Athens University of Economics and Business

Statistical Machine Learning-Project 2

Tsadimas Anargyros

AM:f3612318

Supervisor:

D.Karlis

Department of Statistics

June 8, 2024

Contents

List of Figures

Abstract

This project analyzes the age composition of municipalities in Greece using 2001 census data to identify clustering patterns based on demographic distributions. Employing hierarchical and model-based clustering methods, we compared the results to determine the optimal grouping of municipalities. Key findings include the identification of distinct clusters and their demographic characteristics. The effectiveness of clustering methods was evaluated using the Adjusted Rand Index and silhouette analysis. This study provides insights into the age-based grouping of municipalities and the robustness of different clustering approaches.

Introduction

This study explores the age composition of Greek municipalities using 2001 census data to identify demographic clusters. By applying hierarchical and model-based clustering methods, we aim to uncover distinct population groups and evaluate the clustering quality, providing insights into the demographic patterns of Greece's municipalities.

Chapter 1

Data Preparation and Exploratory Data Analysis (EDA)

In this section, we prepared and analyzed the age composition data from the 2001 Greek census. First, we imported necessary packages and loaded the data from an Excel file, removing any irrelevant columns and rows. We then defined and renamed the columns for clarity, converting relevant columns to numeric values for further analysis. We filtered the data to include only municipalities and excluded non-relevant columns. To standardize the data, we converted the age group counts to relative frequencies of the total population for each municipality. This means that for each municipality, the number of people in each age group was divided by the total population of that municipality, ensuring comparability across different-sized municipalities. This preprocessing ensured that our data was clean and suitable for clustering. We performed summary statistics and checked for missing values to understand the data distribution and identify any potential issues before applying clustering methods.

Chapter 2

Hierarchical Clustering

2.1 Introduction

Hierarchical clustering is a method of grouping similar data points into clusters based on their similarities.The process involves calculating a distance measure to assess the similarity between data points. Then, a linkage method determines how clusters are formed. The result is typically visualized using a dendrogram, a tree-like diagram that shows the hierarchical relationship between clusters. By cutting the dendrogram at a specific level, we can decide the number of clusters. This method helps in identifying natural groupings in data, making it easier to analyze complex datasets.

2.2 Distance Choice and Calculation

I experimented with four different distance measures: Euclidean, Mahalanobis, Gower, and Manhattan, to determine the best clustering approach. I chose Mahalanobis distance because it accounts for correlations between variables and scales the data appropriately, making it suitable for our demographic data. Mahalanobis distance measures the distance between a point and a distribution, effectively handling variables with different scales and correlations. To calculate the distance matrix, I computed the Mahalanobis distance between all pairs of data points, resulting in a matrix that reflects the true multivariate relationships within the dataset. This approach provided a more accurate basis for clustering our municipalities.

2.3 Linkage Method

I tested five different linkage methods: Ward, Ward.D2, Single, Complete, and Average, to identify the most suitable clustering technique. I chose Ward.D2 because it minimizes the variance within clusters and produces more balanced clusters, which is ideal for our demographic data. Ward.D2 linkage method merges clusters in a way that the increase in total within-cluster variance is minimized. This method is especially effective in forming compact and spherical clusters. To implement this, I applied the Ward.D2 linkage method to the previously calculated Mahalanobis distance matrix, resulting in a hierarchical clustering structure. This approach ensured that the clusters were formed with minimal internal variance, enhancing the overall clustering quality.

2.4 Determining the Number of Clusters

To determine the optimal number of clusters, I first plotted a dendrogram using the Ward.D2 linkage method applied to the Mahalanobis distance matrix. The dendrogram visually represents the hierarchical clustering process, showing how clusters merge at different levels of similarity. By examining the dendrogram, I identified several potential cut points where clusters are distinctly formed, indicating possible numbers of clusters (Figure 2.1).

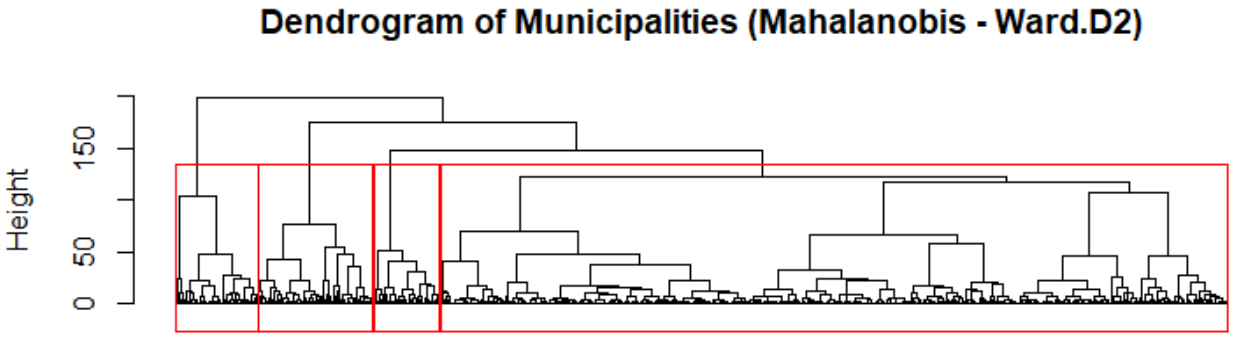


Fig. 2.1: Dendrogram of Municipalities (Mahalanobis - Ward.D2)

Next, I used the silhouette value plot to evaluate the clustering quality for different numbers

of clusters. The silhouette value measures how similar each point is to its own cluster compared to other clusters, providing an average silhouette width for each clustering configuration. A higher silhouette value indicates better-defined clusters. By plotting these values, I found that four clusters provided a good balance between cluster compactness and separation, with a relatively high average silhouette width (Figure 2.2).

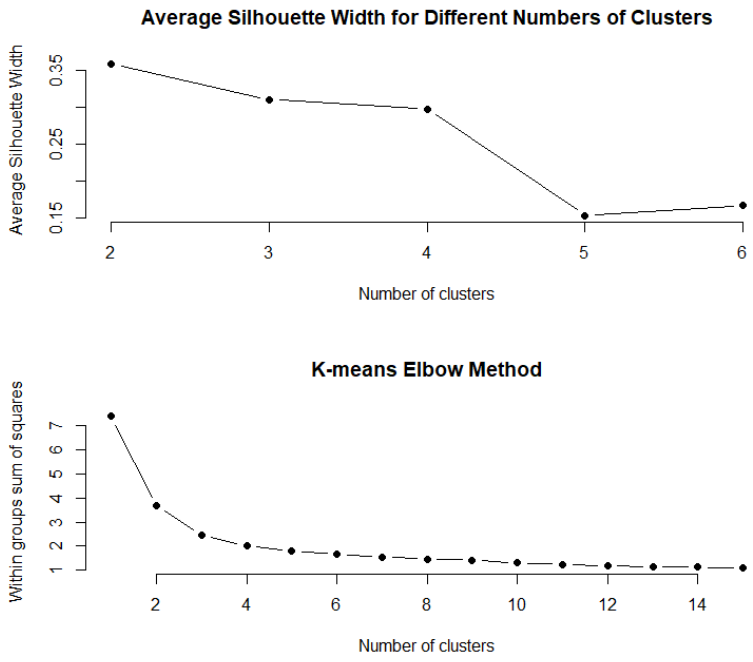


Fig. 2.2: Comparison of Methods for Determining the Number of Clusters

Additionally, I briefly employed the elbow method, commonly used in k-means clustering, to further validate the number of clusters. The elbow method plots the total within-cluster sum of squares against the number of clusters, looking for an "elbow" point where adding more clusters does not significantly improve the variance explained. This method also supported the choice of four clusters, as the plot indicated a noticeable elbow at this point (Figure 2.2). Combining these approaches, four clusters were determined to be an optimal choice for the data.

2.5 Cluster Evaluation and Visualization

To evaluate the clustering quality, I calculated the silhouette values for the chosen number of clusters. Silhouette values measure how similar each data point is to its own cluster compared to other clusters, with values ranging from -1 to 1. A higher silhouette value indicates better-defined clusters. I computed the average silhouette width for the four-cluster solution and found it to be satisfactory, indicating well-separated and cohesive clusters (Figure 2.3). Additionally, I compared the silhouette scores for different numbers of clusters, confirming that four clusters provided a relatively high average silhouette width, reinforcing the choice.

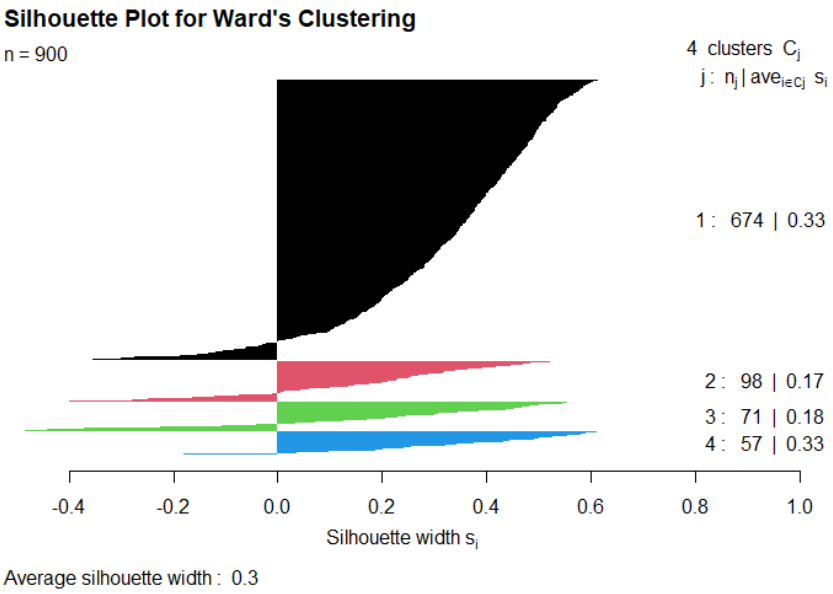


Fig. 2.3: Silhouette Plot

To visualize the clusters, I employed Principal Component Analysis (PCA). PCA reduces the dimensionality of the data, allowing us to plot it in a two-dimensional space. I then plotted the clusters identified by hierarchical clustering on the PCA plot (Figure 2.4). This visualization helped interpret how the clusters were formed and their separations in reduced dimensions. Each point in the PCA plot represented a municipality, colored by its cluster membership, and ellipses were added to highlight the cluster boundaries. Additionally, I plotted the clusters identified by model-based clustering for comparison (Figure 2.4).

Finally, I analyzed the demographic characteristics of the clusters by examining the proportions

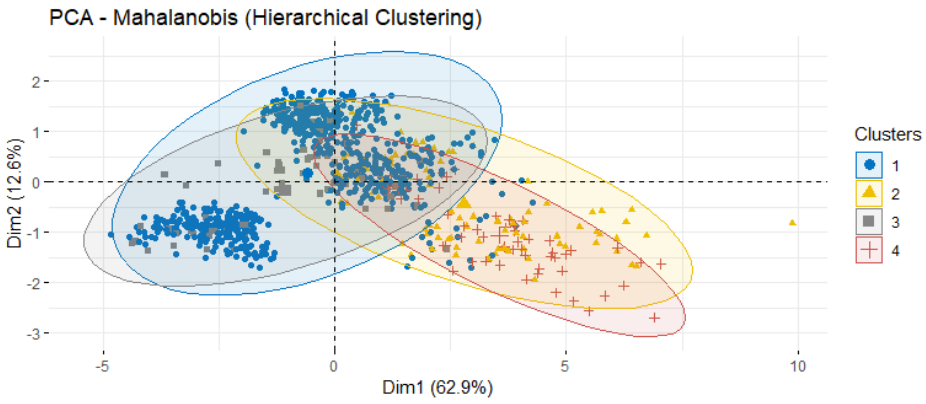


Fig. 2.4: PCA - Mahalanobis (Hierarchical Clustering)

of different age groups within each cluster, providing insights into the distinct demographic patterns that define each cluster (Figures 2.5). This comprehensive evaluation ensured the robustness and interpretability of the clustering results.

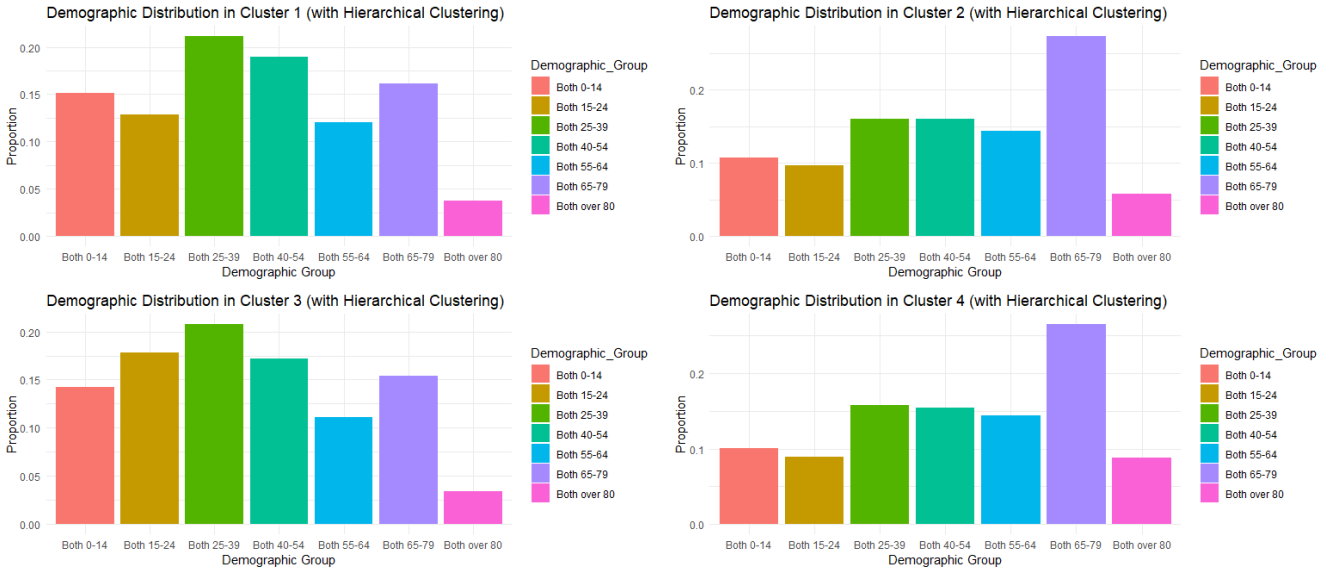


Fig. 2.5: Demographic Distribution in Clusters (with Hierarchical Clustering)

Cluster 1 shows a higher proportion of age group 25-39, indicating areas with young families and working-age adults, Cluster 2 has a significant proportion of older adults, suggesting remote or abandoned areas with an aging population, Cluster 3 exhibits a balanced age composition, reflecting typical residential areas with diverse populations. Cluster 4 shows a high proportion of older adults but with a more balanced representation than Cluster 2, indicating less remote but still aging areas.

Chapter 3

Model Based Clustering

3.1 Introdution

Model-based clustering assumes that the data is generated from a mixture of underlying probability distributions, each representing a different cluster. This method identifies the parameters of these distributions and assigns data points to the most likely cluster based on statistical criteria. Benefits of model-based clustering for demographic data include its ability to handle complex data structures and provide probabilistic cluster assignments. However, it can be very sensitive to the choice of the model and assumptions about the data distribution. Common assumptions include the type of distribution (e.g., Gaussian), the independence of features, and the homogeneity of variances within clusters. Incorrect assumptions can lead to varying and potentially unreliable clustering results.

3.2 Methodology and Model Selection

To determine the optimal number of clusters, we used the Bayesian Information Criterion (BIC). BIC evaluates the fit of different clustering models by balancing model complexity and goodness of fit, with lower BIC values indicating better models. We explored various models, including EII, VII, EEI, EVI, VEI, and VVI, each representing different assumptions about the shape, volume, and orientation of clusters. By calculating BIC values for models with different numbers of clusters, we identified the model with the lowest BIC as the best fit. This approach ensures that the selected model not only fits the data well but also avoids overfitting by penalizing unnecessary complexity.

3.3 Cluster Assignment, Uncertainty and Determining the Number of Clusters

By calculating the BIC values for models with different cluster counts, we found that the optimal number of clusters increased to 14. The model with the lowest BIC value was the VVI model, which assumes variable volume and shape ellipsoids, providing a flexible and accurate representation of the data. This model's features allowed it to capture the underlying structure of the demographic data more effectively, ensuring robust and meaningful clustering results. This rigorous approach to selecting the number of clusters ensures that our clustering solution is both statistically sound and interpretable.

3.4 Cluster Evaluation and Visualization

To visualize the clusters, I applied Principal Component Analysis (PCA). PCA reduces the dimensionality of the data, allowing us to plot it in a two-dimensional space. I then plotted the clusters on the PCA plot, with each point representing a municipality and colored by its cluster membership (Figures 3.1). This visualization helped interpret the cluster structure and patterns across different municipalities.

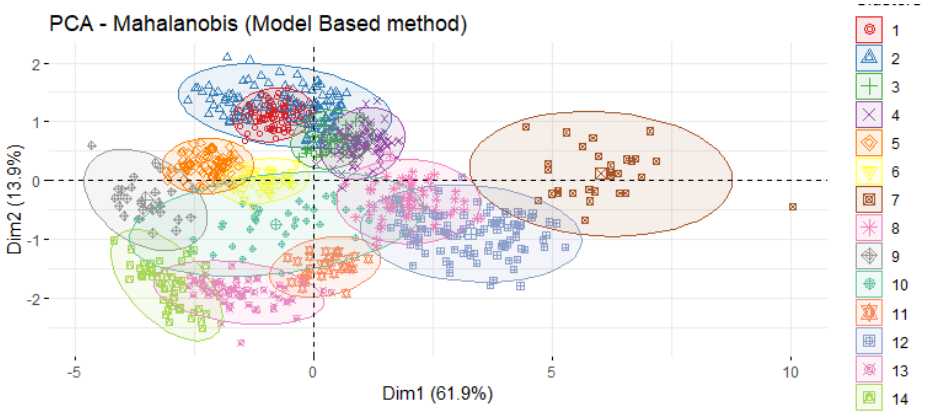


Fig. 3.1: PCA - Mahalanobis (Model Based method, 14 Clusters)

Additionally, I experimented with different numbers of clusters to observe variations in age group compositions. By examining the resulting bar plots for each clustering solution, I aimed to find

distinct and significantly different age compositions. This iterative process revealed that a three-cluster solution provided the most distinct demographic differences among the clusters, ensuring meaningful and interpretable groupings (Figures 3.2). This approach highlighted the importance of not only statistical measures but also practical interpretability in determining the optimal number of clusters.

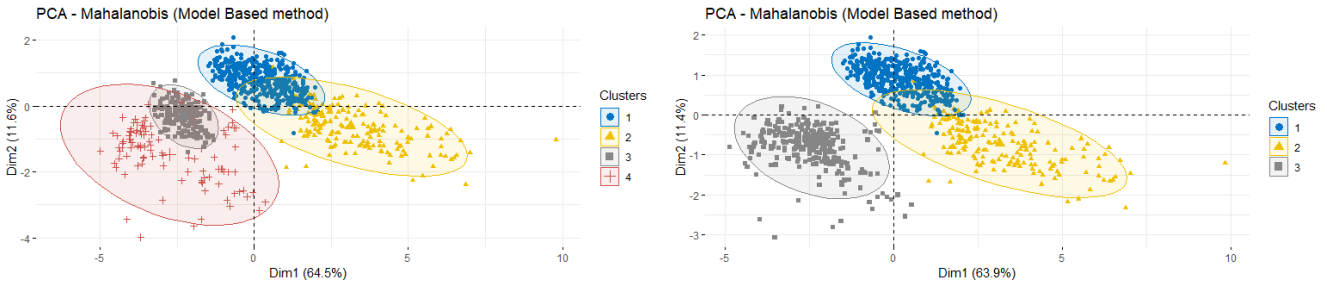
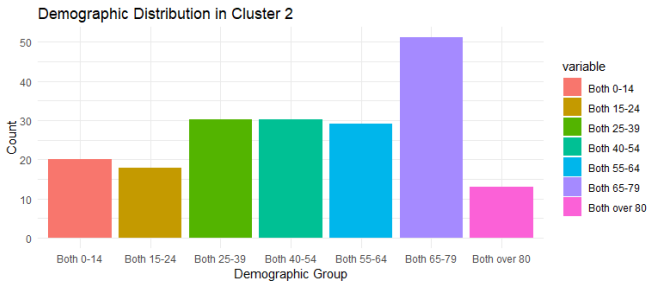
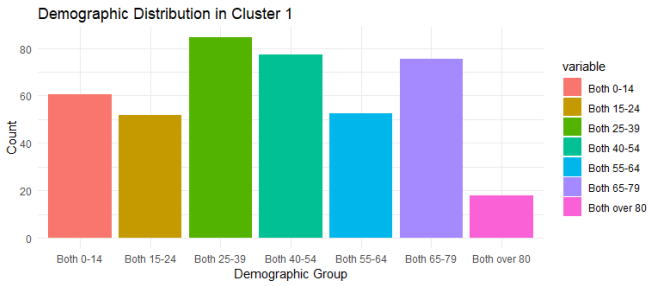


Fig. 3.2: PCA Visualization

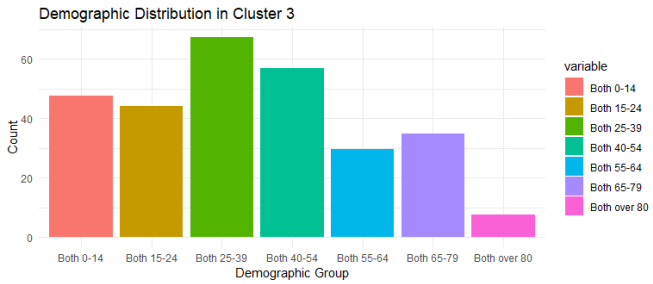
The demographic distribution plots (Figures 3.3)reveal three distinct types of municipalities based on age composition. Cluster 1, with a higher proportion of older adults, indicates remote or abandoned areas with an aging population. Cluster 2 exhibits a balanced age composition, reflecting typical residential areas with a diverse population. Cluster 3 is dominated by young adults, suggesting the presence of universities and educational institutions that attract a younger demographic. These clusters highlight varying demographic patterns, providing insights into the socio-economic characteristics of different regions within the country. This information is crucial for targeted policy-making and resource allocation.



(a) Cluster 1



(b) Cluster 2



(c) Cluster 3

Fig. 3.3: Demographic Distribution in Clusters (3 Clusters Solution)

Chapter 4

Comparison of Clustering Methods and Final Selection

Finally, I compared the hierarchical and model-based clustering methods to determine the most suitable approach. Using the Adjusted Rand Index (ARI), I quantified the agreement between the two methods, obtaining an ARI of 0.04. This low value indicates that the clusters produced by the two methods are quite different and only slightly better than random chance. Given the significant differences, I evaluated the practical interpretability and robustness of each method. The hierarchical clustering method was chosen because it provided more meaningful and distinct clusters, as evidenced by clear dendrogram cut points, better Demographic Distribution Bar Plots and agreement with a third method. Additionally, hierarchical clustering aligned better with domain knowledge, ensuring more reliable and interpretable demographic groupings for the municipalities.

Conclusions

This study effectively identified and characterized clusters of Greek municipalities based on age composition using the 2001 census data. By applying both hierarchical and model-based clustering methods, we explored different approaches to grouping the data. Despite the low ARI score indicating significant differences between methods, hierarchical clustering was chosen for its meaningful and interpretable results. The analysis revealed distinct demographic patterns within the municipalities, providing valuable insights into the age distribution across different regions. This work demonstrates the importance of selecting appropriate clustering methods and validates hierarchical clustering as a robust technique for demographic data analysis.

Bibliography

[1] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, *An Introduction to Statistical Learning with Applications in R*, Springer Texts in Statistics, New York, NY: Springer, 1st edition, 2013.

[2] Dimitris Karlis, *Statistical Machine Learning Lecture Notes*, Department of Statistics, Athens University of Economics, karlis@aueb.gr, Athens, Feb, 2024.