STATISTICAL MACHINE LEARNING

Second Assignment

Professor: D.Karlis

06/06/2024
MSc in Statistics - AUEB
Grammenos Konstantinos
AM : f3612302
Project 2
Title: Clustering in Greek Data Population Data, Census 2001

Contents

List of Figures

Abstract

In this assignment, we explore the age composition of the Greek population across various municipal districts in Greece, based on the 2001 census data. The objective of this assignment is to identify distinct groups of municipalities with similar age compositions through clustering analysis, using hierarchical and model-based clustering methods. The effectiveness and reliability of these clustering approaches are evaluated and compared. The findings from the clustering provide us some useful insights for the demographic segmentation across the country.

Introduction

This report aims to analyze the age composition of Greek municipalities using data from the 2001 census, applying advanced clustering techniques to identify and interpret distinct demographic patterns. To achieve this, we employed both hierarchical and model-based clustering approaches. Hierarchical clustering was utilized to explore the natural grouping of municipalities based on their age distributions. This method provided us with 4 distinct clusters of different municipalities, while the model based one gave us 18 different clusters as result. The combination of these two methods ensured a comprehensive and robust analysis of our dataset.

In the hierarchical clustering phase, we used the Mahalanobis distance matrix to account for the correlations between different age groups and used Ward's method to construct the dendrogram. Also, to validate the optimal number of clusters and evaluate our cluster, we performed silhouette analysis and evaluated the Wilks' Λ statistic. With those metric we understand the quality of our clustering.

For the model-based clustering, we fitted Gaussian Mixture Models with various covariance structures to the data. The Bayesian Information Criterion (BIC) was used to select the optimal model. The analysis revealed that the VEI provided the best fit for our data. The model identified 18 clusters as the optimal number, capturing the diverse age compositions across the municipalities.

To interpret the clustering results, we visualized the age distribution within each cluster and examined the demographic patterns. Principal Component Analysis (PCA) was also employed to reduce the dimensionality of the data and provide a clear visual representation of the clusters. The PCA plot illustrated the separation and characteristics of the clusters, enhancing our understanding of the demographic segmentation.

Data Processing

For this assignment, we had to face several challenges in order to derive the different clusters among the municipalities and understand the demographic composition of each one of them. Firstly, the challenge of the dataset. The initial dataset contained the Greek names of municipalities, and their suburbs and was consisted of 7.231 observations and 27 variables. The variables of the dataset were about the number of people of both genders but also only males and only females were living in these areas during 2001 and all these counts were distributed by the age meaning that we had people at age 0-14, 15-24, 25-39, 40-54, 55-64, 65-79 and over 80 years old. With the use of some necessary libraries we manage to load the Greek data into the R programming language and also to clean the dataset in way to have 7237 observations and 26 variables. However, because our variables of the dataset were written in Greek, we decided for our own convenience to rename these variables-columns of our dataset to new ones more understandable. Also, because our data were in an xls format and some cells in excel were merged into one and R does not recognized this, we had to made a processing in order to derive the new columns.

So after this detailed cleaning and processing of our data, we ended up with these columns Geographical Code" "Municipality", "Both Genders Sum", "Both Genders 0-14" , "Both Genders 15-24", "Both Genders 25-39", "Both Genders 40-54", "Both Genders 55-64", "Both Genders 65-79", "Both Genders 80+", "Males Sum", "Males 0-14", "Males 15-24", "Males 25-39"   "Males 40-54", "Males 55-64", "Males 65-79", "Males 80+", "Females Sum", "Females 0-14", "Females 15-24", "Females 25-39", "Females 40-54", "Females 55-64", "Females 65-79", "Females 80+" . Also, we observed that while all the columns are counts of how many people belong to each municipality – suburbs of the municipality, there were in character format, so we transformed them into the numeric format.

Clustering Methods

First thing we did in order to be able to apply some clustering methods into our dataset was to keep only the rows that were about the municipality like for example the municipality of Drama. Next we discarded the gender information by keeping only the columns that refer to both genders combined population and their sum in each municipality. By applying these steps we ended up with 900 observation, which are the different municipalities recorded in Greece in 2001, and 8 variables – columns, which refer to the combined population of people of both genders distributed by age, who were living in each municipality during that period of time. Also, instead of using absolute counts, we converted the number of people in each age group to relative frequencies. This was achieved by dividing the population count of each age group by the total population of the respective municipality. This approach helps to normalize the data, allowing for a fair comparison between municipalities of varying population sizes. By converting to relative frequencies, we ensure that the age composition of each municipality is represented as a proportion of its total population and this highlights the demographic structure more clearly, without being skewed by the overall population size. In addition, giving the fact that the relative frequencies in each municipality sum up to 1, to simplify our analysis, we excluded one age group (e.g., "Both Genders 80+") from our dataset. Finally, we discarded the variable-column of Both Genders Sum, because we thought that after those steps is not necessary. As a result, our final dataset consists of the relative frequencies of all the age groups except those of 80+, for each municipality. In this way, we effectively capture the age composition of each municipality. All these processing steps are crucial for generating accurate and interpretable clustering results, allowing us to identify meaningful demographic patterns across Greece.

Hierarchical Clustering

The first method of clustering that we applied in order to distinguish how the country's municipalities are grouped based on age composition and in how many such groups, was Hierarchical Clustering. In Hierarchical Clustering, the algorithm take as an input a distance matrix and at each step we join/split the closer observations/clusters. So at the first step, given a set of N observations, in our dataset the observations are the 900 municipalities, the algorithm assign each observation to its own cluster so we have 900 clusters. At the second step the algorithm finds the closest cluster and merge them into a single one and at the third it compute the distances between the new cluster and each of the old ones. Lastly, the algorithm repeats the 2nd and 3rd step until I have 1 cluster only. For our example because our data are quantitative and also correlated we decide to use the Mahalanobis distance, because this distance takes into account the correlations of our data, but we also tried some other distances like the Euclidean and the Manhattan distance in order to check different solutions and results. Also, we have to say that tried different ways (linkages) of calculating the distances between 2 clusters.

So first, we calculated the Mahalanobis distance matrix between each pair of municipalities. With our distances calculated, we then applied hierarchical clustering to cluster the municipalities using a specific method called Ward's method, in which at each step the pair of clusters with the minimum between cluster distance are merged. Ward method also minimizes the total within cluster variance. However, in addition to Ward's method, we explored also other linkage methods for hierarchical clustering to ensure robustness and to compare different approaches and results. We tried the so called Complete Linkage in which the distance between 2 clusters is equal to the longest distance of any member of one cluster to any member of the other cluster, the Average Linkage and the Simple Linkage which respectively refers to the minimum distance between any single point in the first cluster and any single point in the second. The choice of which linkage we will use plays an extremely significant role in the clustering process.

So in our clustering analysis to identify the most appropriate approach for clustering the municipalities based on their age composition we used all these different methods like the Complete Linkage, Single Linkage, Average Linkage, and Ward's Method. For each method we made a dendrogram plot, in order to compare their effectiveness visually. A dendrogram plot show us the join of the clusters. More specifically in the y axis there is the dissimilarity between clusters and in the x axis the individual data points. So, after reviewing the dendrograms of each method we ended up with the Wards Method and below we can see its dendrogram.

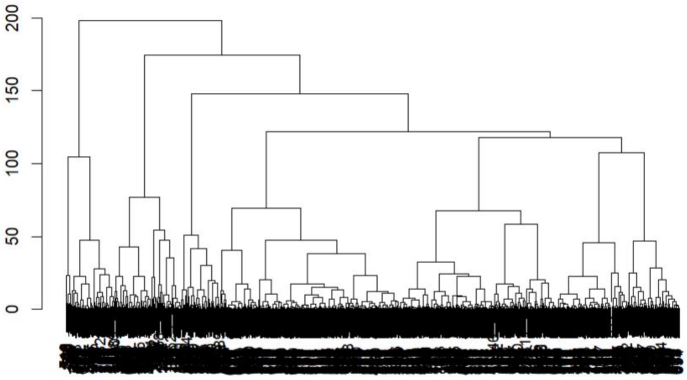

Figure 1:Dendrogram of Hierarchical Clustering using Ward method

In a dendrogram we examine the heights at which large jumps occur and in this way we can have a picture of the optimal number of clusters. Basically, what we do is to cut the dendrogram to form clusters that are distinct. So by examining the plot above we would say that that there are approximately 4 clusters.

Evaluation of Hierarchical Clustering

In addition to visual inspection of the dendrogram, we checked some different metrices to see how good is our clustering and to determine the optimal number of clusters for our dataset.

So, first we performed a Silhouette Values analysis and we calculated 1 Silhouette Value for each observation, this means 1 value for each municipality. These Silhouette Values will provide us with a view of how good is each cluster, so as we understand average Silhouette Values over all data of the entire dataset is a measure of how appropriately the data has been clustered. Silhouette Values can be between -1 and 1. A value close to 1 indicates that the municipality is well matched to its own cluster. Below, the plot represents the average silhouette width for different numbers of clusters, ranging from 2 to 15. Based on this, we determine that a possible number of clusters of Greek municipalities based on their age composition can be 4.

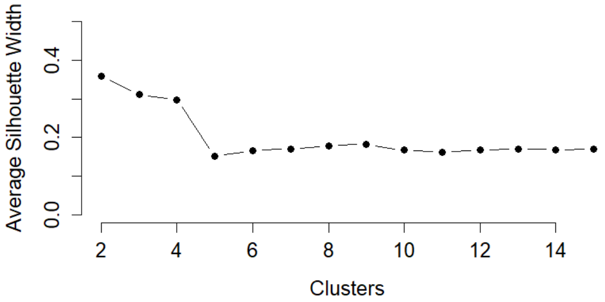

Figure 2: Average Silhouette Values for different number of Clusters using Ward method

Furthermore, we did a Silhouette Values plot for 4 clusters. The plot below is a visual representation of how well each of the 900 municipalities fits into one of the 4 clusters. Each bar in the plot represents a single municipality's silhouette width, which measures how similar that municipality is to its own cluster compared to other clusters. For example we observe that in the first cluster there are 674 municipalities, in the 2nd 96, in the 3rd 71 and in the 4rth cluster 57, while the average Silhouette Value is 0.33.

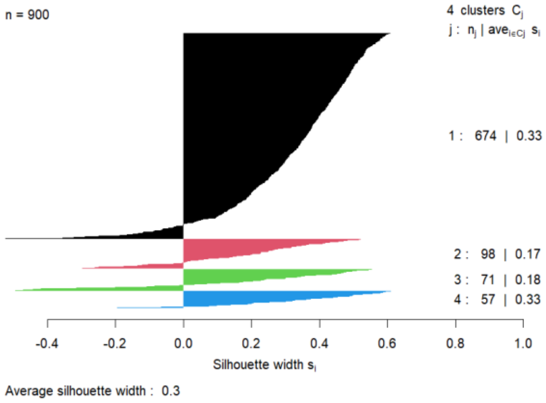

Figure 3: Silhouette Plot of 4 Clusters using Ward method

Another way that we used in order to evaluate our hierarchical clustering was to compute the Wilks λ statistic which provides an indication of the separation between clusters by comparing the within-cluster variability to the total variability. It works only for continuous data and we expect small values of Wilks λ statistic if we have a good clustering. The plot below shows the Wilks' Λ statistic for different numbers of clusters. The sharp decrease in Wilks' Λ for the first few clusters indicates significant improvement in clustering quality. Also, after around 4 to 5 clusters, the Wilks' λ values level off, suggesting that we do not gain anything with adding more clusters.

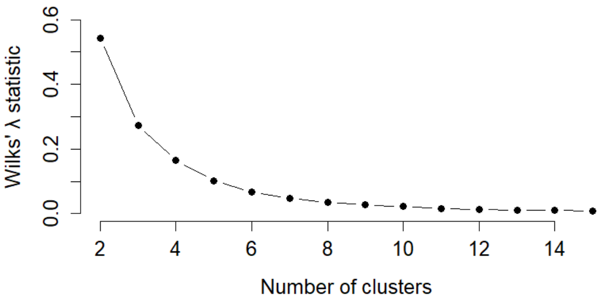

Figure 4:Wilks' λ statistic for different numbers of clusters

Visualization and interpretation of the results of Hierarchical Clustering

After all this analysis we proceeded to visualize the age distribution across the municipalities for each cluster. This step was crucial to interpreting the clustering results and understanding the demographic patterns within each cluster. In the plot below we see the different age distribution for each cluster of the municipalities. Cluster 1 is relatively balanced across most age groups, but we see a spike in the ages of 25 – 39 and 0-14, meaning that maybe in these municipalities there are mostly young people and children under 14. This may be an indication of young families and working-age population. In cluster 2, the majority of people living in these municipalities belong to the age group of 65-79 which suggest a concentration of older, pre-retirement, and retired individuals. The 0-14 age group has the lowest frequency which may indicates that those municipalities may refer to remote areas where young people have left and only the elderly remain and live there. This may be a sign that those remote places need to be under surveillance, in order not to be abandoned in the near future. Maybe the Greek state should take some action and provide some incentives to young people to go and live there. Those incentives may be economical ones. So, those municipalities are rural or less urbanized areas with aging populations. In Cluster 3, the age group of 15-24 stands out and also there is a significant amount of people in the age group of 25-39. This cluster likely includes urban or suburban areas with a significant number of children and teenagers or young adults. Cluster 4 shows a high proportion of older adults at the age group of 64-79 but with a more balanced representation than Cluster 2. In this cluster there may be municipalities with a mix of middle-aged and older populations, possibly transitioning areas with both working-age and plenty of retired individuals.

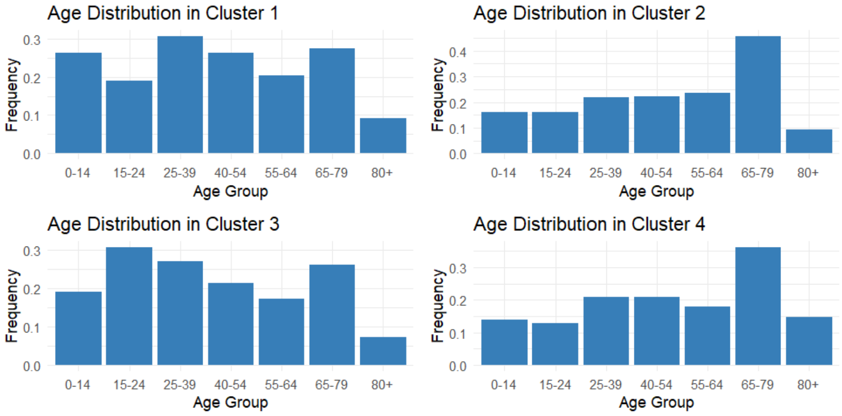

Figure 5: Age Distribution Per Cluster - Hierarchical Clustering

Principal Component Analysis

Last thing that we did regarding the hierarchical clustering that we applied was Principal Component Analysis (PCA). PCA basically is just a mathematical transformation of our correlated data to uncorrelated data. PCA is used for dimension reduction because basically what you do is to transform your variables and create new ones that summarize the information. In our assignment, PCA was applied to visualize and interpret the results of our clustering analysis. We chose to show only the first two principal components (Dim1 and Dim2) as they explain a substantial portion of the variability in the data (64.5% and 13.6%, respectively). Cluster 1is the largest and most densely populated in the plot, indicating that in this cluster the municipalities have mostly similar age distributions, while in cluster 2 age compositions differ significantly from those in Cluster 1. Lastly, cluster 3 is somehow above and far enough especially from the Cluster 1 and along with the second principal component (Dim2), which may shows us an age composition different enough from others. Finally, Cluster 4 overlaps with Cluster 3 and we can also confirm that by seeing that in both clusters' municipalities, there are significant older people of the age of 65-79.

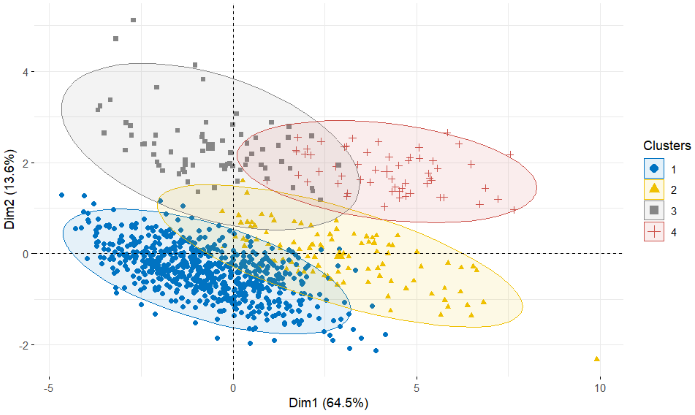

Figure 6:PCA using Mahalanobis Distance and Ward - Hierarchical Clustering

Model Based Clustering

Before saying anything about model based clustering, we have to declare the 2 most significant characteristics that define a cluster. These are the shape and the center of the cluster. The shape is a covariance matrix while a way to calculate the center of a cluster is to find the observation which has the minimum sum of distances from all the other observations in the cluster. Now, since we define these 2 important features, we will describe the model based clustering approach. So, model-based clustering is a sophisticated statistical method that assumes the data is generated from a mixture of underlying probability distributions, in a way that each observation may belong to 1 or more clusters with some probability. Each cluster is represented by a different distribution function, and the objective is to identify the parameters of these distributions as well as the number of clusters. Usually a way to find these parameters is through fitting a likelihood and finding the MLEs. In our dataset we fitted a Gaussian Mixture Models (GMMs) where each component of the mixture is a Gaussian distribution and the models parameters were estimated through the Expectation-Maximization (EM) algorithm. We fitted those mixture models using the transformed data of the relative frequencies of the number of people in each age group of each municipality. We fitted the models and we specified a range for the number of clusters (2 to 20) and allow the algorithm to explore different covariance structures. We fitted six different covariance structures, each representing a different way of modeling the variability and shape of the clusters. First, the EEI (Spherical, Equal Volume and Shape) which assumes clusters are spherical (equal shape in all directions) and have the same volume, the VII in which clusters are spherical but allows each cluster to have a different volume that means cluster of different sizes. Also, we fitted the EEI (Diagonal, Equal Volume and Shape), which allows for different variances (ellipsoids), EVI (Diagonal, Equal Volume, Variable Shape), ellipsoidal shapes with equal volume but different shapes, VEI (Diagonal, Variable Volume, Equal Shape) which lusters have axes-aligned ellipsoidal shapes with equal shape but different volumes basically clusters of different sizes with the same shape. Finally the VVI (Diagonal, Variable Volume and Shape) with ellipsoidal shapes with both variable volume and shape. We can see a visual representation of all of these in the figure below.

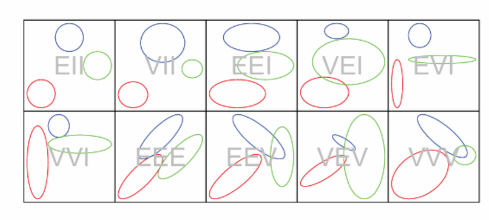

Figure 7: Different Models in Model Based Clustering

After fitting all these models, the model identified the optimal number of clusters based on the lowest BIC value is 18 clusters, while the top 3 models again based on the BIC were the VEI 18 with 17 clusters and third with 20 clusters.

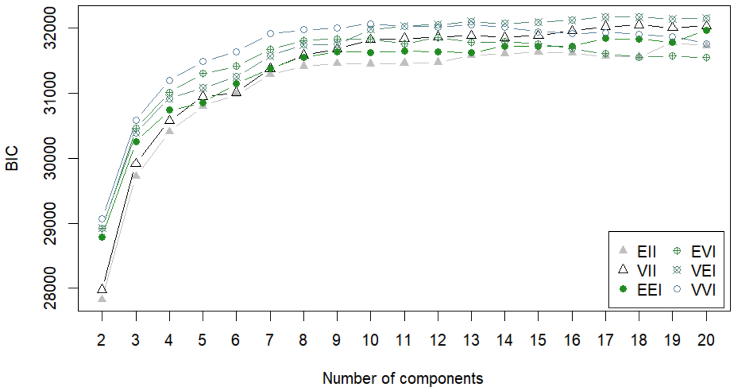

Figure 8:BIC Plot for Model-Based Clustering

In the above plot we see the evaluation of the fit of different Gaussian Mixture Models (GMMs) to the age composition data of the Greek municipalities. The y-axis represents the Bayesian Information Criterion (BIC) values, while the x-axis shows the number of clusters ranging from 2 to 20. Each line corresponds to one of the six different covariance structure models (EII,EVI,VII,VVI,VEI,EEI). Lower BIC values indicate a better fit, so with this criterion we confirm that a VEI model with 18 clusters seems to be the better one.
We also calculated the proportion of municipalities assigned to each cluster to understand the distribution of them across the clusters, so for example in cluster 1 there is 5% of municipalities and in cluster 2 the 8% etc.

Visualization and interpretation of the results of Model Based Clustering

After applying the Gaussian Mixture Models and selecting the optimal one (VEI) and the optimal number of clusters (18) we visualized some of them in order to see how the age distribution changes in all those municipalities. Below, there is the age distribution per cluster plot visually represents the demographic composition across 18 clusters derived from our Model Based Clustering analysis. Each bar plot illustrates the relative frequency of different age groups (0-14, 15-24, 25-39, 40-54, 55-64, 65-79, 80+) within a specific cluster.

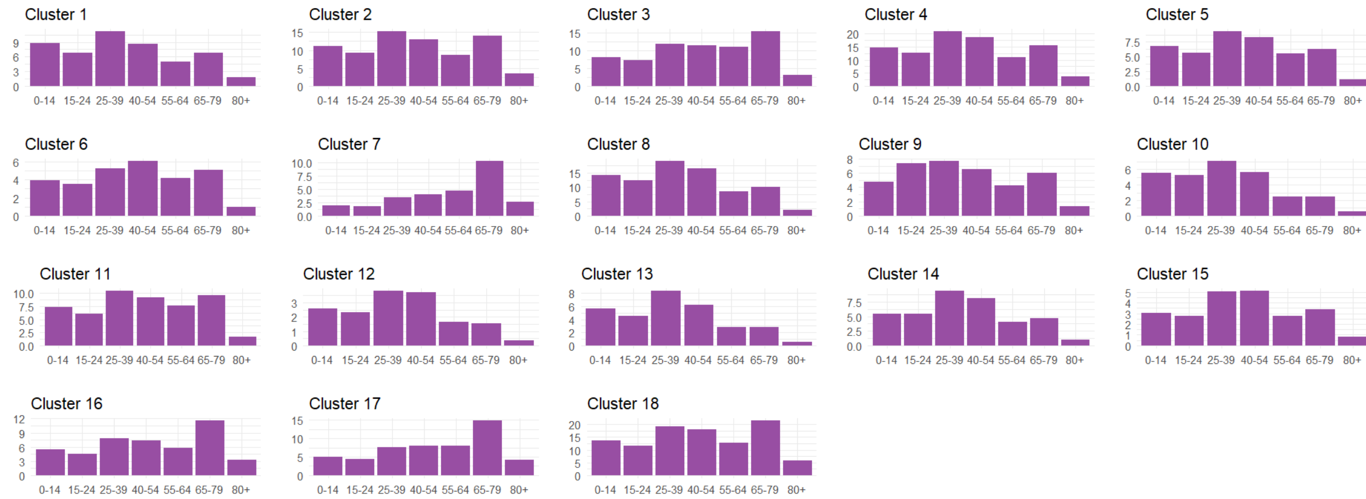

Figure 9: Age Distribution Per Cluster - Model Based Clustering

Principal Component Analysis in Model Based Clustering

Lastly, we performed some PCA in the results of the Model Based Clustering and we plot the first two Principal Components. Together those components explain about 75% of the total variance. In the plot, each point represents a municipality, and points are colored according to their cluster. We observe that some clusters are tightly grouped (e.g., Clusters 6 and 10), suggesting homogeneity within those clusters, while others are more spread out (e.g., Cluster 18), indicating greater diversity within those clusters. In addition, many clusters overlap one with each other like for example the clusters 6,8,9,11 and 10.

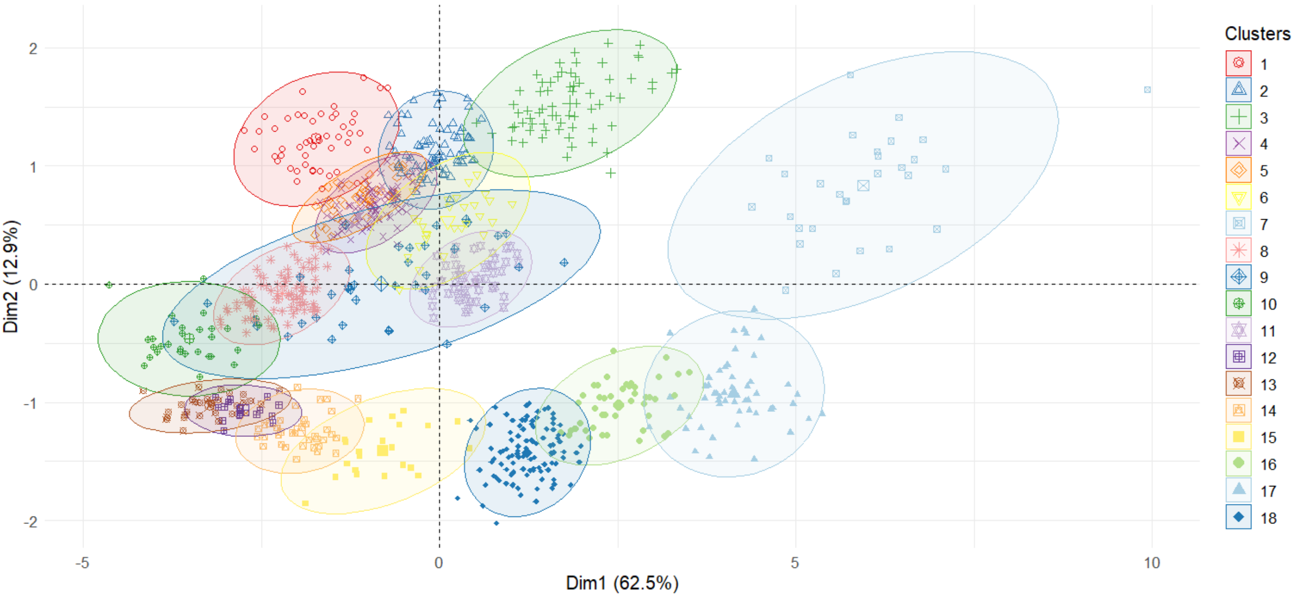

Figure 10:PCA using Mahalanobis Distance and Ward - Model Based Clustering

Comparison of the different Clustering methods and final selection

Model-based clustering and hierarchical clustering use different algorithms and criteria for forming clusters. Model-based clustering relies on statistical models, while hierarchical clustering builds clusters based on pairwise distances, which led our analysis to different cluster results. First of all, the 2 methods identified different number of clusters with Hierarchical identifying 4 using Ward's method while Model Based 18. In Hierarchical clusters tend to be more uniformly sized and shaped while we also we wanted to minimize the total within cluster variance. In Model Based we fitted Gaussian Mixture Models so clusters are formed based on those models, allowing for different shapes and sizes. Which method should we use depend of what is our purpose of our clustering, so the choice between those two methods ultimately depend on the specific objectives of the analysis and the characteristics of the dataset. In Hierarchical Clustering, the clusters were much more distinct but more generalized and we saw a broader, high-level overview of the demographic patterns across Greek municipalities. On the other hand, Model Based Clustering offered a more detailed segmentation of Greek municipalities based on their age groups. This method may allow for more precise targeting of policies and resources.

Conclusions and further research

Our analysis of the Greek municipalities grouped based on age composition, using both hierarchical and model-based clustering methods, provided valuable insights into the demographic landscape. Hierarchical clustering offered a broad overview with four generalized clusters, useful for identifying overarching trends. In contrast, Model Based Clustering, revealed a more detailed segmentation into 18 clusters. This method captured more detailed demographic patterns among the municipalities.

Our clustering analysis may be valuable for designing targeted policy interventions. Each cluster’s unique demographic profile can guide resource allocation and urban planning tailored to the specific needs of each age groups. For example, municipalities in a cluster dominated by elderly people might need and also benefit from increased healthcare services, social support for the elderly, and initiatives to attract younger populations to prevent further aging, decline of population of these municipalities and avoid the danger of getting abandoned in the near future. On the other hand, clusters with a high proportion of young families and children or young adults in the working age would need investments in schools, childcare facilities, and recreational areas.

Finally, for future research, exploring the fit of multinomial mixture models could be beneficial. This approach is well-suited for our dataset because it consists of count data and thus it could further enhance the precision and relevance of our clustering analysis and may provide us with better results.