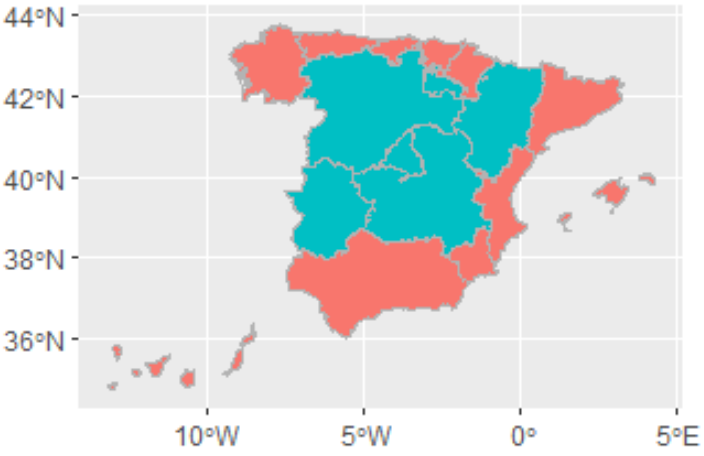
Group 4: Although Group 4 has a higher density than any other group, its level of self-sufficiency is lower than that of Group 2. This could suggest that, although these Autonomous Communities have a high concentration of density, they are more dependent on external sources for their supply.

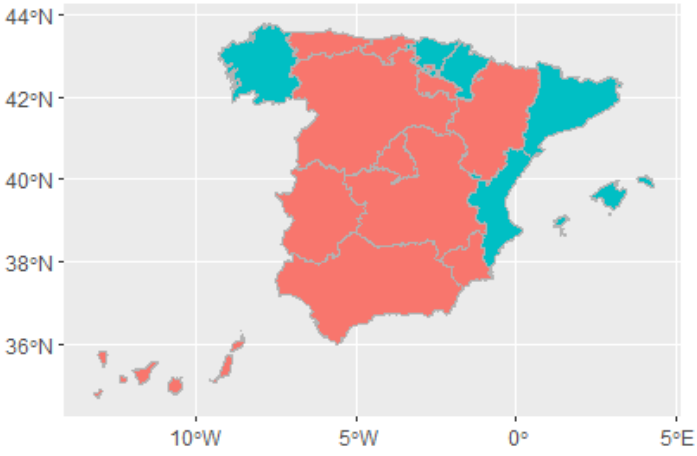
CONCLUSIONS

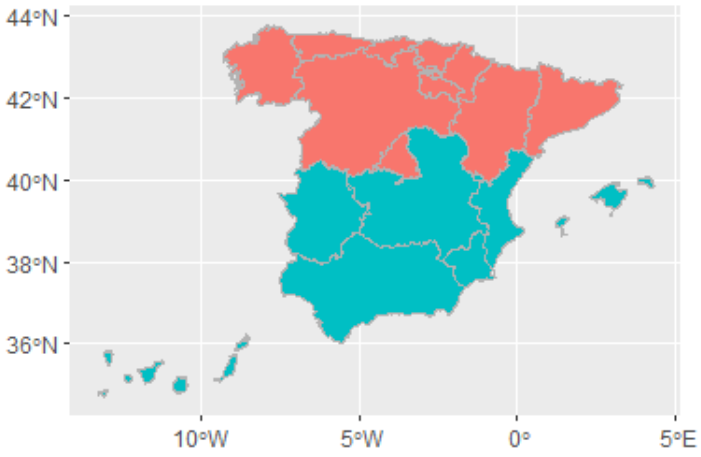
After analysing the different groupings generated, we have observed that they have many differences from each other. However, our choice would be the grouping obtained using the Hierarchical Complete algorithm. This selection is justified by the greater clarity and distinction of the groups, with the exception of Group 1, which is less differentiated. The rest of the groupings are more complicated to explain and are not as uniform. Now we can compare our clusters with the previous ones.

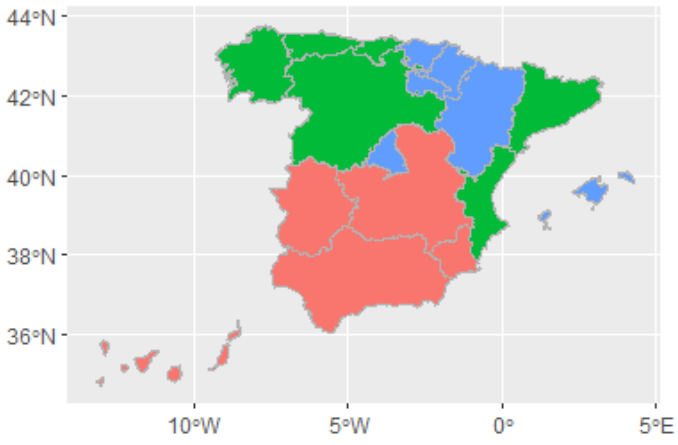
COMPARISONS WITH PRE-EXISTING CLASSIFICATIONS

We have graphed the previous groupings with the mapSpain library, for this we extracted a database already included in the package, *esp_codelist*. and artificially created lists with the number of the group as factors of each criterion used for the maps. The criteria are as follows, in order, by periphery or interior (RED=periphery, BLUE=interior), by own language or not (RED=no own language, BLUE=own





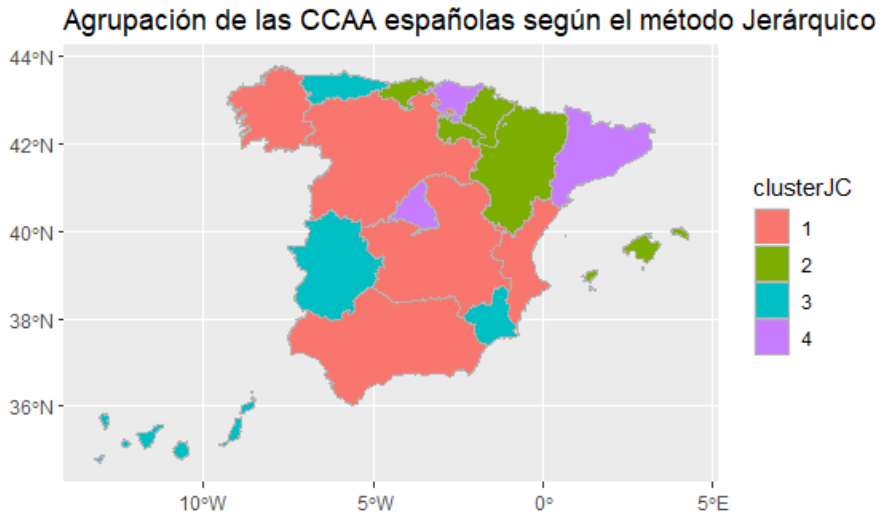




language), by north or south (RED=North, BLUE=South), and according to the level of unemployment (RED= more than 15%, BLUE= less than 10%, GREEN= between 10% and 15%).

To make the graphs we have used *ggplot(data)+geom_sf(aes=fill=cluster)...)*tags.

Our clusterer, on the other hand, looks like this:



The coincidences that we can see are few, perhaps in terms of their own language in cluster 4, which the Basque Country and Catalonia have. In addition, we could see group 1 as inland if we look at the North area. We could see that group 2 is made up of those with less than 10% less unemployment in Madrid and the Basque Country. But, in addition to this, we do not see any clear similarities, especially with group 3.

This may indicate that our country is very diverse and therefore could be classified in many ways according to the objective of the study. In our case, we have opted for a rather economical approach, but any of the above would be just as valid.

BIBLIOGRAPHY

Hernangómez, D. (n.d.). Get started.

https://cran.rproject.org/web/packages/mapSpain/vignettes/mapSpain.html

RPUBS - ML-Assignment 2 Clustering. (n.d.).
https://rpubs.com/sushantgote/ml2clustering

som. (n.d.). https://www.uv.es/mlejarza/datamine/som.html

RPUBS - Good Practice Toolbox for Analyzing Data using Unsupervised Learning Methods. (n.d.).

https://rstudiopubsstatic.s3.amazonaws.com/548392_de6cd39746a641718eb34123421cb11f.html
GeeksforGeeks. (2023, March 21). Making Maps with R.
https://www.geeksforgeeks.org/making-maps-with-r/

RPUBS-Exploratory Analysis by Principal Components and Cluster
https://edimer.github.io/Stat/6_MetodosExploratorios.html

GeeksforGeeks. (2023, April 18). Self organizing maps Kohonen maps.
https://www.geeksforgeeks.org/self-organising-maps-kohonen-maps/