STUDY OF THE SOCIO-ECONOMIC GROUPING OF THE MUNICIPALITIES OF THE METROPOLITAN AREA OF VALENCIA

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 municipalities in the Metropolitan Area of Valencia, based on socioeconomic data collected in the "amValencia1.csv" file. The aim of this study is to use clustering techniques to group municipalities into homogeneous but heterogender groups, which will allow us to better understand the similarities and differences between the various Valencian regions.

We have employed farthest first techniques and hierarchical clustering, including simple chaining and full chaining, as well as Ward's method, to obtain four clusters in both RStudio and Weka.

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 region and can be useful for decision-making and strategic planning in the area.

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, Hierarchical Ward, and Hierarchical Simple methods.

PREVIOUS TASKS

The starting point is the "amValencia1.csv" file, which contains socio-economic information on the 59 municipalities 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.

One of the important first initial steps is the removal of unnecessary attributes. In this case, we remove the "Proximity" and "Deployment" attributes using the "Remove" option. In addition, we identified an additional empty row in the CSV file, resulting in a total of 60 instances instead of the expected 59. This extra row was removed by using the "Edit" option, right-clicking on it and selecting "Delete Selected Instance", thus ensuring that we have 59 instances.

And finally, since we're looking to group municipalities together and not classify them, we set the class to "No Class" so that a class attribute isn't included.

SETTING UP THE GROUPINGS

As we have already mentioned before, the groupings we made in Weka were FarthestFirst, and the hierarchical Ward and simple, and the way of grouping each clustering is as follows:

-FarthestFirst: seeks to create groups taking as a starting point the point furthest from the centroid of the dataset. This approach ensures that groups are as far apart from each other as possible, i.e. that they are as heterogeneous as possible.

-Hierarchical Ward: 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.

-Simple Hierarchical: it is based on the idea of gradually merging the groups by choosing the two groups closest to each other based on the Euclidean distance between their closest elements.

CLUSTERING VISUALIZATION AND ATTRIBUTES

Once the clustering process is complete, it is crucial to understand the composition of the resulting clusters. To do this, we use the "Visualize Cluster Assignments" option, which allows us to see the instances that make up each cluster.

In addition, we create clustering attributes that reflect the belonging of each municipality to a specific group according to the method used. This makes it easier to analyze and interpret the results.

In summary, this clustering process and the previous tasks carried out at Weka provide us with a solid foundation to explore and understand the diversity of municipalities in the metropolitan area, which will allow us to make informed decisions and plan future strategies.

THE CLUSTERING PROCESS IN RSTUDIO

PREVIOUS TASKS

Then, we made the remaining grouping, that of the hierarchical Complete with RStudio, a programming language environment dedicated to statistical computing, among other tasks. Link to the Rstudio script.

The first thing was to import the file that we saved from Weka with the groupings to have them for later study. With this imported file we can create the matrix we call Z without including "muni" or the groupings in the variables, and we use the variable "muni" to name each of the observations and thus make it easier to understand.

SETTING UP THE GROUPING

Once the Z matrix is assembled, we can create the complete hierarchical cluster, which consists of the following:

-Hierarchical Complete: seeks to merge the groups by selecting the two that have the greatest distance between their furthest elements.

To do it in R we use the hclust(method="complete") function, and to visualize it we use the plot(cut(h=320)) and so we can see it with the 4 groups formed.

Just by looking at it you can see how groups 3 and 4 are individual groups, and if you look at the whole tree we can detect what these cities are much faster.

Arbol superior al corte h=250
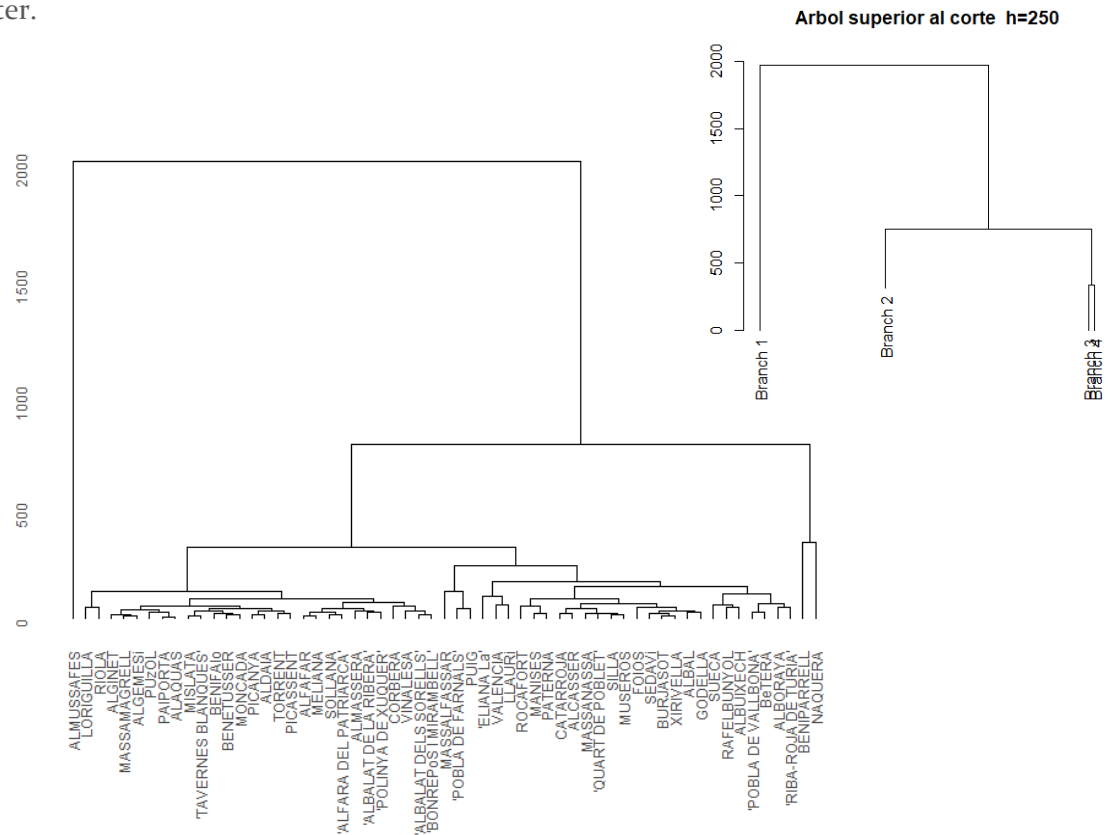
Here we already see that the "marginalized" villages are Beniparrell and Náquera, and then we will study why they are separated from the others.

Once we have visualized the cluster, we can add a fourth to our file that already contained three groupings. To do this, we create the cluster with a cutree(k=4) and add it.

ELEMENTS OF EACH GROUPING

FIRST THE FURTHEST AWAY:

GROUP 1: Mislata, Tavernes Blanques, Burjassot, Valencia, Alboraya, Pobla de Farnals, La Eliana, Massamagrell, Picanya, Almassera, Meliana, Torrent, Rocafort, Quart de Poblet, Manises, Paterna, Museros, Foios, Godella, Puig, Albalat dels Sorells, Rafelbunyol, Moncada, Pobla de Vallbona, Puzol, Bétera, 'Riba-Roja de Turia, Bonrepos i Mirambell, Náquera, Vinalesa.

GROUP 2: Benifaió, Algemesí, Sueca, Alginet, Sollana, Albalat de la Ribera, Albuixech, Alfara del Patriarca, Corbera, Llaurí, Loriguilla, Massalfassar, Polinya de Xuquer, Riola.

GROUP 3: Silla, Aldaia, Almussafes, Picassent, Beniparrell.

GROUP 4: Benetússer, Paiporta, Xirivella, Alfafar, Sedaví, Albal, Alaquàs, Massanassa, Catarroja, Alcàsser.

FULL HIERARCHICAL:

GROUP 1: Benetússer, Mislata, Tavernes Blanques, Paiporta, Burjassot, Xirivella, Valencia, Alfafar, Sedaví, Albal, Alboraya, Alaquàs, Massanassa, Pobla de Farnals, Eliana La, Massamagrell, Picanya, Catarroja, Silla, Almassera, Aldaia, Meliana, Torrent, Rocafort, Alcàsser, Quart de Poblet, Manises, Paterna, Museros, Foios, Godella, Puig, Albalat dels Sorells, Rafelbunyol, Benifaió, Moncada, Algemesí, Pobla de Vallbona, Puzol, Picassent, Sueca, Bétera, Riba-roja de Túria,   Alginet, Sollana, Albalat de la Ribera, Albuixech, Alfara del Patriarca, Bonrepòs i Mirambell, Corbera, Llaurí, Loriguilla, Massalfassar, Polinyà de Xuquer, Riola, Vinalesa.

GROUP 2: Almussafes

GROUP 3: Beniparrell

GROUP 4: Náquera.

SIMPLE HIERARCHICAL:

GROUP 1: Benetússer, Mislata, Tavernes Blanques, Paiporta, Burjassot, Xirivella, Valencia, Alfafar, Sedaví, Albal, Alboraya, Alaquàs, Massanassa, Pobla de Farnals, Eliana La, Massamagrell, Picanya, Catarroja, Silla, Almassera, Aldaia, Meliana, Torrent, Rocafort, Alcàsser, Quart de Poblet, Manises, Paterna, Museros, Foios, Godella, Puig, Albalat dels Sorells, Rafelbunyol, Benifaió, Moncada, Algemesí, Pobla de Vallbona, Puzol, Picassent, Sueca, Bétera, Riba-roja de Túria,   Alginet, Sollana, Albalat de la Ribera, Albuixech, Alfara del Patriarca, Bonrepòs i Mirambell, Corbera, Llaurí, Loriguilla, Massalfassar, Polinyà de Xuquer, Riola, Vinalesa.

GROUP 2: Almussafes

GROUP 3: Beniparrell

GROUP 4: Náquera.

HIERARCHICAL NEIGHBORHOOD

GROUP 1: Benetússer, Paiporta, Xirivella, Alfafar, Sedaví, Albal, Alaquàs, Massanassa, Picanya, Catarroja, Silla, Aldaia, Torrent, Alcàsser, Benifaió, Algemesí, Almussafes, Picassent, Alginet, Beniparrell.

GROUP 2: Mislata, Tavernes Blanques, Burjassot, Valencia, Massamagrell, Almassera, Meliana, Quart de Poblet, Manises, Paterna, Museros, Foios, Albalat dels Sorells, Rafelbunyol, Moncada, Puçol, Albuixech, Alfara del Patriarca, Bonrepòs i Mirambell, Massalfassar, Vinalea.

GROUP 3: Alboraya, Pobla de Farnals, Eliana La, Rocafort, Godella, Puig, Pobla de Vallbona, Bétera, Riba-roja de Túria, Náquera.

GROUP 4: Sueca, Sollana, Albalat de la Ribera, Corbera, Llaurí, Loriguilla, Polinyà de Xuquer, Riola.

As we can see, the most characteristic thing is that the complete hierarchical and the simple hierarchical have ended up being exactly the same in the sense of elements that make up their groups, despite having followed a different process. Another thing to take into account in these groupings is the marginalization of three municipalities into three groups, so we can take into account that when analyzing them we will be talking about specific groups that are different or strange enough 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 and the corresponding model and with FUN=mean. When we analyze them, we see that the Complete Hierarchical and the Single are exactly the same, so we will treat them as such. And we detected differences in the means of the following variables:

-FarthestFirst: Mortality, Growth, Migration, Aging, Ageing, Housing, Taf, Commercial. Iaindus and cenergindus.

-Simple and complete hierarchical: Tnatality, Tgrowth, Migration, aging, nlineastelef, nvehi, segvivienda, vivdesocu, taf, iafinan, iaindus, cenerdomest, cenergindus, iparlabindustria and iparlabconstr.

. Hierarchical District: Mortality, Growth, Migration, Aging, Segvivienda, iacomercial, iaindus, cenerdomest, cenergindus, formacad, iparlabindustria and iparlabconstr.

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 municipalities in Valencia in the variables in which we have found apparent differences between the means of the groups with the aov(variable~cluster) function. To do this, we have converted the variables of the clusters 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 that have a numerical interpretation. The results have been, in summary, as follows:

FIRST THE FURTHEST
- Mortality rate: significant differences were found at all levels of significance.
- Growth rate: significant differences were found at any level of significance.
- Migration rate: no significant differences were found.
- Ageing: significant differences were found between the groups at any level of significance.
- Number of vehicles: significant differences were found between the groups from a significance level of 0.01.
- Second home: we did not find enough evidence to show that there were significant mean differences.
- Female activity rate: significant differences were found between the groups at any level of significance.
- Commercial Activity Index: significant differences were found between the groups at any level of significance.
- Industrial Activity Index: significant differences were found between the groups at any level of significance.
- Industrial Electrical Energy Consumption: significant differences were found at any level of significance.

SIMPLE HIERARCHICAL/FULL HIERARCHICAL
- Migration rate: significant differences were found between the groups at a significance level of 0.05.
- Number of telephone lines: significant differences were found between the groups at any level of significance.
- Second home: significant differences were found between the groups at any level of significance.
- Unoccupied dwellings: significant differences were found between the groups at a significance level of 0.01.
- Domestic electricity consumption: significant differences were found between the groups at a significance level of 0.01.
- Industrial electrical energy consumption: significant differences were found between the groups at any level of significance.
- Industrial labor participation index: significant differences were found between the groups at a significance level of 0.05.
- Labor participation index in construction: significant differences were found between the groups at a significance level of 0.05.

HIERARCHICAL NEIGHBORHOOD
- Mortality rate: significant differences were found between the groups at any level of significance.
- Growth rate: significant differences were found between the groups at any level of significance.
- Migration rate: significant differences were found between the groups at any level of significance.

- Ageing: significant differences were found between the groups at any level of significance.
- Second home: significant differences were found between the groups at any level of significance.
- Female activity rate: significant differences were found between the groups at any level of significance.
- Commercial activity index: significant differences were found between the groups at a significance level of 0.01.
- Industrial activity index: significant differences were found between the groups at any level of significance.
- Domestic electricity consumption: significant differences were found between the groups at any level of significance.
- Academic background: significant differences were found between the groups at a significance level of 0.01.
- Industrial Labor Force Participation Index: Significant differences were found between the groups at any level of significance.
- Construction Labor Participation Index: Significant differences were found between the groups at any level of significance.

MULTIPLE ANALYSIS

Finally, we performed the Tukey analysis with the TukeyHSD(anova) function on the variables in which we 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.

FIRST THE FURTHEST
- Mortality rate: Multiple analysis tests revealed that groups 2 and 4, 3 and 2 and 2 and 1 showed significant differences from each other, and specifically, analyzing the diff (if negative, the first number is lower, and if positive, it is higher), group 2 is significantly higher than 1, 3 and 4.
- Growth rate: The analysis revealed that group 2 had a significantly different growth rate than the other groups. 2 being significantly lower than 1, 3 and 4.
- Aging: Tests revealed that group 2 had significantly different aging than the other groups. Group 2 is significantly larger than group 1, 3 and 4.
- Number of vehicles: The evidence indicated that group 3 had a significantly different number of registered vehicles compared to the other groups. 3 being significantly higher than 1, 2 and 4.
- Female activity rate: Post hoc tests revealed that group 2 had a significantly different female activity rate than the other groups. 2 is significantly lower than 1, 3 and 4.
- Business Activity Index: Post hoc tests indicated that group 4 had a significantly different number of commercial activity premises compared to the other groups. Group 4 was significantly higher than group 1, 2 and 3.

- Industrial Activity Index: Post hoc tests revealed that group 4 had a significantly different number of industrial activity premises compared to the other groups. Group 4 was significantly higher than group 1, 2 and 3.
- Industrial Electrical Energy Consumption: Post hoc tests indicated that group 3 had significantly different consumption compared to the other groups. Group 3 is significantly higher than group 1, 2 and 4.

SIMPLE HIERARCHICAL/FULL HIERARCHICAL
- Migration rate: Tests revealed that group 4 had a significantly different migration rate than the other groups. Group 4 is significantly larger than group 1, 2 and 3.
- Number of telephone lines: The tests revealed that group 4 was significantly higher than 1, 2 and 3 (only if you choose the significance level of 0.05) and group 3 was significantly higher than 1 and 2.
- Second home: Testing revealed that group 4 had a significantly different number of second homes than the other groups. Group 4 is significantly larger than group 1, 2 and 3.
- Unoccupied dwellings: Tests revealed that group 4 had a significantly different number of vacant dwellings than the other groups. Group 4 was significantly smaller than group 1, 2 and 3 (at a significance level of 0.01).
- Household electricity consumption: The tests revealed that group 4 had significantly different household energy consumption than the other groups. Group 4 is significantly larger than group 1, 2 and 3.
- Industrial electrical energy consumption: Tests revealed that group 1 was significantly lower than 2 and 3, and group 4 significantly lower than 2 and 3.
- Industrial Labor Participation Index: The tests revealed that group 4 was significantly lower than groups 2 and 3.
- Construction Labor Participation Rate: Testing revealed that group 4 had a significantly different construction labor participation rate than the other groups. Group 4 is significantly larger than group 1, 2 and 3.

HIERARCHICAL NEIGHBORHOOD
- Mortality rate: Tests revealed that group 4 had a significantly different mortality rate than the other groups. Group 4 is significantly larger than group 1, 2 and 3.
- Growth rate tests revealed that group 4 was significantly lower than 1 and 3, and group 2 significantly lower than 1 and 3.
- Migration rate: The tests revealed that group 3 had a significantly different migration rate than the other groups. Group 4 is significantly larger than groups 1, 2 and 4.
- Aging: The tests revealed that group 4 was significantly older than 1, 2, and 3, group 3 significantly smaller than 1 and 2, and group 2 significantly older than 1.
- Second home: Tests revealed that group 3 was significantly larger than 1, 2 and 4, and group 4 significantly larger than 1 and 2.

- Female activity rate: Tests revealed that group 4 had a significantly different female activity rate than the other groups. Group 4 is significantly smaller than group 1, 2 and 3.
- Business Activity Index: Tests revealed that group 1 was significantly higher than groups 3 and 4.
- Industrial Activity Index: Tests revealed that group 1 was significantly higher than group 3 and 4.
- Household electricity consumption: Tests revealed that group 3 was significantly higher than group 1, 2 and 4.
- Educational background: Tests revealed that group 3 was significantly larger than group 1 and 2.
- Industrial Labor Participation Index: The tests revealed that group 1 was significantly higher than groups 2, 3 and 4, and group 2 significantly higher than groups 3 and 4.
- Construction Labor Participation Rate: Tests revealed that group 4 was significantly higher than groups 1 and 2.

INTERPRETATION OF THE GROUPS

In the last section of our report, we focus our attention on the interpretation of the clusters resulting from our clustering analyses. Up to this point, we have applied various statistical techniques to group the municipalities of the metropolitan region of Valencia into homogeneous groups. However, grouping the data is only the first step. Now, we look at the meaning and distinguishing characteristics of each of these groups. To do this, we use the multiple comparisons, specifically the Tukey test, which we have just provided. Through this analysis, we will unveil the particularities that characterize each group and, in doing so, contribute to a deeper understanding of the socioeconomic and demographic structure of the region.

FIRST THE FURTHEST AWAY:

GROUP 1: Group 1 is emerging as a group of municipalities with a more active demographic dynamic compared to Group 2, lower rates of industrial and commercial activity in contrast to Group 4, and a higher female labor force participation than Group 2. These characteristics suggest a specific demographic and economic profile that distinguishes it from the other groups, but with less differentiating nuances than in the others.

GROUP 2: This group is characterized by having the highest mortality and aging rate, and the lowest growth rate, which leads us to think that they are aging and emptier cities, this would explain why the female activity rate is lower than all the others since in these cities labor activity is usually more masculinized while women stay at home due to more gender roles Traditional.

GROUP 3: This group is characterized by having more industrial energy consumption, although not the highest index of industrial activity, and more registered vehicles than

the others, this may be because its industrial activity is very powerful in terms of the use of electrical energy, and it needs the use of a lot of transport to carry out its activity.

GROUP 4: This group is characterized by having a higher index of commercial activity and industrial activity, this may be due to the fact that they are cities with a large concentration of companies and businesses in the commercial and industrial sectors. This could be a result of its strategic location, favorable infrastructure, access to specific resources or markets, economic development policies, or deep-rooted trade and industrial traditions in the region.

SIMPLE/COMPLETE HIERARCHICAL:

GROUP 1: Group 1 shows a profile that is characterized by lower telephone connectivity, which may occur because it is not a very commercial or industrial area, lower industrial energy consumption and less migration, second homes, domestic energy consumption and construction compared to Group 4 despite having more unoccupied homes than the latter. These differences may indicate that these areas, grouped in this way, are of little interest in the sense of the study since they do not yield a very coherent interpretation.

GROUP 2: Group 2 is distinguished by its higher consumption of industrial electrical energy and a higher rate of labor participation in industry, indicating a possible greater industrial or residential activity. Despite this, it has fewer phone lines than other groups, suggesting a less developed communication infrastructure. In addition, Group 2 has less migration, fewer second homes, lower household energy consumption, and lower labor participation in construction compared to Group 4. These differences could reflect a more stable population and less construction activity in these areas, as well as a higher number of unoccupied homes compared to Group 4.

GROUP 3: Group 3 is distinguished by having more telephone lines than Groups 1 and 2, although still less than Group 4, shows a higher industrial energy consumption compared to Groups 1 and 4, and a higher labor participation in industry compared to Group 4, which could indicate greater industrial activity in these areas that requires telephone lines. On the other hand, this group has less migration, fewer second homes, lower domestic energy consumption and lower labor participation in construction compared to Group 4, but has more unoccupied homes than Group 4 (although with less significance than in groups 1 and 2).

GROUP 4: Group 4 is the most interesting of all in terms of clear and interpretable differences, it is characterized by a high level of migration, which suggests a greater mobility of the population in these areas. In addition, this group has more telephone lines and second homes compared to Groups 2 and 3, which could indicate a greater communications infrastructure and a greater presence of vacation properties. On the other hand, Group 4 has fewer unoccupied dwellings, higher household energy consumption but lower industrial energy consumption compared to Groups 2 and 3.

This could reflect higher energy use in households, possibly due to a higher number of occupied homes, but lower industrial activity in these areas. In addition, Group 4 has a lower industrial labor participation rate and a higher construction participation rate compared to the other groups. This could suggest less dependence on industry and greater participation in the construction sector in these areas by the construction of holiday homes.

HIERARCHICAL NEIGHBORHOOD:

GROUP 1: Group 1 is characterized by having a higher population growth compared to other groups, also showing moderate aging. It has a lower presence of second homes in the region and stands out for a higher rate of commercial and industrial activity. However, educational background is lower in this group compared to Group 3. In addition, Group 1 has a higher labor participation rate in the industrial sector, but a lower participation in the construction sector compared to Group 4. These characteristics suggest greater commercial and industrial economic development, but not in the building district, and a growing population, but with slightly lower levels of education compared to other groups.

GROUP 2: Less growth than 3 and 1, more ageing than 1, but less than 4 and 3, fewer second homes than 4 and 3, less education than 3, more industrial labour participation rate than 3 and 4, but less than 1, and less construction labour participation rate than 4.

GROUP 3: Group 2 is characterized by having a lower population growth rate compared to Groups 1 and 3, and although it presents population aging, it is moderate and is in an intermediate range between groups 1 and 4. In addition, it shows a lower presence of second homes in the region and lower levels of educational education compared to Group 3. This group stands out for its labor participation rate in the industrial sector, although it is surpassed by Group 1 in this aspect. On the other hand, it has a lower labor participation in the construction sector than Group 4. These characteristics suggest a group with moderate population growth and ageing and a greater presence in industrial activities compared to other economic sectors.

GROUP 3: Group 3 is characterized by a higher population growth rate compared to Groups 2 and 4, indicating more dynamic population growth. In addition, this group has a higher migration rate, suggesting a higher population flow. In addition, population ageing in this group is lower than in Groups 1, 2 and 4, reflecting a younger population. This group also shows a greater presence of second homes in the region, which could indicate a greater influx of tourism or residential areas, which adds to the fact that it has higher domestic energy consumption and a higher educational background compared to Groups 1 and 2. On the other hand, its labor force participation rate in the industrial sector is lower than that of Groups 1 and 2, suggesting a less industry-centric economy. In summary, Group 3 appears to be experiencing sustained population growth, with a younger population, a diversified economy, and more residential areas.

GROUP 4: Group 4 is characterized by a higher mortality rate, a lower population growth rate compared to Groups 1 and 3, and greater aging, suggesting a less dynamic demographic profile and an aging population. The presence of second homes in this group is higher than in Groups 1 and 2, which could indicate an influx of tourism or a greater number of residential areas. The female participation rate in this group is lower compared to Groups 1 and 2, which could reflect a lower participation of women in the labor force due to more traditional roles. Economically, Group 4 has a lower index of commercial and industrial activity than Group 1, suggesting an economy that is less oriented towards these sectors. However, its labour force participation rate in the construction sector is higher than that of Groups 1 and 2. In summary, Group 4 appears to be experiencing an aging population with slower growth, a less industrialized economy, and a greater presence of second homes.

CONCLUSIONS

After analysing the different groupings generated, we have observed that they have many differences from each other. However, our choice would be the clustering obtained using the Farthest First algorithm. This selection is justified by the greater clarity and distinction of the groups, with the exception of Group 1, which is less differentiated. On the other hand, the Complete and Single hierarchical groupings focus too much on Group 4 as the most distinctive and additionally have the three "strange" municipalities separated into three unique distinct groups, while the Ward grouping generates more complex sets to explain due to the difference relationships between groups. In summary, we consider that the Farthest First grouping presents the most interesting and clear interpretation for our study.