NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

A Socio-Economic Portrait of the Autonomous Region of the Azores

Sofia de Medeiros Cabral

Dissertation presented as partial requirement for obtaining the Master's degree in Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

A SOCIO-ECONOMIC PORTRAIT OF THE AUTONOMOUS REGION OF THE AZORES

by

Sofia de Medeiros Cabral

Dissertation presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Business Intelligence and Knowledge Management

Advisor Prof Dr. Paulo Gomes

July 2021

ABSTRACT

The present study aims to deepen the knowledge in the Autonomous Region of the Azores' sub-regional areas. By applying Principal Component Analysis and Cluster Analysis to a set of essential variables of this region's census data, one can study the relation between those sub-regions and the chosen variables at the municipality level. This type of analysis is useful in the sense that by characterizing a sub-region, one can withdraw the significant influencers of its socio-economic outcomes. Moreover, due to its natural dispersion, being able to group the subregions or municipalities by similarity might be a pivotal factor to apply the right governmental policies to each group by playing an important decision-making criterium for territorial planning and economic development.

KEYWORDS

Sub-regional; Azores; Principal Component Analysis; Cluster Analysis; census data; socio-economic outcomes; territorial planning; economic development

INDEX

FIGURE INDEX

TABLE INDEX

LIST OF ACRONYMS AND ABBREVIATIONS

INE	Instituto Nacional de Estatística
GDP	Gross Domestic Product
NUTS	Nomenclatura das Unidades Territoriais para Fins Estatísticos
PCA	Principal Component Analysis
KMO	Kaiser-Meyer-Olkin
MCD	Minimum Covariance Determinant

1. INTRODUCTION

The Autonomous Region of the Azores is composed of nine spread-out islands divided into three groups: Eastern Group (São Miguel and Santa Maria), Central Group (Terceira, Faial, Graciosa, São Jorge, and Pico), and Western Group (Flores and Corvo) with a total of 2 333 km^2 of land that is home to 246 746 habitants, according to the 2011 census' data (Instituto Nacional Estatística-INE). The Gross Domestic Product (GDP) oscillates between 3 to 4 billion euros, where each island's relative contribution is significantly distinct. Even though in 2011, according to the Regional statistics data, São Miguel contributed for around 59% of total regional GDP, it is not the island with the highest GDP per capita, being topped by Faial, Corvo, and Santa Maria. The lowest contributor was Corvo island, equaling 0.2% of total GDP, with GDP per capita higher than São Miguel (16 427€/habitant > 16 063€/habitant). Santa Maria was the island with the highest GDP per capita of 18 625€/habitant while it only contributes for 2.80% of total regional GDP. This means that populational distribution is uneven when compared to the islands' production. When comparing some of the significant economic or social indicators, a visible distinction is seen between the different islands and between cities of the same island and municipalities of the same town. Take, for instance, the case of the per capita purchasing power. If one looks to an island level, one will say that Corvo is the island with the lowest purchasing power per capita (63.1%). However, there are at least four cities outside Corvo with lower purchasing power: Nordeste, Povoação, Vila Franca do Campo in São Miguel (55.9%, 57.8%, and 59.2%) (data from INE).

An extensive analysis using smaller geographical units might prove to be beneficial for economists and politicians to better formulate regional level policy by withdrawing patterns in data otherwise unseen before.

A socio-economic portrait aims to go beyond the available research results and provide an in-depth view of a geographical area. The method and variables used highlight this work compared to other studies done for the Azores region. Other regional analyses have been done recently, using an economic model to estimate the major determinants of employment and other socio-economic variables (Pavão et al., 2020). However, the analysis is made at the island level. As was presented before by Soares et al. (2003), it is more useful to perform such study at a smaller economic unit, as is the municipality's case. As such, what is proposed is a

socio-economic characterization at the municipal level of every sub-region of each island of the Azores.

This work's expected contribution relies on going to the crux of why some statistical outcomes are the way they are or the major statistically significant reasons for them to be like that by finding the hidden relationships between socio-economic outcomes and each municipality. In this sense, one can deepen the knowledge around how the regional structure functions by finding the not so apparent reasons why some islands or subsections of certain islands are lesser or more developed than others. One should not treat an island as a homogeneous geographical area because it most certainly is not. Some studies conclude that intra-regional dissimilarities are untreated when using a bigger geographical unit that generalizes a sub-regions outcome, especially sprawling areas (Boldea et al., 2012; Zambon et al., 2017). Some of these associations might be "common sense" or seem to be "just like that", but the idea is to quantify these relations and find statistical meanings to better comprehend their impact.

This work will be divided into seven sections beginning with a theoretical background, where previous studies of regional socio-economic disparities will be analyzed, followed by the research model used and respective methods. Subsequently, the data analysis results will be presented along with a discussion of its implications. Lastly, conclusion notes will be drawn, as well as this work's limitations and suggestions for future works.

2. LITERATURE REVIEW

2.1. THEORETICAL BACKGROUND

Regional planning and regional policy-making are aided by the diversity of studies provided with a particular area's data. Regional socio-economic portraits might analyze several indicators from economic outcomes like GDP or unemployment rate; health as in the number of hospitals, for instance; professional qualification translated in the Index of tertiarization or the Theil Index; demography indicators like the average population age or the elder dependency index, amongst others. They can use a smaller geographical unit than a city or region. The degree of complexity of such portrait using smaller economic and geographic units depends on the variety of variables and indicators chosen or methodology used to study them.

The ultimate goal of such a portrait is to characterize each sub-region so that it is possible to group sub-regions by similarity and find common ground upon which their characterization is being influenced. For example, some indicators might be more significant for some sub-regions while others might be completely insignificant, and by finding these relations, a more interpretable insight can be drawn for each sub-region. Seeing these differences is crucial to better understand the socio-economic outcomes of sub-regions and improve the policies applied to them.

Many socio-economic studies have been made throughout the years following different processes since researchers do not always agree on the methodologic procedures. However, it has been established before that policy-makers should go beyond the study of differences between regions and start to look at intra-region socio-economic disparities to understand better a region's specificities (Lipshitz & Raveh, 1994). It is wrong to treat countries, cities, or sometimes municipalities as homogeneous regions due to uneven development inside the same region, translating into a fragmented landscape. Even though the living conditions might decrease by how significant the distance from the major metropolitan area is, the low income in areas further from the leading economic points, like interior or rural areas, is somehow compensated by the low housing costs, for instance. Fundamentally, one can classify a region as being developed with good socio-economic outcomes when considering a bigger geographical unit, like a city, but still find "pockets of poverty" within that same city, being the reverse also true (Pettersson, 2001).

This heterogeneous landscape can be explained by different development rates, as explained before, but also an uneven distribution of technical and social infrastructures followed by different resource accessibility and demographic imbalances. Then, the productivity level will be different for each sub-region (Boldea et al., 2012). Considering all this, the indicators to use in a portrait like the one proposed should be diverse. They should account for the different socio-economic areas that affect a determined region.

As it is mentioned in Soares et al. (2003), the use of smaller geographical units and a diverse set of indicators can characterize a sub-regions degree of development, showing weaknesses on the NUTS II classification broadly used.

This type of reasoning also diverges from some European Commission methods of classifying regions by only using its GDP (Cziráky et al., 2003). As it can be imagined, when classifying a sub-region, its GDP value is sometimes hard to find or even inexistent for a smaller unit like a municipality. As for remote places, one cannot merely withdraw a sub-regions socio-economic portrait by merely comparing its proximity to its core region's bigger classification unit, for example, a city, since locality proximity is quite different from socio-economic proximity (Rovan & Sambt, 2003). Furthermore, sprawling areas tend to demonstrate higher socio-economic disparities than compact settlements (Zambon et al., 2017). This can be very important when studying the Azores case since it is an autonomous region divided by islands where there are fewer compact settlements when compared to the sprawled ones. Also, the access between villages inside the same island is sometimes limited, revealing the importance of each sub-region's institutional factors, from the natural conditions to the actual geographical location (Wang, 2016).

The use of smaller geographical units to characterize sub-regions has proven useful before while using different study methods. For example, this is the case of an exploratory and factor analysis model used to study Croatian municipalities (Cziraky et al., 2002), an expert and population poll done in Russia using a direct estimation Ball method (Sayfudinova et al., 2016), cluster analysis applied to municipality data from rural Sweden (Hedlund, 2016) or in Slovenia (Rovan & Sambt, 2003), a Theil index decomposition method for China (Wang, 2016) and, for instance, a Composite Index of Infrastructure that compares the degree of development between infrastructure services in India (Patra & Acharya, 2011).

Below is a summarized table with all indicators studied.

Study's Name	Country	Indicators
A multivariate methodology for modelling regional development in Croatia (Cziraky et al., 2002)	Croatia	Income per capita, Population share of income, Municipality income per capita, Employment rate, Unemployment rate, Social aid per capita, Age index, Density, Vitality Index, Distance, Population trend
Methodological basis of the regional systems socio-economic profile using survey method (Sayfudinova et al., 2016)	Russia	Expert poll (more than 30 economic, demographic, social and environmental indicators) and population poll (consumer moods, current status of economy, consumer expectations, consumer activity, independence worthiness and manpower mobility)
Mapping the Socioeconomic Landscape of Rural Sweden: Towards a Typology of Rural Areas (Hedlund, 2016)	Sweden	Share of the working population aged 18–64 working with: agriculture, forestry, mining, manufacturing, tourism, and finance and other sectors requiring university education, Share of population aged 18–64 established in the job market, Share of population aged 18–64 with a university degree, Females aged 15–45 as a share of the population, Share of the population aged 65+, Population difference 1985–2008
Socio-economic Differences Among Slovenian Municipalities: A Cluster Analysis Approach (Rovan & Sambt, 2003)	Slovenia	Aging Index, Index of population growth, Index of daily migration, Income tax base per capita, Share of agricultural population, Unemployment, Number of students per 1 000 inhabitants, Number of cars per 100 inhabitants
Analysis on the Regional Disparity in China and the Influential		GDP per capita, Urban household disposable income per capita, Rural household net income per capita

Factors(Wang, 2016)	China	
Regional Disparity, Infrastructure Development and Economic Growth: An Inter-State Analysis (Patra & Acharya, 2011)	India	Percentage of villages electrified, Per capita consumption of electricity, Length of road, Length of railway route, Vehicle density, Percentage of villages connected by roads, Number of post offices, Number of banks, Number of mobile consumers, Registered motor vehicles

Table 1 Adoption models at the individual level

3. STUDY'S ADAPTATION

The method to be applied to Azores' municipality data follows the INE (2004) Socio-Economic Portrait of the Metropolitan Area of Lisbon. However, it utilizes a more tailored set of variables, including more than 20 indicators going through housing, education, health, amongst others, to be studied with a multivariate principal component analysis followed by cluster analysis. The data used is derived from census data from 2011, which is a type of data with high reliability since the major statistics entities in Portugal verify it, and it is revised according to several accuracy parameters by external evaluators. The use of several confirming methods leads to a cohesive data source, which leads to a more enriched study. As for this study's analysis, the following variables will be used:

Active Variables	
Prop of buildings not exclusively residential	(Buildings partially residential + Buildings principally not residential)/ Total of buildings*100
Prop of leased or sub-leased classic family accommodation	[Rented conventional dwellings (with fixed-term contract, contract without-term, social or support income or sub-rented)/ Conventional dwellings of usual residence]*100
Prop of own housing with charges	[(Owner occupied dwellings (with mortgage due to to acquisition of the dwelling))/ (Owner occupied dwellings)]*100
Prop of overcrowded accommodation	(Overcrowded dwelling(lacking one room, two rooms, three rooms or more)/ Homestays of habitual residence)*100
Average age of resident population	Sum of ages of the resident population/ Resident population
Average age of buildings	[(Number of buildings aged in class j * middle point of class j)/ Total buildings]
Prop of resident population working or studying in another municipality	(Resident population that works or studies in other municipality/ Resident population that works or studies)*100
Proportion of car use when traveling	((Car-driving or passenger)/ (Resident population that works or studies))*100
Prop of resident population with 15 and more years old whose main livelihood is work	(Resident population with 15 and more years old whose main livelihood is work/ Resident population with 15 and more years old)*100
Prop of resident population that 5 years previously lived outside the municipality	(Resident population that 5 years before inhabited outside of municipality/ Resident population)*100
Prop of single-person classic families	(Single-person families/ Classic families)*100
Prop of classic families with 5+	(Classic families with 5 or more members/Classic families)*100
Prop of family nuclei of couples with children	(Family nuclei of couples with children/ Family nuclei of couples)*100
Prop of buildings needing major repairs or degraded	(Buildings with large repair needed or most degraded/ Buildings)*100
Average households per accommodation	Classic families/Conventional dwellings of usual residence
Prop of resident population of foreign nationality	(Resident population of foreign nationality/ Resident population)*100
Prop of socially most valued professionals	[(Employed population (CPP=1 ou CPP=2))/ Employed population]*100
Unemployment rate	(Unemployed population/ Active population)*100
Elderly dependency index	[(P(65, +)/ P(15, 64))]*10^n
Prop of resident population (Who has lived abroad for a continuous period of at least 1 year)	(Resident population (Who has lived abroad for a continuous period of at least 1 year)/Resident population)*100
Proportion of dwellings with heating	(Dwellings with heating/Conventional dwellings of usual residence)*100
Prop. Of population 15+ with no school level completed	(Resident population with 15 and more years old without any level of education completed/ Resident population with 15 and more years old)*100

Supplementary Variables	
Prop buildings with 3+ accomodations	(Buildings with 3 or more accomodations/ Buildings)*100
Index of tertiarization	[(Population working on the first sector*Proportion of first sector workers)+(Population working on the second sector*Proportion of second second workers)+(Population working on the third sector*Proportion of third sector workers)]/3
Theil Index	[-(sum((proportion of socio-economic group j)*ln(proportion of socio-economic group j))/ln(number of socio-economic groups)]

Table 2 Variable Set for this Study

It is crucial to have a varied set of indicators to tackle the heterogeneity of a region and the chosen variables reflect this need. For a region like the Azores, some indicators play an important influence, for instance, the average age of buildings, since some municipalities have a higher rate of newly constructed buildings while others are fairly old. Others go to the crux of the differences in family structures like single-person families, families with more than five members, or even family nuclei with children. Historically, Azores tends to have a more considerable amount of family nuclei with more members. However, even though that number has been converging to the national average, Azores still has a higher teenage pregnancy rate of 10,8% against 6% of all Portugal (Santos, 2014). This also influences the housing and urbanization matters like the average number of households per accommodation or the building's overall condition.

These are all factors that might influence the living conditions of each sub-region and consequent poverty dissimilarities. As stated in Diogo (2019), different levels of poverty and inequalities in income distribution might be reflections of "poli-insularity". This concept relies upon the fact that the region receives different amounts of governmental social income redistributions due to uneven population and economic activity distribution. The Regional Government introduced this "poli-insularity" concept in order to contest the fact that one island, São Miguel, retains most of the population of the region and consequently results in the blasting of several issues, and introduce cohesive policies that intent to aid families and companies from smaller and more remote islands. So, besides the political and economic matters, it also considers the social issues that lead to the region's fragmented territorial landscape.

Another important factor for the fragmented socio-economic landscape is that services and state infrastructures are not equally available for all islands or even some municipalities of the bigger islands. For instance, Corvo island has no social work activities or social action services. The detachment of certain services might contribute to higher poverty levels and the need for social income redistribution (Diogo, 2019). Access to such social aid services, along with education and health infrastructures and job market offers, pay a strong influence on the birth rate and consequent populational density of certain regions (Santos, 2014).

As stated before, certain islands' demographic weight might influence economic tendencies while compared to the others, but also social propensities. For instance, larger

families tend to be more dependent on social incomes, and the region is characterized by having more large families and families with children receiving this aid. Naturally, bigger families have a higher risk of needing social assistance since work income is shared amongst more people. Another example is that many men who receive these incomes are workers with lower job market-specific qualifications reflecting a lower income from work. The existence of these types of jobs, characteristic of the region, like agriculture, fisheries, and construction work requiring fewer qualifications (easier access), is attractive to a younger population, and might influence their early dropout from school. Some sub-regions combine these two examples, which means that income per capita in such households might be below the poverty threshold, hence the need for social aid in the first place.

As can be seen, a socio-economic portrait of a region this fragmented justifies the need to take more indicators into account than just the economic ones. The goal is then to be able to combine these indicators in a useful and insightful manner such that each municipality can be portraited and grouped to find which sub-regions need which aid or why some regions behave the way they do.

4. RESEARCH MODEL

The goal of this study is to be able to find patterns unseen before. The usage of a smaller geographical unit comes as a tool to be able to segment sub-regions in order to tackle the possible heterogeneity of a certain region.

As explained before, the nature of the Azores region can influence certain sub-regions' remoteness. Some areas of bigger islands can be as remote as areas of smaller islands depending, amongst other things, on their territorial land access. Not every sub-region is as attractive as the main city of São Miguel or Terceira islands. These are some examples of differences intra-region but similarities between sub-regions. The question is, are these factors distinct for some sub-regions in order to motivate a municipality analysis and consequent municipality-driven governmental investment or aid.

Does the regional government need to pay attention to the heterogeneous landscape of some regions?

Is there a Socio-Economic Portrait capable of describing the sub-regions' heterogeneity?

These questions want to answer the fact that the socio-economic dimensions that will be used in this work, characteristic of the population's living conditions, might be more predominant in certain sub-regions. If this is true, then policies and economic aid given at the island level or even at the city level are unfitting. By undertaking those sub-regional differences and trying to answer their specific needs, those areas' living conditions might increase since they finally receive the fair aid they need.

This means that, for the indicators considered, different grouped influences and relationships can possibly describe and differentiate territorial characteristics.

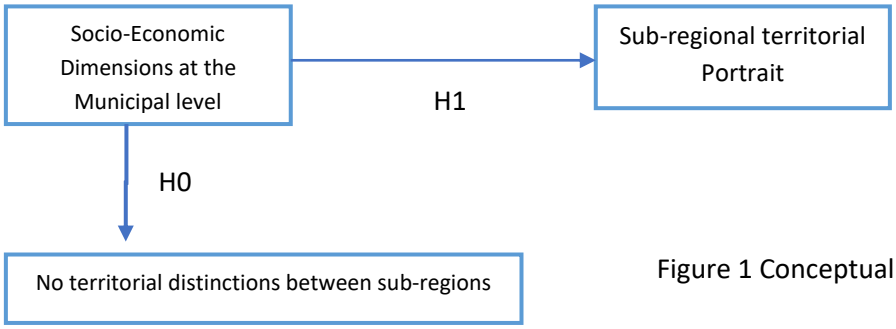


Figure 1 Conceptual Model

Hypotheses:

H_0: Indicators do not disclose groups of sub-regions with distinct characteristics

H_1: Indicators possibly describe groups of sub-regions with distinct territorial characteristics

	PCA and Cluster Analysis
H_1	Clear territorial distinction that motivates a Socio-Economic Portrait Municipal level analysis is justified

Table 3 Instrument Table

5. METHODS

In terms of data, it mainly came from the census of 2011 made available by INE, suffering minor transformations into proportions or major ones as is the case of indexes, for instance, the Index of tertiarization or Theil Index, both used as supplementary variables. Other indicators, as is the case of the Proportion of Population with Foreign Nationality, Proportion of car use when traveling or School Dropout Rate, amongst others, were collected from the INE database, under the condition that it was stated that the values were collected at the date of the 2011 census. Thus, it ended up with 22 active variables for 156 municipalities. *A posteriori*, three supplementary variables will be studied, and nine supplementary individuals representing the region's nine islands will also be studied according to the methodology's outcomes.

Before starting any analysis, the data will be checked in order to confirm its adequacy to the proposed methods. As such, a Bartlett's test of sphericity is going to be performed. This test checks if the data is redundant enough to apply a factor analysis by comparing its correlation matrix to the identity matrix. In order to reject the null hypothesis of having a matrix too close to the identity, this value should be lower than the significance level. Adding to this indicator is the Kaiser-Meyer-Olkin measure. This one checks the proportion of underlying factors common to the variables used in the study. It is expected to have values closer to 1 (or higher than 0.5) since this indicates that variables are suitable for factor analysis and observations can be grouped (Ul Hadia et al., 2016). R Studio was used to measure both indicators.

As for the methods themselves, a univariate data analysis was first performed to check for outliers and abnormal variability, followed by a bivariate analysis, studied through the scatter plot and correlation matrix. Due to the outlier behavior of some municipalities in several variables, it was decided to do a multivariate outlier analysis before any further analysis since the standardization process used while computing the multivariate analysis chosen cannot resolve this multivariate outlier behavior, being highly influenced by it in return.

To identify possible multivariate outliers, the Mahalanobis distances are going to be calculated. According to its size, these distances tell how far an observation can be from the center of the observational cloud. The limitations regarding this indicator rely on the *masking*

effect it can suffer from the proximity of possible outliers. When some outliers have a strong influence, they can skew the mean and covariance towards them, resulting in a smaller distance between those outliers and the mean, as well as outliers closed to them (L, 2017). A robust estimator like the Minimum Covariance Determinant (MCD) can be calculated to tackle this issue. This estimator is less sensitive to the outlier behavior explained previously, distinguishing the outliers with greater influence upon the study. The algorithm used on the MCD method is called *FAST-MCD*, which reliably computes a robust distance without the extensive calculations done with other algorithms (Hubert et al., 2005). Once the multivariate outliers are found, they will have a passive status until the end of the multivariate analysis, where they will be studied à posteriori. R Studio was used to perform this analysis.

After that, the indicators correlated with each other were used to apply a Principal Component Analysis (PCA). This analysis can be performed on the correlation matrix since the data does not have the same units.

A Principal Component Analysis creates a set of new variables (components) as a linear combination of the initial set of centered variables that potentially preserve a good percentage of the initial data variability, thus, not losing too much inertia even if the final set is smaller than the initial (Jr et al., 2018). Furthermore, each new component (latent variables) is correlated with some dimensions considered to comprehend better the differences between municipalities of different Azorean islands or even from the same island.

Principal Components have a decreasing variance meaning that the first one retrieved by the software (JMP) is the component that explains the most variance of the initial variables. The second component, not correlated with the first one, explains most of the remaining variance not explained by the first, and so on. In this way, one can retain a group of non-correlated components that explain a critical percentage of the initial variance, which will reduce the initial set of variables into a smaller one, easier for interpretation.

According to the Kaiser criterium, from the total number of components given by the software (JMP), the optimal number will be chosen through a scree plot graphical analysis or by selecting the components whose eigenvalues are greater than one. Then, all components will be named after a varimax matrix rotation since this method maximizes the sum of the

variances of the squared correlations between variables and factors. Thus, it is easier to draw relationships between a group of variables and identified components.

Afterwards, principal planes can be studied, and some relevant municipalities and variables on those planes can be identified. The idea is to study the relationship between a sub-group of municipalities and variables and the principal components in order to determine each municipality's specific characteristics and its relation to all variables. Two different scores will be used to identify the relevant municipalities or variables: the partial contribution (CTR) and the squared cosine (COS^2). The CTR, either of variables or municipalities, gives the contribution of a single variable or municipality to the component's inertia, summing to 1. In this case, the total amount of inertia is equal to the number of variables, 22. The COS^2 gives the part of inertia or variability of the municipalities or variables explained by the retained components. Municipalities or variables with a CTR above average are considered relevant for a certain component's representation since they replicate a considerable amount of the component's inertia. However, it was added to the relevant group for some cases, the municipalities or variables with a high percentual COS^2 for a specific component. That is, even if a municipality or variable had a contribution below average, it might be important to analyze in a certain component's representation if the percentage of squared cosine explained by that component is high since most of the variability of that municipality or variable is explained by that component. In this particular study, this procedure gave adequate visibility to municipalities with smaller weight in the principal component analysis.

After analyzing each new dimension's representativeness in each municipality of the region, a Cluster Analysis will be applied to the principal components score retrieved for each municipality. Several multivariate procedures are applied to perform a Cluster Analysis. The idea is to classify each municipality by observing similarities and dissimilarities between them. Thus, municipalities can be segmented into mutually exclusive classes, more homogeneous intra-group and more heterogeneous between groups.

The procedure consists of grouping observations according to the existing data. The units belonging to one group are as similar as possible or more identical to the other units in that group than to units from other groups. The methodology used to group the active municipalities starts with an ascending hierarchical aggregating method followed by a K-means sub-optimal method. The K-means is applied secondly, since choosing random K's for the

analysis can influence the results, and thus, an informative K will be found before, on the ascending aggregation method. The aggregation method used at the ascending hierarchical aggregation was the Ward's method, where each step fuses classes where the loss of between variability is minimum (Gan et al., 2007).

Each cluster will translate a set of municipalities with similar characteristics regarding the socio-economic dimensions considered (principal components). Therefore, each cluster will represent a regional socio-economic class with distinct relationships with those dimensions, allowing to draw a portrait of the region according to the considered indicators.

Finally, the heterogeneity of each island is going to be quantified through a coefficient of dispersion. This coefficient translates the division of the standard deviation by the mean for all variables with CTR higher than average for each principal component at the island level. Then, the minimum and the maximum coefficient were taken for all principal components for each island. These values were presented on a graph that could easily show on which principal component an island presented more dispersion.

6. DATA ANALYSIS AND RESULTS

The first step of this analysis is checking if the dataset is appropriate for the chosen methods. Non-correlated variables do not motivate the factor analysis proposed, while a significant correlation between pairs of variables indicates that variables can be grouped by similarity and still be significant for the study. Thus, the connection between variables needs to be checked beforehand.

In order to confirm the relationship between variables, a Bartlett's test of sphericity was performed on the data to check the hypothesis where the correlation matrix is equal to the identity matrix at the population level. In this case, a result of a p-value of approximately 0 leads to the rejection of the hypothesis of having non-correlated variables. It is also part of the bivariate analysis to have a look at the correlation matrix, which in this case appears to have some pair of variables with a high positive or negative correlation between each other, meaning that some variables might influence others or simply behave the same way.

To check the analysis's adequacy to the data, a Kaiser-Meyer-Olkin (KMO) statistic was calculated to find the proportion of underlying factors common to all variables. The KMO statistic was 0.80, meaning that variables can be grouped, and the chosen analysis is adequate and useful.

As for the univariate data analysis, most variables exhibited outliers, either above or below the mean. By looking at each variable's variance, it was decided to keep all variables, despite their outliers, using the Proportion of Buildings with three or plus accommodations, Theil Index, and Index of tertiarization as supplementary variables.

The multivariate outlier analysis began with the calculation of the Mahalanobis distances for each municipality. The resulting graph (Figure 2) is presented next.

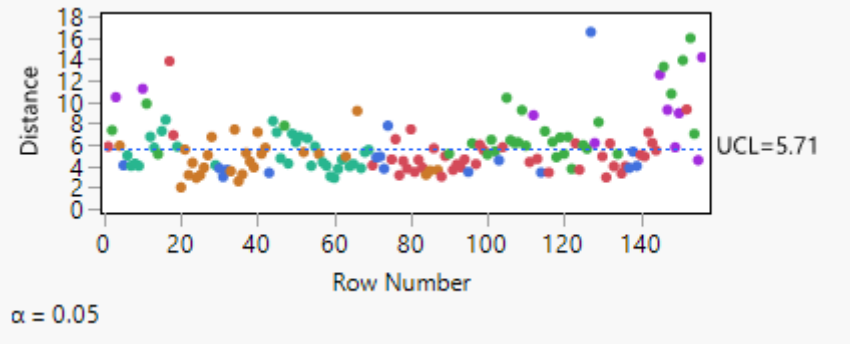


Figure 2 Mahalanobis distances of each municipality

According to this representation, there are several potential multivariate outliers. Even though the reference line is below many municipalities, that does not mean all of those are noteworthy outliers to remove and only analyze *a posteriori*. This representation is suffering from the *masking effect* explained in Section 5. As such, a Minimum Covariance Determinant (MCD) was calculated in order to have a more robust outcome.

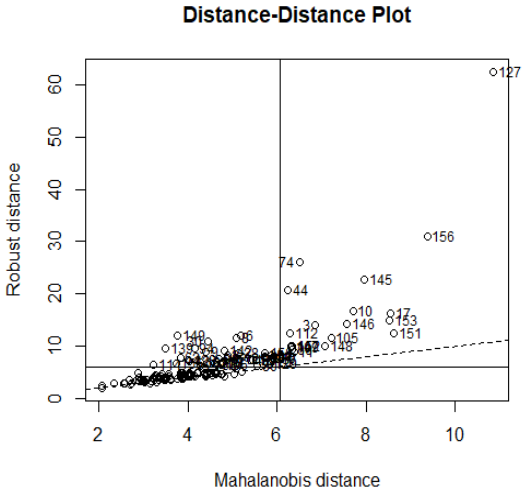


Figure 4 Distance-Distance Plot

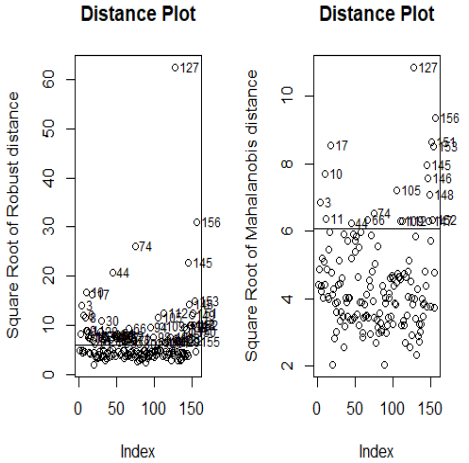


Figure 3 Comparison between MCD and Mahalanobis distances

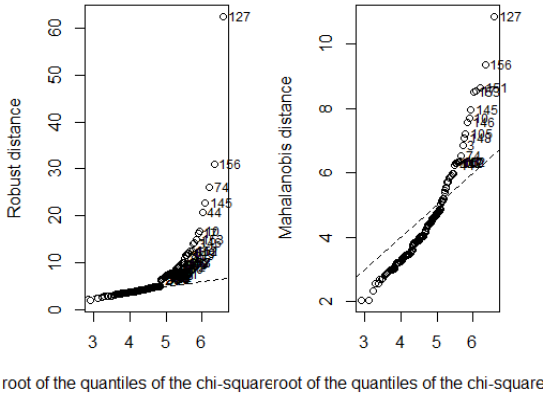


Figure 6 Comparison between the quantiles of the chi-square between the robust and Mahalanobis distances

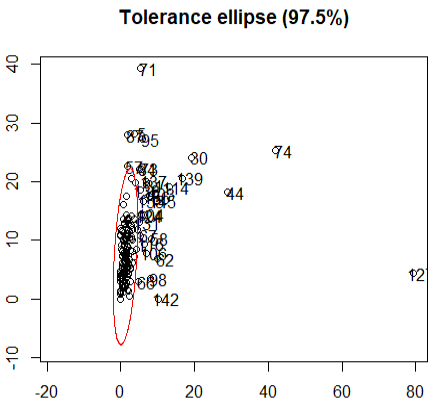


Figure 5 Tolerance ellipse

All the shown representations have the same group of five municipalities being outlined in the robust representation. They are Prainha (Pico) (ID127), Corvo (ID156), Angra (Sé) (Terceira) (ID74), Fajã Grande (Flores) (ID145) and Água Retorta (São Miguel) (ID44). Due to the consistency of these municipalities being represented further away from the others on all representations, it was decided to consider them as passive municipalities, not participating in the principal component and clustering processes.

As for the multivariate analysis, the first decision is to choose the number of components to keep in the Principal Component Analysis. There are several criteriums to select the optimal number of components to keep. The chosen ones rely on a graphical representation of each component's eigenvalue and also the percentage of cumulative variance explained by them.

The graphical analysis suggests that the difference between eigenvalues is reduced from the sixth component onwards, meaning that five might be the optimal number of components to keep. This is reinforced by looking at the actual eigenvalues and choose the ones above 1 (Kaiser's criterium). Being above one means that the explained variance of these new set of variables is superior to the average explained variance of an initial standardized variable. The decision is then to keep the first five components.

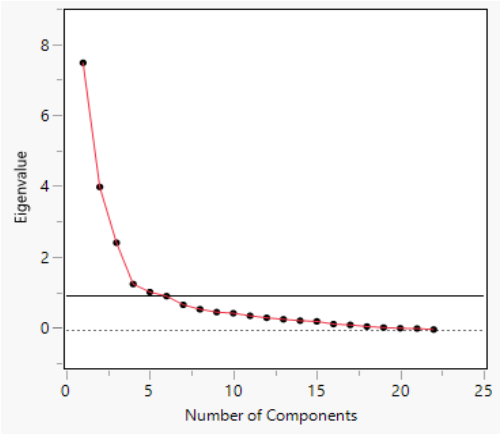


Figure 7 Graphical Analysis Scree Plot

Eigenvalues

Number	Eigenvalue	Percent	Cum Percent
1	7.5502	34.319	34.319
2	4.0462	18.392	52.711
3	2.4733	11.242	63.953
4	1.3067	5.940	69.893
5	1.0809	4.913	74.806
6	0.9669	4.395	79.201
7	0.7207	3.276	82.477
8	0.5952	2.706	85.182
9	0.5156	2.344	87.526
10	0.4832	2.196	89.722
11	0.4108	1.867	91.589
12	0.3552	1.615	93.204

Table 4 Eigenvalues

In the following Table 5, the percentage of variance of the initial variables explained by the retained components is presented. As can be seen, there is an overall high percentage of variance retained by the chosen principal components for most variables, which is what was expected. There is only one exception for the proportion of buildings needing major repairs or degraded; however, it was decided to be kept in the study due to the variable's relevance.

	∑COS2
Prop of buildings not exclusively residential	0.58585
Prop of leased or sub-leased classic family accommodation	0.78403
Prop of own housing with charges	0.67531
Prop of overcrowded accommodation	0.80048
Average age of resident population	0.92612
Average age of buildings	0.83277
Prop of resident population working or studying in another municipality	0.78936
Proportion of car use when traveling	0.81892
Prop of resident population with 15 and more years old whose main livelihood is work	0.83587
Prop of resident population that 5 years previously lived outside the municipality	0.84115
Prop of single-person classic families	0.88155
Prop of classic families with 5+	0.85873
Prop of family nuclei of couples with children	0.89887
Prop of buildings needing major repairs or degraded	0.36237
Average households per accommodation	0.59265
Prop of resident population of foreign nationality	0.49054
Prop of socially most valued professionals	0.79438
Unemployment rate	0.74605
Elderly dependency index	0.89965
Prop of resident population (Who has lived abroad for a continuous period of at least 1 year)	0.65603
Proportion of dwellings with heating	0.75466
Prop. Of population 15+ with no school level completed	0.63193

Table 5 Indicators and their percentage variability explained by the five principal components

In order to deepen the analysis of the variables on the principal planes, the correlation matrices were studied as well. Below is presented the first correlation circle (first and second components)(Figure 8), and the others are present on the annexes. Looking at the representation below, a clear behavior can be seen when looking at the relationship that some variables have with the first two principal components, and some variables seem to be closer to the correlation circle, which means the two first axes quite explain their variance. Table 6 presents the variables and their loadings, which in this case are the correlations between variables and each principal component. For example, looking at the table, one can see that the Average age of resident population has a significantly high positive correlation with the first component but a really low one with the fifth. However, when looking into the Average age of buildings, it has a significantly high positive correlation with the fifth component but a low one with the first. This suggests a further analysis to understand the relationship between each variable and the component it relates the most to comprehend better the behavior that each component translates.

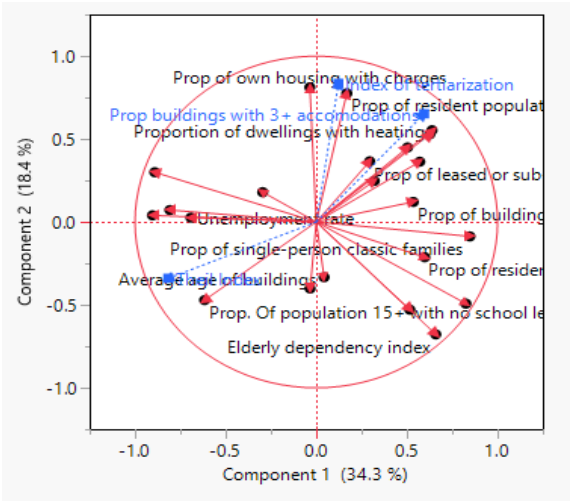


Figure 8 Correlation circle for the first and second principal components

	Prin1	Prin2	Prin3	Prin4	Prin5
Prop of buildings not exclusively residential	0.53483***	0.121	0.44959***	(-0.27424)**	0.088
Prop of leased or sub-leased classic family accommodation	0.29434**	0.36563***	0.63036***	(-0.31545)***	(-0.25858)*
Prop of own housing with charges	-0.036	0.81167***	-0.065	-0.105	0.004
Prop of overcrowded accommodation	(-0.80779)***	0.072	0.37726***	0.003	0.021
Average age of resident population	0.82496***	(-0.49219)***	-0.038	-0.041	-0.015
Average age of buildings	0.043	(-0.33309)***	0.204	(-0.24421)*	0.78649***
Prop of resident population working or studying in another municipality	(-0.29592)**	0.180	0.2923**	0.7528***	0.132
Proportion of car use when traveling	0.50312***	0.44944***	(-0.6012)***	0.008	0.048
Prop of resident population with 15 and more years old whose main livelihood is work	0.168	0.77543***	(-0.43147)***	0.085	0.114
Prop of resident population that 5 years previously lived outside the municipality	0.57051***	0.36321***	0.52369***	0.32975***	0.028
Prop of single-person classic families	0.84674***	-0.087	0.34084***	-0.197	0.047
Prop of classic families with 5+	(-0.90359)***	0.040	0.144	-0.048	-0.133
Prop of family nuclei of couples with children	(-0.89138)***	0.29991**	0.035	-0.037	0.109
Prop of buildings needing major repairs or degraded	-0.034	(-0.40097)***	-0.175	0.021	0.41129***
Average households per accommodation	0.31964***	0.24758*	0.46472***	0.39762***	0.23479*
Prop of resident population of foreign nationality	0.59828***	-0.213	0.106	0.178	(-0.21033)*
Prop of socially most valued professionals	0.62578***	0.52845***	0.31428***	-0.154	0.033
Unemployment rate	(-0.69174)***	0.02668***	0.4893***	-0.158	-0.050
Elderly dependency index	0.65894***	(-0.67773)***	0.037	0.007	-0.069
Prop of resident population (Who has lived abroad for a continuous period of at least 1 year)	0.51566***	(-0.52881)***	0.014	0.27922**	-0.180
Proportion of dwellings with heating	0.63823***	0.55304***	-0.176	-0.097	0.036
Prop. Of population 15+ with no school level completed	(-0.61646)***	(-0.47112)***	0.100	-0.091	-0.108

Table 6 Loadings for each principal component[1]

Considering the need to find the specific relationships between each variable and the components, an orthogonal rotation was applied to variables and consequent components (varimax method) to obtain a "simple structure" (so, a gain of interpretability). By applying such a technique, one can attribute the most adapted label to each component by looking at its associated variables since each variable will have a higher value for the component(s) it relates the most.

[1] From this point onwards: *** p-value < 0.0001 ** p-value < 0.001 * p-value < 0.01

Rotated Factor Loading

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Prop of buildings not exclusively residential	0.233964	0.081561	0.709263	0.017457	0.145236
Prop of leased or sub-leased classic family accommodation	-0.045189	-0.059231	0.846078	0.015296	-0.249810
Prop of own housing with charges	-0.523805	0.524318	0.269059	0.050991	-0.225898
Prop of overcrowded accommodation	-0.680930	-0.556356	-0.064232	0.148268	-0.034409
Average age of resident population	0.912050	0.105936	0.179875	-0.125139	0.187203
Average age of buildings	0.017199	-0.152410	0.127174	-0.002476	0.890544
Prop of resident population working or studying in another municipality	-0.227047	-0.135279	-0.192448	0.820812	-0.093520
Proportion of car use when traveling	0.145475	0.879451	-0.050327	-0.119062	-0.087189
Prop of resident population with 15 and more years old whose main livelihood is work	-0.308740	0.842150	0.001250	0.087778	-0.153718
Prop of resident population that 5 years previously lived outside the municipality	0.242636	0.223447	0.601002	0.600158	-0.104705
Prop of single-person classic families	0.622755	0.158970	0.663760	-0.002747	0.166994
Prop of classic families with 5+	-0.700015	-0.523334	-0.254512	-0.038166	-0.169162
Prop of family nuclei of couples with children	-0.881620	-0.241005	-0.247402	0.037882	-0.029813
Prop of buildings needing major repairs or degraded	0.147128	-0.080194	-0.301501	-0.039444	0.491768
Average households per accommodation	0.098360	0.110218	0.373162	0.650303	0.093234
Prop of resident population of foreign nationality	0.646581	0.058688	0.190885	0.123900	-0.131338
Prop of socially most valued professionals	0.104747	0.452438	0.746007	0.138604	-0.054556
Unemployment rate	-0.591227	-0.615334	0.121599	0.035368	-0.042733
Elderly dependency index	0.910822	-0.142110	0.075821	-0.111112	0.178219
Prop of resident population (Who has lived abroad for a continuous period of at least 1 year)	0.790564	-0.114948	-0.066307	0.110550	-0.034874
Proportion of dwellings with heating	0.141452	0.758114	0.390199	-0.008729	-0.087134
Prop. Of population 15+ with no school level completed	-0.191219	-0.676608	-0.321419	-0.182248	0.032241

Table 7 Rotated Factor Loading

According to the table presented (Table 7), the first component is named Demography since it has higher values for variables age-dependent or family matters. The second component is Socio-Economic, since it has higher values for variables related to work and living conditions, and so it goes.

The following table summarizes the names proposed for each component regarding their rotated factor loadings:

Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Demography	Socio-Economic	Residential Attractiveness	Mobility	Building Condition

Table 8 Principal Component Names

In order to better understand the relationship between individuals, variables, and principal components, a JMP output analysis was done. The idea is to represent the major variables and municipalities in each principal component or ax to see how they behave.

The individual space comprises 151 active municipalities, which is hard to analyze in one factorial representation. To choose the most explicative municipalities to represent on each principal component, their CTR-Partial Contribution for the component's variability- was evaluated. Adding to the above-average CTR criterium is the fact that some municipalities might have a low contribution to the component but be relevant to study since they might have a high percentual COS[2]. This means that some municipalities might have a high percentage of variability explained by solely one component. Thus, the most important municipalities to represent are those with CTR above average or the ones with percentual COS[2] higher than 50%. From all these municipalities, the 10 most to the left and 10 most to the right of each component's representation will be studied. Finally, their contribution to the component's inertia will be calculated (including the municipalities with CTR below average but high percentual squared cosine).

1st Principal Component

The first principal component represents 34.3% of total inertia, being the most significant municipalities: Rabo de Peixe, Fenais da Ajuda, Ponta Garça, Ribeirinha (Ribeira Grande), Angra (São Pedro), Lajes das Flores, Ponta Delgada (São Sebastião) and Horta (Matriz), which are also some of the municipalities with higher percent contribution for the variability of the first principal component or high percentual COS[2].

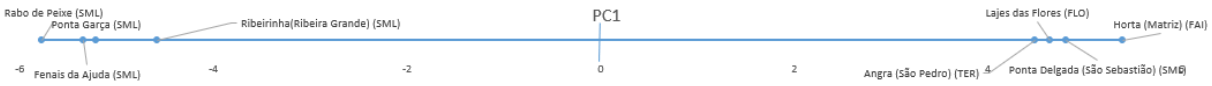


Figure 9 Representation of the first principal component and main municipalities[2]

As one can see on Figure 9, there's an opposition between these groups of regions. As for Horta (Matriz), Ponta Delgada (São Sebastião), Lajes das Flores and Angra (São Pedro), they behave quite positively with the first principal component while Ribeirinha (Ribeira Grande), Ponta Garça, Fenais da Ajuda and Rabo de Peixe behave negatively.

[2] From this point onwards: STM: Santa Maria; SML: São Miguel; TER: Terceira; GRA: Graciosa; SJO: São Jorge; PIC: Pico; FAI: Faial; FLO: Flores; COR: Corvo (Island abrevviations)

The 10 municipalities at the furthest right-hand size (green) and the 10 at the furthest left-hand size (red) are represented on the following table by order of their contribution to the first principal component:

Major Municipalities (by order of contribution)			
Rabo de Peixe (SML)	16.00406	Ponta Delgada (São Sebastião) (SML)	4.961646
Ponta Garça (SML)	5.227323	Horta (Matriz) (FAI)	4.041642
Ribeirinha(Ribeira Grande) (SML)	2.678631	Angra (São Pedro) (TER)	3.779642
Água de Pau (SML)	2.226341	Angra (Nossa Senhora da Conceição)(TER	2.748979
Fenais da Ajuda (SML)	1.748914	Horta (Angústias) (FAI)	2.293196
Feteiras (SML)	1.463359	Madalena (PIC)	2.12294
Água de Alto (SML)	1.169397	Santa Cruz das Flores (FLO)	1.595735
Covoada (SML)	0.859459	Horta (Conceição) (FAI)	1.009513
Sete Cidades (SML)	0.789052	Lajes das Flores (FLO)	0.733159
Santa Bárbara(Ponta Delgada) (SML)	0.640146	Fazenda (FLO)	0.191703

Table 9 Major Municipalities of the First Principal Component

Considering all the municipalities whose CTR is above average (above-average contribution to the inertia) or with a high percentual COS^2, they represent around 88.15% of the first component's inertia.

The highlighted municipalities (orange) are part of the particular case where their component's contribution is below average; however, these municipalities were added to the component's representation due to its high percentual COS^2 value. This means that even though they might contribute less to the inertia of the first component, this component explains the majority of the municipalities' variability retained by the overall components. In Santa Bárbara's case, the overall retained variability is 0.82 (sum of square cosines); however, out of this value, 54% is solely explained by the first component. As for Fazenda, the sum of square cosines is only 0.58, being 56% of this value explained by this component. Surprisingly, these were two municipalities that stood on the left-hand side and right-hand side extremes of this component's representation.

As for the variable analysis, the most significant variables for this component are: Proportion of family nuclei of couples with children, Proportion of classic families with 5 and more members, Proportion of overcrowded accommodation, Elderly dependency index, Proportion of single-person classic families, and Average age of the resident population.

Figure 10 Representation of the first principal component and main variables

The first principal component interpretation can be confirmed from this graphical representation since variables more positively correlated with the component are age-dependent, like the population's average age and the elderly dependency rate. This also means that municipalities with higher coordinates of the first component will most likely have higher values for these variables. What can now be associated is the significant negative correlation between the first component and variables like the proportion of family nuclei of couples with children, larger families, and overcrowded accommodations. From this, one can deduce that municipalities that behave negatively to the first principal component will probably have higher positive values for these variables since they have lower coordinates of this component.

For instance, one of the municipalities contributing more to the first principal component inertia is Rabo de Peixe (São Miguel). This municipality behaves quite negatively with the first principal component, which translates into the lowest average age of the resident population (28.39 years old) and one of the lowest elderly dependency index (7.7). It also has the highest proportion of classic families with 5 or more members (34.95%) and higher values for the other two variables with a negative relationship with the first component.

Below are all the variables with an above-average contribution for the inertia of the first component (contributing for 80.86%):

	Coord(Corr)
Prop of single-person classic families	0.84674***
Average age of resident population	0.82496***
Elderly dependency index	0.65894***
Proportion of dwellings with heating	0.63823***
Prop of socially most valued professionals	0.62578***
Prop of resident population of foreign nationality	0.59828***
Prop. Of population 15+ with no school level completed	(-0.61646)***
Unemployment rate	(-0.69174)***
Prop of overcrowded accommodation	(-0.80779)***
Prop of family nuclei of couples with children	(-0.89138)***
Prop of classic families with 5+	(-0.90359)***

Table 10 Variables with a high contribution for the first principal component

2nd Principal Component

As for the second principal component, it represents around 18.4% of total inertia; the most significant municipalities are represented in Figure 11:

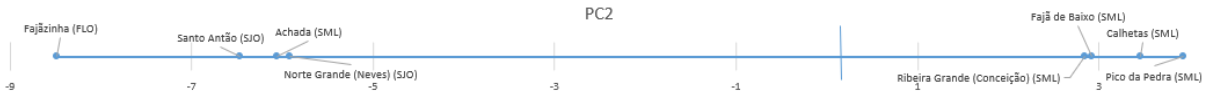


Figure 11 Representation of the second principal component and main municipalities

As can be seen, there's a clear opposition between Ribeira Grande (Conceição), Calhetas, Fajã de Baixo, Pico da Pedra and the municipalities on the left-Norte Grande (Neves), Achada, Santo Antão and Fajãzinha. This means that municipalities on the right-hand side of the representation behave positively with the second component, most likely having higher values for the variables that it represents, and the ones on the left-hand side will have lower values. Below are the 10 municipalities furthest to the left (red) and furthest to the right (green). When considering all the municipalities with CTR higher than average or significant percentual COS^2, they contribute around 85.51% of the component's inertia.

Major Municipalities (by order of contribution)			
Santo Antão (SJO)	3.157876	Ponta Delgada (São Pedro) (SML)	5.182465
Norte Grande (Neves) (SJO)	1.694218	Pico da Pedra (SML)	4.539316
Achada (SML)	1.619769	Fajã de Baixo (SML)	4.358659
Fajãzinha (FLO)	1.471942	Rosto do Cão (Livramento) (SML)	2.394808
Manadas (Santa Bárbara) (SJO)	1.295542	Ribeira Grande (Conceição) (SML)	1.982485
Ponta Delgada(FLO)	1.271001	Terra Chã (TER)	1.725842
Calheta de Nesquim (PIC)	1.104424	São Vicente Ferreira (SML)	1.575764
Faial da Terra (SML)	1.047146	Calhetas (SML)	1.190176
Norte Pequeno (SJO)	0.719242	São Bartolomeu de Regatos (TER)	0.898224
Cedros(Flores) (FLO)	0.438982	Fenais da Luz (SML)	0.896631

Table 11 Major Municipalities of the Second Principal Component

For this component, another municipality with a high percentual of its variability being explained by the second component is added to the study, which is the one that is the furthest to the left of all considered municipalities, Cedros (Flores). This municipality has a sum of square cosines of around 0.3, being 72% of that variability explained by the second component. Due to this percentual variability explained by the component, it was added to the graphical representation.

As for the variable space analysis, all the variables with higher than the average contribution for this component's inertia contribute for around 80.33% of its inertia and are the ones represented in the following figure:

Figure 12 Representation of the second principal component and main variables

As can be seen, the variables with a significant positive correlation with the second principal component are:

- Proportion of own housing with charges (0.81167***)
- Proportion of resident population with 15 and more years old whose main livelihood is work (0.77543***)
- Proportion of dwellings with heating (0.55304***)
- Proportion of socially most valued professionals (0.52845***)
- Proportion of car use when traveling (0.44944***)

This means that municipalities with higher coordinates for this principal component will tend to have higher values for these variables, thus being areas more urbanized. The variables with a significant negative correlation with the second component are :

- Proportion of population with 15 and more with no school level completed(-47112***)
- Average age of resident population (-0.49219***)
- Proportion of resident population who has lived abroad for a continuous period of 1 year (-0.52881***)
- Elderly dependency index (-0.67773***)

This means that municipalities with higher coordinates for this component will probably be characterized by a higher level of urbanization/qualification and a younger population. That being said, this principal component is related to socio-economic matters, thus its previous interpretation.

One example is one of the municipalities that contribute more to the second component inertia, Pico da Pedra (São Miguel). When analyzing this municipality, it is possible to confirm a high Proportion of own housing with charges (74.17%), people whose main livelihood comes from work (58.89%), and houses with heating (53%). It also presents lower values for the variables with a negative correlation with the second principal component.

3rd Principal Component

The third principal component represents around 11.2% of total inertia, being the most important municipalities:

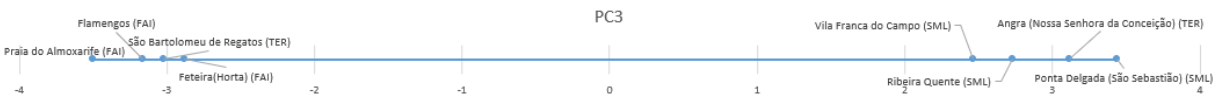


Figure 13 Representation of the third principal component and main municipalities

For the third component, there's a clear opposition between Vila Franca do Campo, Ribeira Quente, Angra (Nossa Senhora da Conceição) and Ponta Delgada (São Sebastião); and

Feteira (Horta), São Bartolomeu de Regatos, Flamengos and Praia do Almoxarife. The municipalities in the extremes of the representation (10 for each side) are represented below:

Major Municipalities (by order of contribution)			
Relva (SML)	3.093313	Ponta Delgada (São Sebastião) (SML)	7.716762
São Bartolomeu de Regatos (TER)	3.006044	Rabo de Peixe (SML)	6.238495
Flamengos (FAI)	2.660681	Angra (Nossa Senhora da Conceição) (TEI	5.977611
Feteira(Angra do Heroismo) (TER)	2.604155	Lagoa (Nossa Senhora do Rosário) (SML)	3.676706
São Vicente Ferreira (SML)	2.56763	Ponta Garça (SML)	3.2345
Praia do Almoxarife (FAI)	1.690292	Água de Pau (SML)	2.912115
Feteira(Horta) (FAI)	1.552311	Vila Franca do Campo (SML)	2.67397
Posto Santo (TER)	1.348374	Horta (Matriz) (FAI)	1.059098
Cinco Ribeiras (TER)	0.906188	Ribeira Quente (SML)	0.943536
Pedro Miguel (FAI)	0.810767	Fenais da Ajuda (SML)	0.864391

Table 12 Major Municipalities of the Third Principal Component

Considering all the municipalities with contributions above average, they represent around 81.05% of its inertia.

As for the variable space analysis, the variables that contribute the most for its inertia are:

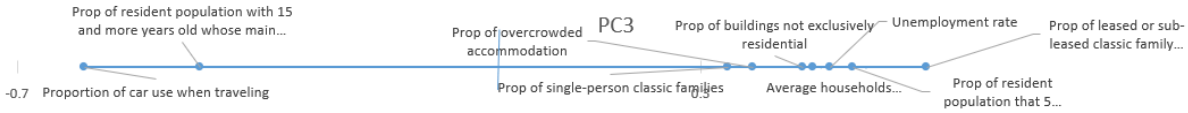


Figure 14 Representation of the third principal component and main variables

Almost all the variables that contribute more to the inertia of the third principal component have a significant positive correlation with it. This means that each municipality that behaves accordingly to this component will most probably have higher values of either:

- Proportion of leased or sub-leased classic family accommodation (0.63036***)
- Proportion of resident population that 5 years previously lived outside the municipality (0.52369***)
- Unemployment rate (0.48930***)
- Average households per accommodations (0.46472***)
- Proportion of buildings not exclusively residential (0.44959***)
- Proportion of overcrowded accommodation (0.377726***)
- Proportion of single-person classic families (0.34084***)

The variables with significant negative correlation with the third component are the Proportion of resident population with 15 or more years old whose main livelihood comes from work (-0.43147***) and the Proportion of car usage (-0.6012***). All these variables contribute to around 86.33% of the inertia of the third principal component.

One example of this is Ponta Delgada (São Sebastião) (São Miguel). It is the municipality that contributes the most to the inertia of the third principal component. While analyzing this municipality, it can be confirmed that it has the highest Proportion of buildings not exclusively residential (19.13%) and high values for the other positively correlated variables.

4th Principal Component

The fourth component represents 5.9% of total inertia being the municipalities represented at the extreme of the component's representation:

Figure 15 Representation of the fourth principal component and main municipalities

For the fourth component, the municipalities Lagoa (Nossa Senhora do Rosário), Lajes das Flores, Pico da Pedra, and Ribeira Chã are expected to have higher values for the variables positively correlated with this component. In contrast, Terra Chã, Fajã de Cima, Fenais da Ajuda and Ponta Delgada (São Sebastião) are expected to have lower values. The other municipalities at the extremes of the graphical representation not presented above are the following:

Major Municipalities (by order of contribution)			
Ponta Delgada (São Sebastião) (SML)	4.714556	Lagoa (Nossa Senhora do Rosário) (SML)	16.66388
Fajã de Cima (SML)	2.31868	Pico da Pedra (SML)	11.27435
Arrifes (SML)	2.185249	Água de Pau (SML)	5.846986
Terra Chã (TER)	1.90384	Calhetas (SML)	2.697578
Vila do Porto (STM)	1.287654	Biscoitos TER)	2.285485
Velas (SJO)	1.212392	Lajes das Flores (FLO)	2.176341
Fenais da Ajuda (SML)	1.119429	Porto Martins (TER)	2.072353
Horta (Matriz) (FAI)	0.902959	Ribeira Chã (SML)	1.76497
Santa Cruz da Graciosa (GRA)	0.890311	Santa Luzia (PIC)	0.989494
Feteiras (SML)	0.72763	Bandeiras (PIC)	0.844394

Table 13 Major Municipalities of the Fourth Principal Component

According to where they are in the right (green) or left (red) of the representation, these municipalities behave the same way as the groups explained above. Adding the other municipalities with a contribution above average, the total amount of inertia explained by them is around 82.35%.

As for the variable space, all the variables with a contribution to inertia above average are represented below, representing 87.69% of the inertia of the fourth principal component:

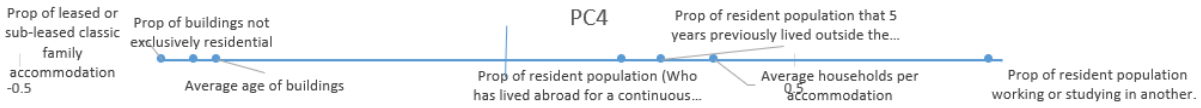


Figure 16 Representation of the fourth principal component and main variables

This component was named as Mobility since, as can be seen, it has a significant positive correlation with the following variables:

- Proportion of resident population working or studying in another municipality (0.7528***)
- Average households per accommodation (0.39762***)
- Proportion of resident population that 5 years previously lived outside the municipality (0.32975***)
- Proportion of resident population who has lived abroad for a continuous period of 1 year (0.27922**)

This means that municipalities with higher coordinates for this component will most likely have higher values for these variables, meaning that their resident population is or was moving, probably residential areas. These municipalities would most likely have lower values for the variables on the left-hand side of the representation - Average age of buildings (-0.24421*), Proportion of buildings not exclusively residential (-0.27424**), and Proportion of leased or sub-leased classic family accommodations (-0.31545***).

One of the municipalities contributing more for the fourth component inertia is Lagoa (Nossa Senhora do Rosário) (São Miguel). Analyzing this municipality confirms a high

proportion of population working or study in another municipality (34.35%), presenting high levels on the other variables as well.

5th Principal Component

Finally, the last principal component represents around 4.9% of total inertia, and the municipalities represented at the extremes of the component's analysis are:

Figure 17 Representation of the fifth principal component and main municipalities

As for the last principal component, Salga, Norte Grande (Neves), Remédios, and Ribeira Chã are expected to have higher values for variables more positively correlated with this component since they have positive coordinates for the component's representation. Castelo Branco, Ribeirinha (Horta), Salão, and Bandeiras, since they are on the left-hand side, with negative coordinates, are expected to have lower values for the variables more negatively correlated with the fifth principal component.

The major municipalities, that is, the 10 most represented to the right (green) and the 10 most represented to the left (red), are presented below by order of contribution:

Major Municipalities (by order of contribution)			
Rabo de Peixe (SML)	10.83873	Remédios (SML)	3.327738
Angra (Nossa Senhora da Conceição) (TER)	4.859738	Fenais da Ajuda (SML)	2.601249
Terra Chã (TER)	3.413151	Vila Nova (TER)	2.346866
Madalena (PIC)	2.433083	Santo Amaro(Velas) (SJO)	1.917532
Castelo Branco (FAI)	1.731577	Santa Bárbara(Ponta Delgada) (SML)	1.90137
Flamengos (FAI)	1.654697	Ribeira das Tainhas (SML)	1.691526
Bandeiras (PIC)	1.171071	Norte Grande (Neves) (SJO)	1.623388
Praia do Almoxarife (FAI)	1.051713	Ribeira Chã (SML)	1.513797
Ribeirinha(Horta) (FAI)	0.711421	Salga (SML)	1.21669
Salão (FLO)	0.678127	Santo Espírito (STM)	1.118938

Table 14 Major Municipalities of the Fifth Principal Component

These municipalities have the same behavior explained before according to the associated color (reflecting the sign of their coordinates). Considering all the municipalities

with an above-average contribution for the inertia of its principal component, they are able to explain 79.97% of this component's inertia.

As for the variable space, only four variables contribute the most for its inertia (84.16% of total inertia), which are:

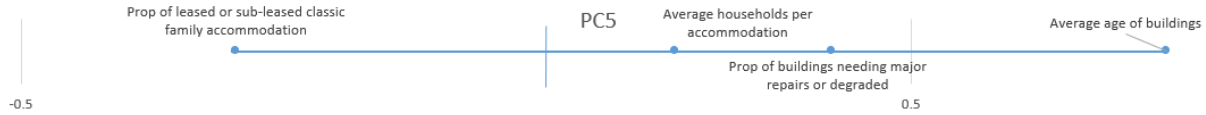


Figure 18 Representation of the fifth principal component and main variables

This principal component was quite correlated with the building condition variables like the Average age of buildings (0.78649***) and the Proportion of buildings needing major repairs or degraded (0.41129***) or the Average households per accommodation (0.23479*). However, it has a significant negative correlation with the Proportion of leased and sub-leased classic family accommodations (-0.25858*), which suggests that the municipalities with higher coordinates for this component are characterized for having older buildings and the predominance of residents with owned houses.

As for the final principal component, one of the major contributors is Ribeira Chã (São Miguel). Analyzing this municipality, it has high values for the Average age of buildings (47.47) and the Proportion of degraded buildings or needing repairs (15.58%). As for the Proportion of leased or sub-leased accommodations, it has a low percentage of 10.94%, confirming the component's behavior.

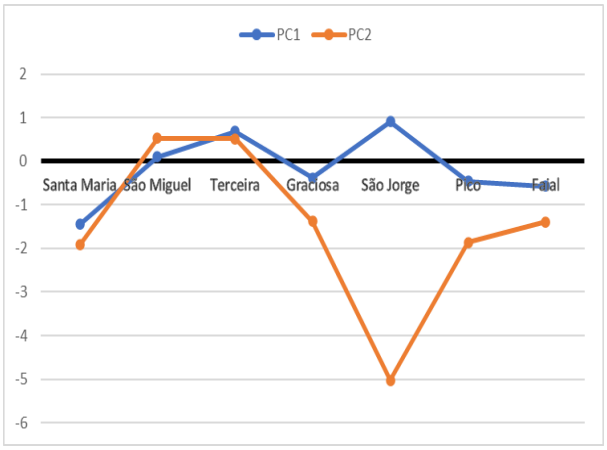
Island Level

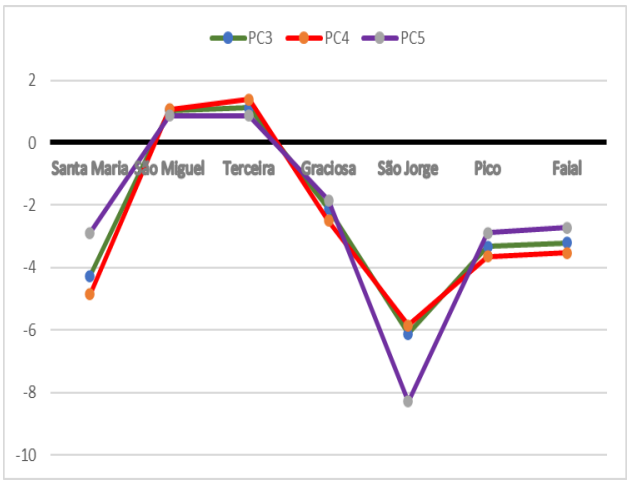
One of the work's propositions was that an island could have sub-regions with different attributes. As such, the exploratory statistical analysis will now be done at the island level so that these divergences can be withdrawn. There are nine islands and five components and their graphical representation is on the annexes (Figure 30).

According to the graphical analysis, for all islands, except Corvo, the distribution of the municipalities in the three principal planes explored suggests that within the same island, even the smaller ones, there are sub-regions that have different characteristics. This goes along with the hypotheses suggested at the beginning of the work, where it was stated that according to

the socio-economic variables selected, it was possible to group municipalities by similarity, thus, finding sub-regions that differ from their main region, suggesting that regional policies should take into account some dissimilarities within the same island.

As for the generalized island distribution, each variable was calculated at the island level, taking into account the proper denominator to calculate its weighted average. Then, each island was centered and projected into the major principal planes. Finally, these projections were introduced into a stacked line chart to understand each island's distribution better (Figure 19).





	PC1	PC2	PC3	PC4	PC5
Santa Maria	-1.44532	-1.92501	-4.27213	-4.85271	-2.89394
São Miguel	0.097461	0.528083	1.037451	1.062259	0.863492
Terceira	0.686091	0.513285	1.111684	1.389396	0.873625
Graciosa	-0.38386	-1.37933	-2.11351	-2.49354	-1.83705
São Jorge	0.90548	-5.02933	-6.11066	-5.85682	-8.28528
Pico	-0.46565	-1.8587	-3.33029	-3.64382	-2.89424
Faial	-0.57073	-1.38894	-3.20867	-3.52134	-2.72087
Flores	11.6632	10.54088	26.82459	31.96393	20.02307

Figure 19 Representation of each island and principal component with the respective coordinates

With a focused graph (without the supplementary island of Corvo and Flores, which has an extreme distribution), it is possible to distinguish the islands' behavior regarding the different principal components retrieved. Some principal components do not show a high variation between islands, as are the first and third components. This means that what differentiates the social-economic outcomes between islands is not so dependent on demography matters, like the average age of population, or residential attractiveness, like housing, employment, and family cradle matters. The distinctions are related to the second and fifth components, as well from the fourth, that is, living conditions, building conditions, and mobility matters. When looking at the second component, two islands stand out due to their negative coordinates, which are Santa Maria and São Jorge. For these islands, it is expected to have lower socio-economic outcomes between its residents, for instance, an overall lower proportion of socially most valued professionals (16.19%), or lower car usage (63.23%) or a lower proportion of residents whose livelihood comes from work (48.31%). As for the fifth component, São Jorge stands out as the island having older and more degraded buildings (average age of buildings of 41.5 years old and a 5.28% of degraded buildings) when compared to the other islands, topped by Santa Maria, which presents a percentage of 7.58% degraded buildings. Looking at the coordinates given by the table in Figure 19, one can see that Flores island has higher coordinates for all components, being the third and fourth components the ones that stand out the most. So, it is expected that the overall island behavior is in accordance with variables significantly correlated with these components, which can be seen by a high proportion of single-person classic families (25.21%) and a high proportion of buildings not exclusively residential (4%) when compared to the other islands; or a high proportion of resident population that 5 years previously lived outside the municipality

(10.41%) or has lived abroad for a continuous period of 1 year (15.58%), for instance. This island also has higher values for the other variables with a significant positive correlation to the principal components, translating into a more mobile island extremely attractive for residential stay.

Supplementary variables

In order to find the relationship of the supplementary variables and the found principal components, they were represented on the principal planes.

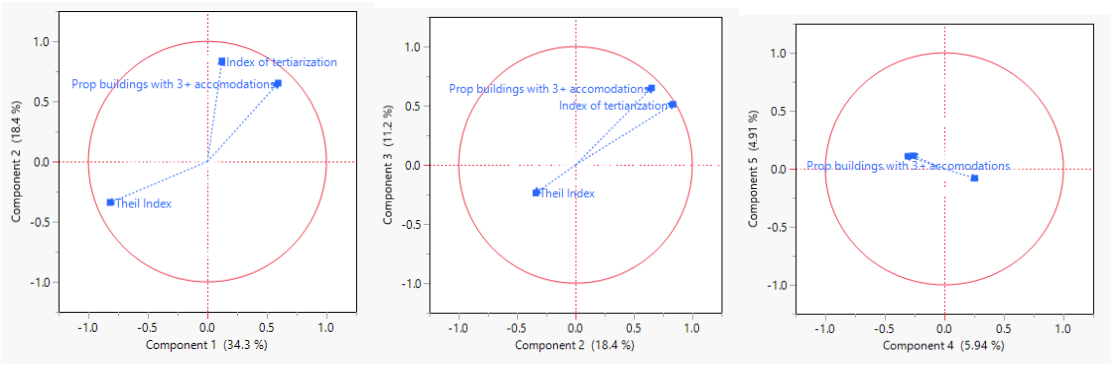


Figure 20 Representation of the supplementary variables on the principal planes

	Prin1	Prin2	Prin3	Prin4	Prin5
Prop buildings with 3+ accomodations	0.59485***	0.65193***	0.65106***	(-0.30632)*	0.10845
Index of tertiarization	0.12190	0.83536***	0.51256***	-0.26197	0.11345
Theil Index	(-0.81625)***	(-0.33976)**	-0.23737	0.25537**	-0.08068

Table 15 Correlation between the principal components and the supplementary variables (loadings)

Prop buildings with 3+ accomodations	
Max	8.605
Median	0.275
Min	0.000
Mean	2.142
Std Dev	0.196

Index of tertiarization	
Max	2248.448
Median	218.041
Min	7.171
Mean	729.057
Std Dev	43.832

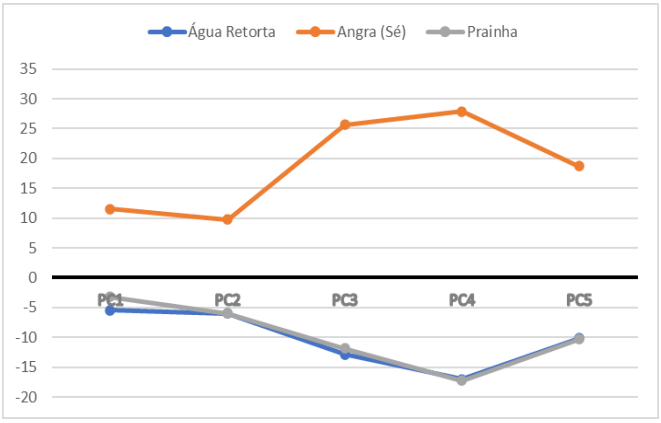
Theil Index	
Max	0.947
Median	0.876
Min	0.747
Mean	0.863
Std Dev	0.004

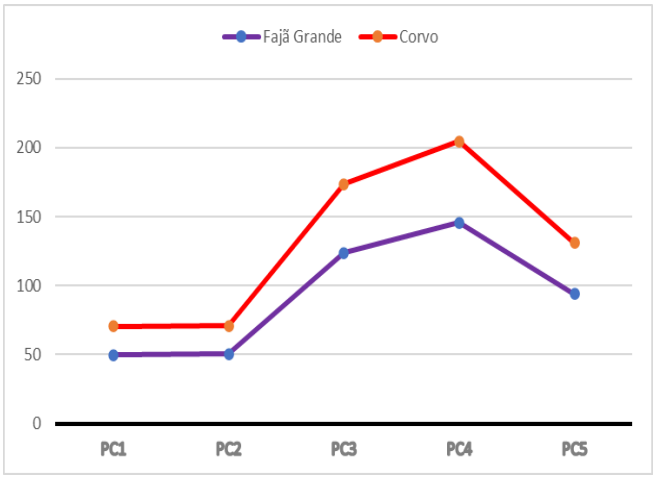
Table 16 Major statistics of the supplementary variables

The supplementary variable Index of tertiarisation measures a region's propensity to have more jobs in the third sector. It is calculated as the weighted average of employment in each sector, being the weight the proportion of the overall population working on that sector of activity. According to the representations, this variable behaves positively with all components except the fourth, being more represented on the second and third components, especially on the second, which shares a highly significant positive correlation with (0.84). This means that municipalities with higher coordinates on these components will most likely have more professionals working in the third sector. It is a variable that varies a lot going from a sector employment propensity of 7 to 2248. The Theil Index calculated measures the social diversity of a region according to its residents' socio-economic groups, and it only has a positive relationship with the fourth component although weak. It is strongly negatively correlated with the first component (-0.82), meaning that municipalities behaving accordingly with this component will most likely have lower values of Theil Index, hence, a lower socio-economic diversification, which translates into a more specialized population in a certain area. However, it is important to keep in mind that the overall Theil Index of Azores is high (>0.70). As for the Proportion of buildings with 3 or more accommodations, it is better represented on the third and predominantly with the second components, meaning that the municipalities that are more explained by this component will most likely have a higher number of buildings with a lot of accommodations. This variable presents a maximum value of 8.6%, which is exceptionally high for the municipality in question (Ponta Delgada (São Pedro)) since it has around 2115 buildings.

Supplementary individuals

The multivariate outliers that were turned into supplementary municipalities should now be represented by their projection into the principal planes in order to check their relationship with the retrieved principal components. The same procedure done at the island level was done to these individuals by representing them in a stacked line chart to better visualize the differences between individuals and principal components.





	PC1	PC2	PC3	PC4	PC5
Água Retorta	-5.45864	-5.96762	-12.8571	-17.0264	-10.1408
Angra (Sé)	11.5119	9.74062	25.6513	27.8758	18.7252
Prainha	-3.20535	-6.00231	-11.8438	-17.2276	-10.3166
Fajã Grande	49.9127	50.3522	123.71	145.54	93.9468
Corvo	70.4534	70.7447	173.515	204.51	130.899

Figure 21 Supplementary individuals and coordinates for the principal planes

Looking at the representation and the coordinates for other planes, one can see that Prainha (Pico) and Água Retorta (São Miguel) have a closer similarity when compared to the other individuals. They have negative coordinates for all principal components, suggesting that it is expected to have lower values for variables with significant positive correlation with some components and higher ones for the variables with a significant negative correlation with other components. There is another similarity between Fajã Grande (Flores) and Corvo when looking at the other municipalities since they exhibit considerably higher coordinates for all components, highlighting Corvo. This leaves Angra (Sé) (Terceira), which also has positive

coordinates for all components, even though they are not as high as the ones for Fajã Grande (Flores) and Corvo.

When comparing the values of these municipalities to the mean of each variable, some distinctive behavior can be withdrawn, for instance, for the Proportion of buildings not exclusively residential, Água Retorta (São Miguel), Angra (Sé) (Terceira), and Prainha (Pico) have a considerably higher value when comparing to the mean or maximum value for this variable of the rest of the data. When only considering the data for the active municipalities, the mean proportion is 3.15%, being the maximum of 19.13%. For these supplementary municipalities, it is 29.01% for Água Retorta (São Miguel), 41.93% for Angra (Sé) (Terceira), and 79.31% for Prainha (Pico), which are percentages outstandingly higher than the "normal" for the Azores region. Even though they present this outlier behavior for this variable, which has a significant positive correlation with the third principal component, this does not mean that their representation will be positive for this component. For instance, Prainha (Pico) and Água Retorta (São Miguel) have a negative coordinate for this principal component due to their below-average behavior with some of the other variables with a significant positive correlation with this component. An example of this is the Proportion of leased or sub-leased accommodations for Prainha (Pico), where the proportion is 4.33% for a mean of 12.76%, amongst other variables. This happens because the contribution for the component's inertia is higher for these other variables when compared to the proportion of buildings not exclusively residential.

Along with this specific variable's behavior, other distinct values can be found. Starting with Corvo, it has a considerably higher Proportion of resident population that 5 years previously lived in another municipality, which is 21.16% compared to the 6.50% mean and 13.97% maximum for the other data; as well as a high Proportion of single-person classic families, which is 41.40% compared to a mean of 16.32% and maximum value of 33.33%. Additionally, it also has the highest Average of households per accommodation, 1.21, compared to the 1.01 mean and 1.08 maximum. This is seen in the presented graphs by looking at Corvo's high coordinates for the third and fourth principal components. Another supplementary individual with a higher average is Fajã Grande (Flores), with 1.15. As for Angra (Sé) (Terceira), it presents another distinctive outlier behavior for the variable Proportion of socially most valued professionals, which is 41.67% when compared to a mean of 17.87% and

maximum value of 39.75% for the other data. This can also be seen by the high coordinates of this municipality for the third principal component.

Considering these examples and the rest of the variables, these municipalities are proven to have a distinguishing behavior that might jeopardize the study if added as active in the principal component analysis.

Cluster Analysis

The ascendant hierarchical aggregation method suggested a 16-cluster division firstly (Cubic Clustering Criterium). Evidently, this is not an optimal solution for this case since 16 is too high a division for 151 municipalities, not making it interpretable. As such, looking at the dendrogram, a possible solution could be of 5, 6, or 7 clusters since they tackle a reasonable amount of distance difference (presented below the dendrogram).

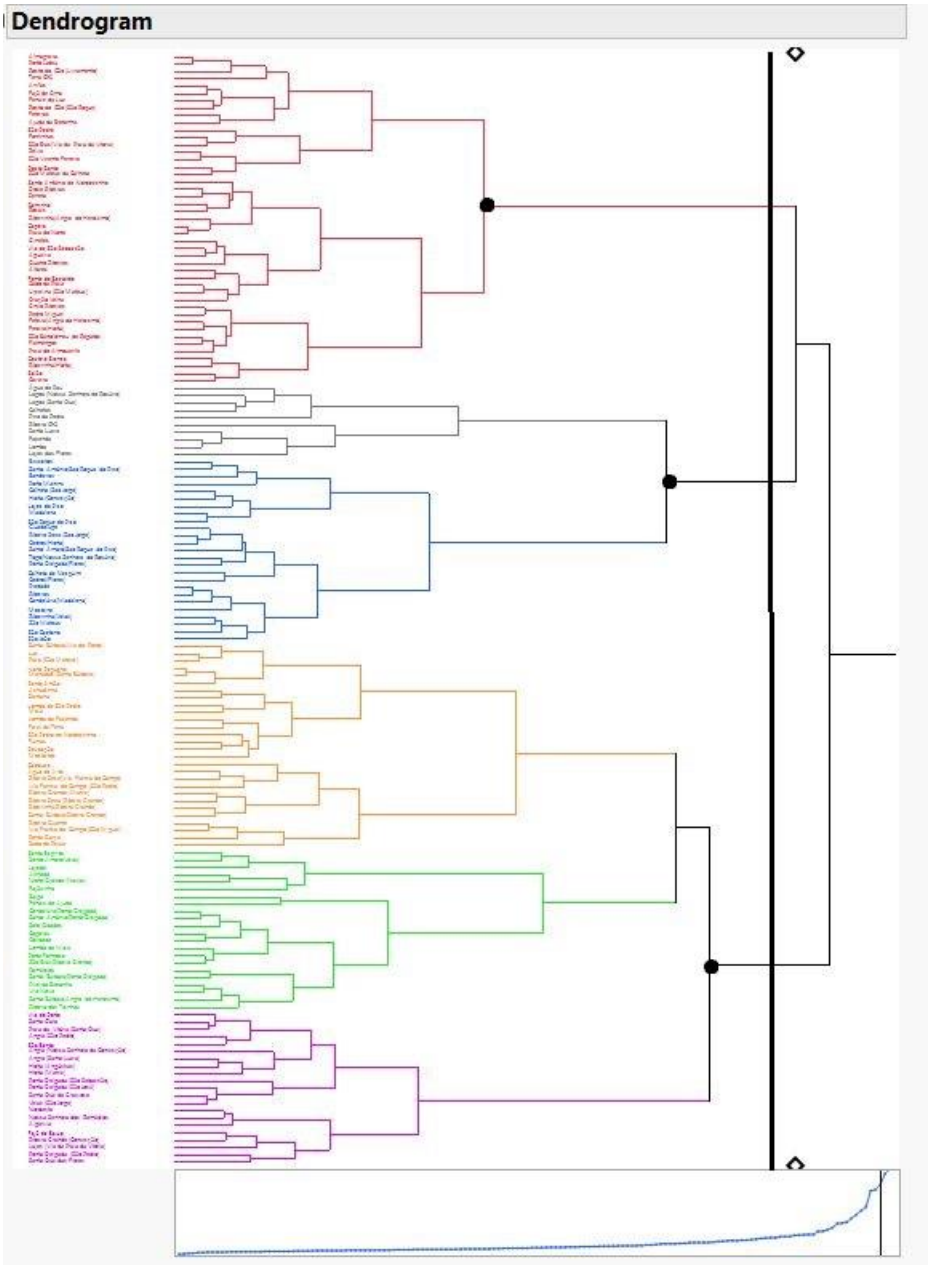


Figure 22 Ascendant Hierarchical Aggregation Dendrogram

Looking at the dendrogram (Figure 22), one can see three major groupings being done from right to left. Looking closely, clear divisions are being made according to their variable behavior. For instance, for the first one (counting from above), it groups municipalities with young residents with smaller families though with a high proportion of couples with children whose primary source of income comes from work and live on newer buildings. As for the second group, it assembles older residents with low mobility inter-municipality, with smaller families of single-person character and lower unemployment. Finally, the third group joins municipalities with younger residents with families with many members, where the unemployment is higher while the residents work on lesser valued professions. Therefore, this major division already accounts for a lot of distinguishing characteristics that motivate a more precise municipality grouping to tackle the differences and similarities between the sub-regions considered in the study.

As such, the constellation plot, the standard deviation table per cluster aggregation, and the k-means optimal solution (presented in Annexes: Figure 31 and Tables 22 and 23) were studied additionally in order to find a sub-optimal stable solution for the number of clusters. The criterium used to aid was a mapping of the distribution of each municipality using the principal component quantiles to measure their distribution. After that mapping was done, each cluster was analyzed by principal component behavior (below, in, or higher than the interquartile range) and classified according to the component's name. According to the mapping, the division that made more sense and had clusters with clearer differentiation was the 6-cluster grouping, as well as it is the number of clusters that provides a big gap in the difference of distances between clusters (graphical representation of the dendrogram).

The final classification (at the ascendant hierarchical aggregation level) is the following:

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
	Demography	Socio-Economic	Residential Attractiveness	Mobility	Building Condition
Cluster 1	High	Moderate to high	High	Low	Mixed
Cluster 2	Mixed	Moderate to low	Moderate	Mixed	High
Cluster 3	Low	Moderate	High	Moderate	Mixed
Cluster 4	Moderate to high	Moderate to low	Moderate	Moderate to High	Moderate to low
Cluster 5	Moderate	High	Low	Moderate	Moderate
Cluster 6	Moderate	High	High	High	High

Table 17 Cluster preliminary classification by principal components[3]

As can be seen, each cluster brings additional information about a group of municipalities, considering each dimension. For instance, clusters 1 and 3 have the same behavior regarding the second, third, and fifth components when considering the variables that correlate the most with them; however, the first and fourth components bring a differentiation between the two groups. This means that municipalities belonging to each group will be more easily distinguished by their levels of demography attributes, like the average age of population or the family cradle size or mobility attributes like the intra-municipality commutes or movings. This suggests that municipalities can be grouped in a way that translates their socio-economic characteristics into a generalized territorial portrait that differentiates all groups.

Since it appears that 6 is the sub-optimal number of clusters, K was set as 6 on the K-means clustering analysis. To name these clusters into territorial geographical classifications, a deeper study was made at the cluster level. As such, the major statistic summaries were calculated for each principal component value on each cluster and the overall behavior to characterize each cluster according to their relationship with the variables was studied.

[3] Demography describes the average age of population and family cradle matters (low value: young population, bigger families; high value: older population and smaller families)
Socio-Economic describes the major indicators of living conditions (low value: lower living conditions or lower urbanized lifestyle; high value: higher living conditions)
Residential Attractiveness summarizes the aspects that describe a residential area (low value: low attractiveness; high value: high attractiveness)
Mobility translates the intra-municipality mobility for working or studying or starting to live in another municipality (low value: most things happen solely in the same municipality; high value: high mobility intra-municipality either for moving or commuting)
Building condition means an old building or needing repairs (low value: good condition; high value: bad condition)

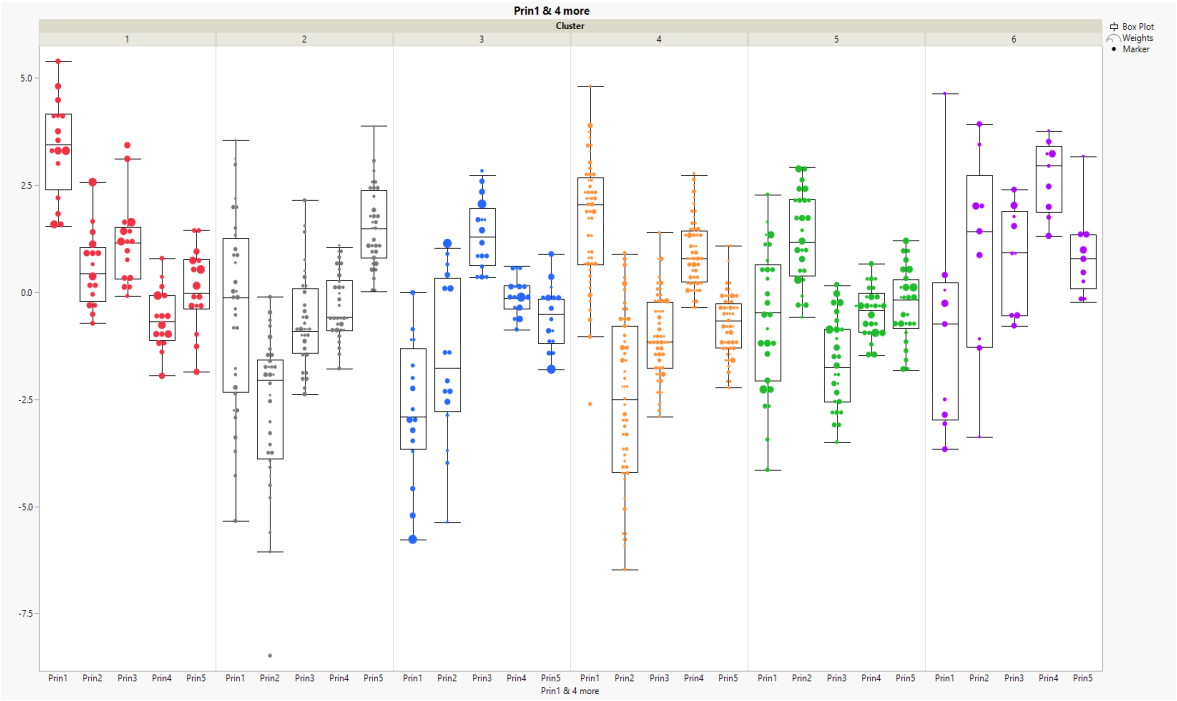


Figure 23 Representation of each Principal Component distribution by Cluster

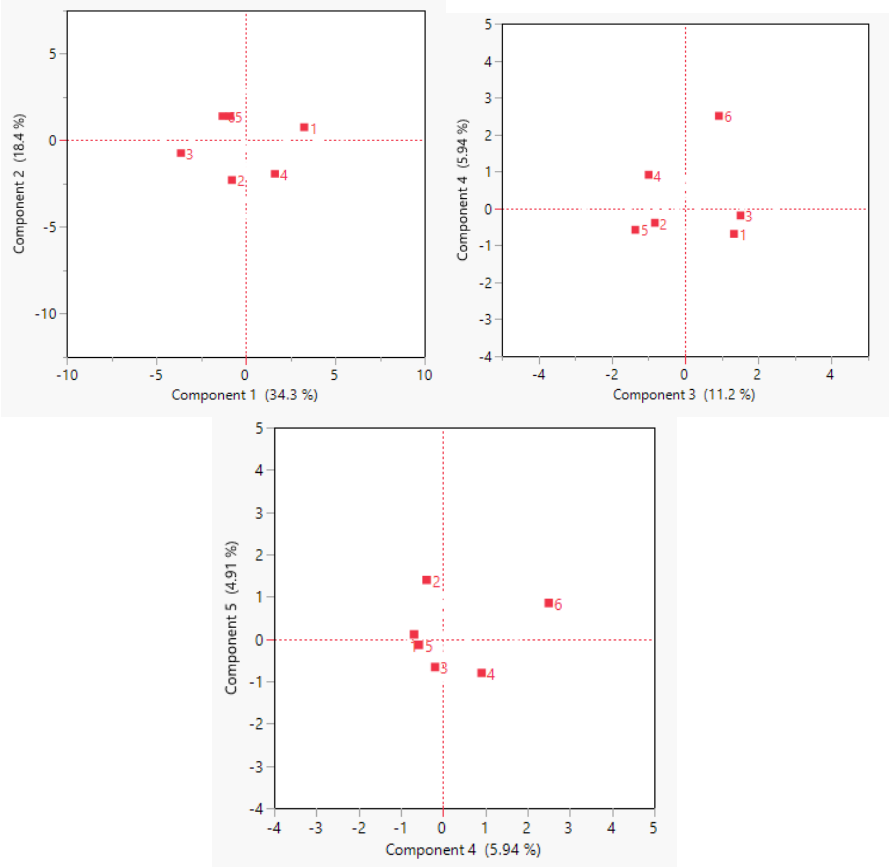


Figure 24 Representation of the cluster on the principal planes

	Number of Municipalities per Island									
	Santa Maria	São Miguel	Terceira	Graciosa	São Jorge	Pico	Faial	Flores	Total Municipalities	Percentage of resident population
Cluster 1	1	4	5	1	1	0	3	1	16	22.08%
Cluster 2	2	19	5	2	2	0	0	2	32	11.36%
Cluster 3	0	16	0	0	0	0	0	0	16	15.04%
Cluster 4	1	4	11	1	8	16	4	6	51	15.78%
Cluster 5	1	13	7	0	0	0	6	0	27	26.44%
Cluster 6	0	7	1	0	0	0	0	1	9	9.30%

Table 18 Cluster Summary

	PC1	PC2	PC3	PC4	PC5
Cluster 1	16	3	6	8	8
Cluster 2	7	15	5	1	12
Cluster 3	10	9	6	0	5
Cluster 4	6	22	4	8	8
Cluster 5	7	10	17	6	8
Cluster 6	4	5	3	9	5

Table 19 Significant municipalities for each principal component

Considering all the outputs presented, including an analysis of the distribution of each cluster by variable, all the relationships between each cluster and variable were retrieved, including what distinguishes each cluster. As can be seen in the previous Tables 18 and 19, having more municipalities in a cluster does not mean it agglomerates a higher percentage of the resident population, like cluster 5 that has lesser municipalities than cluster 4 (27<51), but it includes 26.44% of the population against the 15.78% of the fourth cluster. Another important aspect is the number of significant municipalities for each principal component belonging to each cluster. In Table 19, it is reunited the municipalities with CTR higher than average or a significant percentage of COS^2 being explained by each component and allocated to its respective cluster. For instance, for the first and fourth clusters, there is a high number of municipalities being representative for a certain component, first for cluster 1 and second for cluster 4. This means that these clusters reflect positively the characterization done for the principal components since they englobe many municipalities that contribute significantly to them, respectively.

In the Annexes, Table 24 shows the municipalities belonging to each cluster.

The summary of each cluster will be presented next.

Cluster 1 Urban Working Residential: The municipalities on this cluster are characterized by a population mainly of active age, whose main income source comes from work. Residents have an urbanized lifestyle with high car usage when traveling and high number of households per accommodation. Additionally, the elderly dependency rate is not so high, and a considerable proportion of the resident population has foreign nationality. The cluster stands out due to the high proportion of buildings not exclusively residential accompanied by leased and sub-leased accommodations. The building conditions are good since they have a high proportion of dwellings with heating and the proportion of buildings needing repairs or degraded is relatively low; however, it lodges many accommodations per building. The population is characterized by a high percentage of single-person classic families and/or working in socially most valued professions, mainly in the third sector, since the Index of tertiarization is also high.

Furthermore, this population is qualified, having a low percentage of population with no school level completed. This all translates into a lower Theil Index (in fact, the lowest), suggesting a lower social diversification, that is, lower socio-economic contrast. A considerable proportion of residents from this population are residents that 5 years previously lived outside the municipality; however, the family cradle is translated into a smaller classic family size, reflecting less overcrowded accommodations.

Looking at the graphical representation of all clusters (Figure 24), one can see that cluster 1 has the higher coordinates regarding the first principal component, having negative coordinates only for the fourth component. This means that this cluster will tend to have higher values for the variables that the first component has a significant positive correlation with since it also has many municipalities being significant for this component's representation (table 19). This can be illustrated by looking into some examples of municipalities belonging to this cluster. This is the case of Ponta Delgada (São Sebastião) (São Miguel), which has the highest proportion of buildings not exclusively residential (19.13%) and socially most valued professionals (39.75%), being three other municipalities from this cluster an outlier on this variable; Angra (Nossa Senhora da Conceição) (Terceira), which has the highest proportion of leased and sub-leased accommodations (39.35%); Ponta Delgada (São Pedro) (São Miguel) has the highest proportion of buildings with 3 or more accommodations (8.6%) and Index of tertiarization (2248.45). In contrast, it includes municipalities with the lowest Theil Index, as is the case of Ponta Delgada (São José) (São Miguel) (0.747).

Cluster 2 **Unqualified Sub-Urban Residential:** This cluster has an active to old resident population with higher elderly dependency index. What characterizes this cluster compared to the others is a higher proportion of overcrowded accommodations, though the accommodations tend to have lesser households per accommodation or buildings with lesser accommodations. This might be due to a family cradle characterized by a higher proportion of couples with children. Adding to this is the higher building age and proportion of buildings needing major repairs or degraded, suggesting overcrowded older buildings, in this case, a mix between own housing with charges and leased accommodations, being the proportion of leased accommodations relatively low. The resident population is characterized by being lesser qualified, having a higher proportion of residents with no school level completed, working mainly on the first and second sectors (Low Index of Tertiarization), which in this case are socially less valued professions. This translates into a high Theil of Index, suggesting that this cluster is the one that groups municipalities with higher social-economic contrasts between its residents. This cluster includes municipalities with low mobility inter-municipality and lower urbanized lifestyles.

Looking at this cluster's representation of the principal planes (Figure 24), one can see that it only behaves positively with the fifth principal component. This suggests that the municipalities in this cluster will most likely have higher values for variables with a significant positive correlation with that component since, as shown in Table 19, this cluster has the highest number of municipalities being significant for the fifth component's representation. This can be illustrated by a few examples of municipalities that reflect this behavior, as is the case of Fenais da Ajuda (São Miguel), which has the highest proportion of overcrowded accommodations (38.87%); Lajedo (Flores), which has the oldest buildings (average of 78.29 years old); Fajãzinha (Flores), which has the highest proportion of single-person classic families (33.33%); Santo Amaro (Velas) (Pico), which has the highest proportion of buildings needing repairs or degraded; Nossa Senhora dos Remédios (São Miguel), with the highest proportion of resident population that lived abroad for a continuous period of at least 1 year (28.78%); and Algarvia (São Miguel), which has the highest proportion of dwellings with heating (98.11%). As for the ones it behaves negatively like the second principal component, which is also relevant to study since this cluster gathers a high number of municipalities significant for this component's representation, an example is the lowest proportion of residents whose main

livelihood comes from work, in Fajãzinha (Flores) (28.57%) and also includes the lowest proportion of own houses with charges in Lajedo (Flores) (12.12%).

Cluster 3 Young Unqualified: This cluster is characterized by a less qualified younger population, which translates into a high Theil Index (high social diversification) and a considerably higher unemployment rate. The classic families have many members, including family nuclei with children, and do not rely as much on the car when traveling. This also reflects on a lower proportion of single-person classic families. Their accommodations are mainly owned by residents and overcrowded, though in good conditions, having a lower proportion of buildings needing repairs or degraded, it has a lower proportion of dwellings with heating and lesser accommodations per building. The resident population tends to work on lesser valued professions, mostly on first and second sectors (low Index of Tertiarization), working or studying in another municipality, and whose main source of income comes from other activities rather than from work.

The third cluster only has positive coordinates with the third principal component, while it has a high negative correlation with the first component (Figure 24). Complementing with Table 19, one can see that this cluster has a high number of municipalities being significant for the first principal component's representation, which suggests that it is expected for the municipalities on this cluster to have lower values for the variables positively correlated with this component. For instance, Ribeira Seca (Vila Franca do Campo) (São Miguel), which has the highest proportion of own houses with charges (77.52%); Rabo de Peixe (São Miguel), which has the highest proportion of classic families with 5 and more members (34.95%); Ponta Garça (São Miguel) has one of the highest proportions of overcrowded accommodations (34.11%), and proportion of population with no school level completed in Faial da Terra (São Miguel) (25.26%). As for the ones it behaves negatively with, it includes municipalities with the lowest usage of car when traveling, like Ribeira Quente (São Miguel) (17.11%); Rabo de Peixe (São Miguel), which has the lowest average age of resident population (28.39 years old); and Santana (São Miguel), which has the lowest proportion of dwellings with heating (6.67%).

Cluster 4 Aged Middle Class: The resident population on this cluster is relatively old, also having a high elderly dependency index. It has a relatively high proportion of foreign nationality residents as well residents who lived abroad for at least 1 year. It is also characterized by a higher proportion of single-person classic families and families with fewer members, whose main livelihood comes from other activities rather than work with low mobility inter-municipality. This includes the lowest proportions of family nuclei of couples with children, though it is higher in some municipalities. Buildings are partly degraded with a lower overcrowded accommodation and lower leased or sub-leased accommodations, and fewer accommodations per building. It also has a low Index of Tertiarization, meaning that the population works mainly on the first and second economic sectors, translating into a lower proportion of residents working in socially valued professions and a lower unemployment rate.

The fourth cluster has positive coordinates for the first and fourth principal components; however, it has a high number of municipalities being significant for the second component's representation (Table 19), meaning that it is expected that municipalities belonging to the cluster to have lower values for variables more positively correlated with this component since it has negative coordinates in its representation. Examples of this are Santo Amaro (São Roque do Pico) which has the highest average age of population (50.32 years old); Cedros (Flores), with a high average age of buildings (77.2 years old); and resident population with foreign nationality (12.5%); the highest elderly dependency index in Calheta do Nesquim (Pico) (47.3); Norte Pequeno (São Jorge), which has the highest proportion of residents with no school level completed (28.21%) and finally the highest Theil Index of 0.947 in Santa Bárbara (Vila do Porto) (Santa Maria). As for the lowest values, it includes municipalities with the lowest proportion of couples with children like Mosteiro (Flores) (33.33%). This municipality has the lowest proportion of overcrowded accommodations, with a percentage of 0. This cluster also includes the municipality with the lowest average age of buildings in Ribeirinha (Horta) (Faial) (12.47 years old).

Cluster 5 Young Qualified Middle Class: This cluster groups young residents whose lifestyle is more urbanized, since their main livelihood comes from work, the usage of car is higher, they work on not so valued professions, mainly in the second and third sectors (medium Tertiarization Index), the proportion of residents with no school level completed is low, and there is a higher proportion of family nuclei with children. These family nuclei have many members living in younger buildings, more exclusively residential with lesser accommodations, with heating and not degraded, though not so overcrowded, mainly own houses with charges, having lower proportions of single-person classic families. The unemployment rate is medium to high while mobility inter-municipality is low. There is also a considerably lower proportion of residents who lived abroad for at least 1 year, and the elderly dependency index is also lower.

The fifth cluster has a positive relationship with only the second component, having high negative coordinates for the third component, meaning that the municipalities on this cluster should have higher values for variables with high positive correlation with the second component and lower ones for variables more positively correlated to the other components, like the third, which has a high number of municipalities being significant for this component's representation (Table 19). Some examples are São Bartolomeu dos Regatos (Terceira), which has one of the highest proportion of own houses with charges (71.1%); Feteira (Horta) (Faial), which has a car use of 55.56% when traveling; and Feteira (Angra do Heroísmo) (Terceira), which has one of the highest proportions of resident population whose main livelihood is work (58.92%). It also includes the municipality with the highest car use when traveling, which is in Praia do Almoxarife (Faial) (87.68%). This municipality also has the lowest proportion of resident population with no school level completed (2.45%). As for the variables expected to be lower, there are the examples of municipalities like Praia do Norte (Faial), which has a percentage of 0 leased or sub-leased accommodations; Relva (São Miguel), which has one of the lowest proportions of single-person classic families (7.61%); and Fenais da Luz (São Miguel), which has one of the lowest average age of resident population (32.74 years old).

Cluster 6 Attractive Residential: This cluster groups the municipalities with young residents with high levels of mobility. This means that this cluster has a higher proportion of residents working or studying in another municipality or that 5 years previously lived outside the current municipality. Their main livelihood is work, and the unemployment rate is relatively medium to high, even though there is a low proportion of residents with no school completed. They are also characterized by a high average household per accommodation due to a high proportion of family nuclei with children and families with more members, contrasting with a low proportion of single-person classic families. As for the building conditions, it has a relatively higher proportion of overcrowded accommodation on owned houses, more exclusively residential, while the proportion of dwellings with heating is lower and buildings less degraded. Taking all this into account, the Theil Index is high, which means that there is a higher social-economic contrast between residents.

The sixth cluster has negative coordinates for the first principal components and high positive ones for the fourth. This means that it is expected to have higher values for variables with a significant positive correlation with the fourth principal component while lower ones for the ones positively correlated to the first since it has a high number of municipalities being significant for these components' representations (Table 19). Examples of this are Pico da Pedra (São Miguel), which has the highest proportion of resident population working or studying in another municipality (58.32%); Calhetas (São Miguel), which has the highest proportion of population that 5 years previously lived in another municipality (13.97%); and Lajes das Flores, which has the highest average of households per accommodations (1.08 households). As for the lowest ones, Calhetas (São Miguel), which has one of the lowest average age of resident population (30.57 years old) and lowest elderly dependency index (7.5%); Pico da Pedra (São Miguel), which has one of the lowest proportions of single-person classic families (9.51%); and Ribeira Chã (São Miguel) which has a virtual percentage of 0 buildings not exclusively residential.

The distribution per island is presented next (1-Santa Maria; 2-São Miguel; 3-Terceira; 4-Graciosa; 5-São Jorge; 6-Pico; 7-Faial; 8-Flores and 9-Corvo).

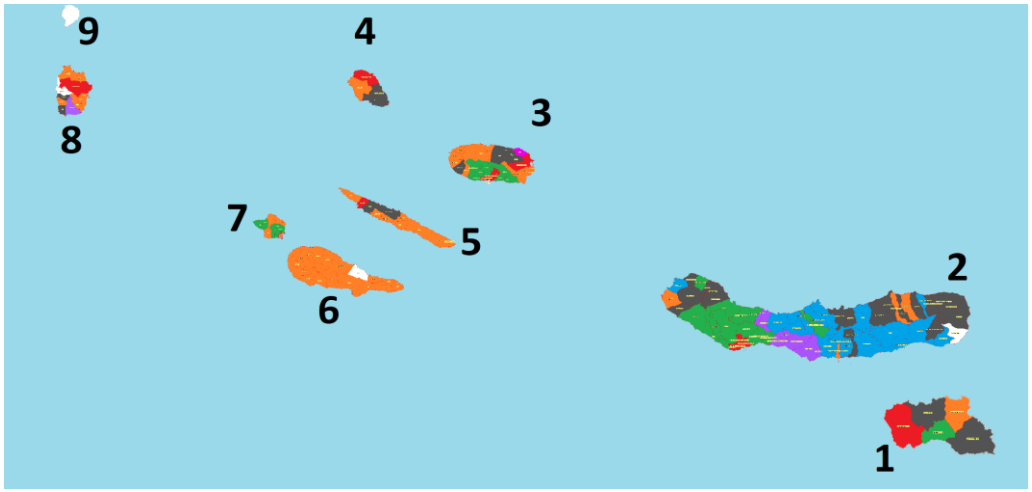


Figure 25 Azores cluster distribution

Even though some islands are more homogenous, there are still sub-regions that differ from those surrounding them, confirming the early suggestions. In order to quantify the heterogeneousness of each island, a differentiation coefficient was calculated. The study of these coefficients might give some insight into which island presents a higher socio-economic divergence.

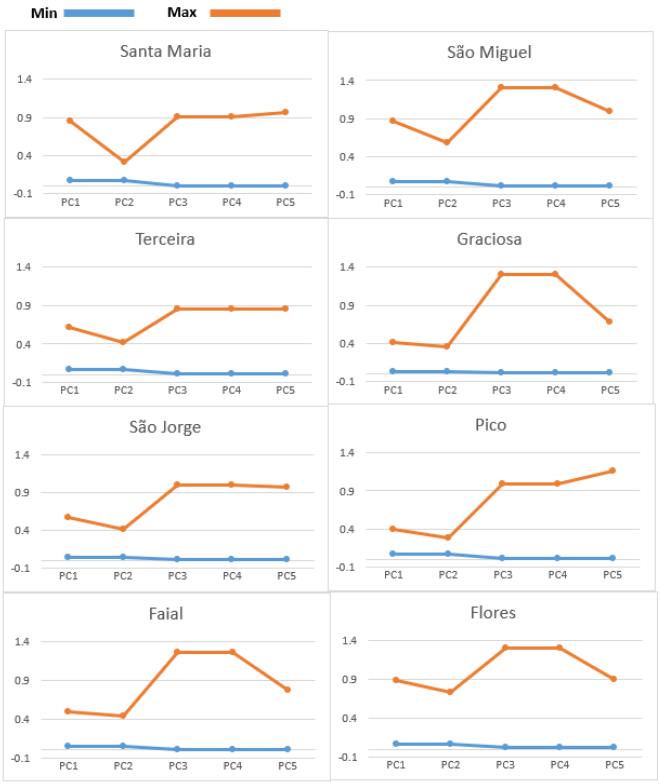


Figure 26 Coefficient of dispersion min-max representation for each island

As can be seen, the principal components that present a higher differentiation for their significant variables are the third, fourth and fifth. The islands that present a higher contrast for the third and fourth components are São Miguel, Graciosa, Faial, and Flores. As for the fifth, the island that stands out for their contrast between municipalities is Pico island. As for the first and second components, the contrast is significantly lower, especially for the second principal component. Taking this into account, these islands present a higher contrast between the residential attractiveness, mobility, and building condition dimensions. The islands that show a higher homogeneity are São Jorge, Terceira, and Santa Maria. This shows that even though some municipalities from the same island might behave similarly, some sub-regions do present divergences at the dimensions considered.

An interesting observation of this outcome is that when the island projection was considered, the third component did not appear as this relevant while the second did; however, at the island level, the third component is one of the components that contribute to intra-municipality dispersion while the second component does not. This means that when comparing islands, differences do not appear at the residential attractiveness level and rather at the living conditions or socio-economic matters more related to the resident's lifestyle. However, when comparing municipalities of the same island, sub-regional divergences appear at the residential attractiveness level and not solely on living conditions. In both cases, building conditions and mobility interfere with the sub-regional differentiation at island and arquipelago levels.

7. DISCUSSIONS

Two important aspects arise when reading the literature about sub-regional territorial disparities and looking at this study's results, namely, what role the demographic unbalance between sub-regions and the socio-economic disparities shown have when characterizing this territory. Typically, economic centers agglomerate a higher proportion of population in a certain region, as is the case of the main cities in the Azores like Ponta Delgada (São Miguel), Angra do Heroísmo (Terceira), and Horta (Faial). Therefore, it is expected to have in these areas a higher amount of population qualified and working in more developed sectors, as well as being the regional areas that generate more employment. However, when looking at the cluster groups formed, especially cluster 1, which includes the major municipalities of Ponta Delgada, Angra, and Horta, it also includes municipalities from other islands that one would not initially associate with them, as is the case of Vila do Porto (Santa Maria) or Santa Cruz das Flores. This suggests that more important than being an overall economic center, some municipalities act as a development pole for their surrounding areas.

Additionally, as the years go by, the tendency for the average age of a region to increase leads to some concerns regarding the demographic influence upon the aged sub-regions. According to this work's outcomes, the disparities between islands and inside the same island are not so dependent on age-dependent variables. Logically, these results were obtained using data from 2011. What would be interesting to see is if the age-dependent indicators would increase their importance with the census data of 2021. With the overall population's aging, it is natural that some socio-economic outcomes might be jeopardized by the inherent needs of an older population.

Moreover, when comparing the leading indicators at the island level, the differences between municipalities from the same island appear mainly at the residential attractiveness, mobility, and building condition. In contrast, between islands, the dispersion appears to be related to socio-economic matters, mobility, and building condition. This means that when considering age-dependent variables and socioeconomic status indicators, sub-regions close to each other tend to behave similarly. However, at the same time, their residential attractiveness and mobility or building conditions vary. This can explain why some municipalities are the development poles of some sub-regions. On the one hand, there is a high level of mobility added to the fact that some areas are more attractive to long-term

housing, which means it is easier for that population to commute and work in another municipality with better jobs and services. Furthermore, the municipalities with lower mobility or lower residential attractiveness group people that work and live in the same municipality on less valued jobs.

When comparing municipalities between islands, the differences appeared at the socio-economic level and not at the residential attractiveness level, suggesting that some islands might have better-living conditions even though that does not mean population would change island only because of it. The "pockets of underdevelopment" generated in the more remote sub-regions lack social support from the government. This situation could potentialize the creation of social employment in areas where it is clearly needed. This goes in line with the need explained before to look at social indicators when characterizing a region. When considering social matters, more importantly than uni-dimensional GDP measures, are the indicators of wealth and social progress. An example of the usefulness of considering such indicators is the Stiglitz-Sen-Fitoussi Commission in 2008, where the well-being and life conditions were studied using indicators beyond GDP. In Portugal, the importance of measuring the overall conditions of life of the population reflected on the release in 2004 of a new indicator by INE called the well-being index. Even though the Portuguese Statistical Institution already provides many socio-economic indicators, there was a need to account for the multiple social factors that contribute to the population's conditions of life.

The island-level socio-economic dispersion goes in line with the "poly-insularity" concept referred to before. Using a smaller geographic unit, one can distinguish some municipalities of the same island as quite different from their surrounding sub-regions. In some cases, the remoteness explained at the beginning is translated into a "born here live here" way of thinking, seen for instance in cluster 2, where mobility is low, and the resident population lives in its own whole-family house and is less qualified, working in lesser valued professions. Looking at the municipalities of this cluster, like Lajedo or Fajãzinha (Flores), Fenais da Ajuda or Nordeste (São Miguel), and so on, they are characterized by having a more remote land access which in this case complicates the existence or condition of some services which combined with the population low-qualified work propensity might explain their higher proportion of residents with no school level completed. There are reports of shortages of main necessity goods for some grocery stores or even gas in remoter areas where land or sea access is hampered by the weather conditions. The outlier municipality Água Retorta (São Miguel) can

be seen as an example of this, since it is a municipality that provides the basic needs of its population, translating into a lack of mobility towards other municipalities since its population tends to be born, work and live at the same place in a long period of time. These are sub-regions also characterized by a well-known phenomenon where larger households rely on the existence of jobs with easier access like agriculture, fisheries, and construction work which are attractive to a younger population, even if it translates into a lower income. Surrounding the municipalities of this cluster, there are sometimes other municipalities from clusters 1 or 4, for instance, with entirely different socio-economic outcomes. This proves the need to avoid geography generalizations and consider the specificity of municipality's behavior to manage their governmental funds better.

8. CONCLUSIONS

This work distinguishes itself from other regional studies by using a smaller geographic unit to study a wide range of indicators provided by the census data collected with methodologic support to characterize Azores' sub-regions better.

As could be seen, the Azores region is not homogenous from a socio-economic point of view. This heterogeneity was shown at the municipality level, revealing “pockets of underdevelopment” in some sub-regions. At a first look, the tendency seems to be that the surrounding areas of a municipality are similar to it; however, this appears to be different for some municipalities that stand out by their behavior regarding some socio-economic indicators. This work provided evidence that some groups of municipalities are either considerably more remote or work as a development pole for their sub-region. An example of this is the expected behavior of the municipalities considered as capitals for their island, as is Ponta Delgada for São Miguel or Angra do Heroísmo for Terceira.

A third of the municipalities of this region are characterized by an aged population living in aged own houses, whose main livelihood comes from other activities rather than work. There is a prevalence of the first and second sector activities, even though the third sector is predominant in the municipalities of the main cities. The population is portrayed as having lower school levels and an overall smaller family cradle, living alone, or mainly having fewer children. There is also a pattern for some of the remote areas to be attractive for a foreign population to move in or simply have a population who went to work or live abroad. Due to the methods of study chosen, it is now possible to pinpoint deviations from this portrait, which was the purpose of this work. Deviations start to appear in the two second-highest municipality groupings where the population is either younger or families are bigger with a higher number of couples with children. This is also the municipality grouping where the main livelihood comes from work, in more valued professions by more qualified residents. Adding to this are other distinct municipalities where the population is either younger with a big family cradle or young but living alone, working on the main economic areas. Finally, a third portrait, even more distinct, is drawn for a smaller number of municipalities with a high residential attractiveness where the population is young, mobile, qualified, and has bigger families with more children. This thorough characterization was only possible due to the scrutinizing

methodology used to study the region, which corroborated the hypotheses of existing disparities within the same region.

Overall, the purpose of this dissertation is then fulfilled by being able to scrutinize sub-regional outcomes, creating a distinctive territorial portrait of the different sub-regions. This shows that using a smaller geographical unit and reliable statistical methods, one can better grasp the socio-economical differences felt by some municipalities of the same region. As for the Azores case, what leads the sub-regions apart on this complete and detailed portrait seems to be the overall family cradle nature along with the job propensity on some sub-regions, which is also dependent on the type of professions and sectors predominant. The development poles described before act as job creators for the family living in those areas, their surroundings, and on more distant residential areas with higher mobility. As for the more remote municipalities, they are dependent on their mobility to the closer main city, or they are left to endure by developing their small businesses providing what they need for the enclosed population.

9. LIMITATIONS AND FUTURE WORK RECOMMENDATIONS

This work studies the general context of a varied gathering of variables in a multivariate statistical study. However, one limitation of it is the fact that the fixed attributes are the individuals, in this case, the municipalities, and the variable attributes are the indicators chosen since they were influenced in the beginning by the indicators chosen by similar studies and by the author's judgment. One interesting future work regarding this limitation would be to check how much the cluster formation would change if the indicators would change to slightly different ones that would still be able to characterize each sub-region from a socio-economic point of view. In this sense, using the indicators as the randomly selected attributes, one can see if the characterization done in this work is accurate or more dependent on the variables selected to study.

A robust PCA study could be the solution for another possible limitation involving the increasing importance of studying the demographic evolution of these sub-regions and how they affect socio-economic outcomes, especially of more extremely older municipalities. As mentioned before, the average age of the Portuguese population tends to increase throughout the years. In this study, age was used as a simple fixed indicator, either the average age or the derived indicator of elderly dependency. In a future work, it would be interesting to tackle this limitation by studying the age pyramid distribution of each municipality and see how they behave according to the different socio-economic indicators, considering any outlier behavior, which is accounted by the robust method of study. This study was done using the classic PCA methods since a multivariate outlier analysis was done beforehand.

Additionally, one of the major drawbacks of using a data source as the census one is that there is an absence of variables related to the well-being of the population, as the presence of health or educational institutions, in a way that does not necessarily measure the investment done to those areas but the outcomes from it; the social relations of different communities, especially particular in remote areas; matters of security, environment or public participation, and so on. All these indicators help understand how a population evolves, more than some economic indicators like the variation of GDP. As such, a suggestion for the future would be to confront the census data with auxiliary variables of this sort, which better describes the socio-economic portrait of a region.

10.REFERENCES

Boldea, M., Parean, M., & Otil, M. (2012). Regional Disparity Analysis: The Case of Romania. *Journal of Eastern Europe Research in Business & Economics*, *2012*, 1–10. https://doi.org/10.5171/2012.599140

Cziráky, D., Puljiz, J., Jurlin, K., Malekovi, S., & Poli, M. (2003). *An Econometric model for development level assessment with an application to municipality development classification*.

Cziraky, D., Puljiz, J., Jurlin, K., Maleković, S., & Polić, M. (2002). A multivariate methodology for modelling regional development in Croatia. *Croatian International Relations Review*, *8*(26/27), 35–52.

Diogo, F. (2019). Algumas Peculiaridades Da Pobreza Nos Açores. *Sociologia on Line*, *2018*(19), 81–101. https://doi.org/10.30553/sociologiaonline.2019.19.4

Gan, G., Ma, C., & Wu, J. (2007). Data Clustering: Theory, Algorithms, and Applications, Second Edition. In *Data Clustering: Theory, Algorithms, and Applications, Second Edition*. https://doi.org/10.1137/1.9781611976335

Hedlund, M. (2016). Mapping the Socioeconomic Landscape of Rural Sweden: Towards a Typology of Rural Areas. *Regional Studies*, *50*(3), 460–474. https://doi.org/10.1080/00343404.2014.924618

Hubert, M., Rousseeuw, P. J., & Van Aelst, S. (2005). Multivariate Outlier Detection and Robustness. *Handbook of Statistics*, *24*(December), 263–302. https://doi.org/10.1016/S0169-7161(04)24010-X

Jr, J. F. H., Black, W. C., Babin, B. J., Anderson, R. E., Black, W. C., & Anderson, R. E. (2018). *Multivariate Data Analysis*. https://doi.org/10.1002/9781119409137.ch4

L, M. O. . R. (2017). Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers. *Environment and Planning B: Urban Analytics and City Science*, *44*(3), 441–463. https://doi.org/10.1177/0265813516638849

Lipshitz, G., & Raveh, A. (1994). Application of the Co-plot Method in the Study of Socio-economic Differences between Cities: A Basis for a Differential Development Policy.

Urban Studies, *31*(1), 123–135. https://doi.org/10.1080/00420989420080071

Patra, A. K., & Acharya, A. (2011). Regional Disparity, Infrastructure Development and Economic Growth: An Inter-State Analysis. *Research and Practice in Social Sciences Patra, A & Acharya*, *6*(2), 17–30.

Pavão, G., Rocha, N., & Maciel, R. F. (2020). *Caracterização da Dinâmica Demográfica Recente dos Açores e das Qualificações da População . Cenários de Evolução até 2030 e Estratégias para o Desenvolvimento Económico , Social e Recuperação Populacional das Ilhas Açorianas*.

Pettersson, Ö. (2001). Microregional fragmentation in a Swedish county. *Papers in Regional Science*, *80*(4), 389–409. https://doi.org/10.1007/PL00013630

Rovan, J., & Sambt, J. (2003). Socio-economic Differences Among Slovenian Municipalities : A Cluster Analysis Approach. *Developments in Applied Statistics*. http://www.stat-d.si/mz/mz19/rovan.pdf

Santos, L. (2014, May 4). *Açores com maior número de famílias numerosas*. https://www.dn.pt/portugal/acores-com-maior-numero-de-familias-numerosas-3843402.html

Sayfudinova, N. Z., Timofeev, R. A., & Makhiyanova, A. V. (2016). Methodological basis of the regional systems socio-economic profile using survey method. *Journal of Economics and Economic Education Research*, *17*(SpecialIssue2), 325–333.

Soares, J. O., Marquês, M. M. L., & Monteiro, C. M. F. (2003). A multivariate methodology to uncover regional disparities: A contribution to improve European Union and governmental decisions. *European Journal of Operational Research*, *145*(1), 121–135. https://doi.org/10.1016/S0377-2217(02)00146-7

Ul Hadia, N., Abdullah, N., & Sentosa, I. (2016). An Easy Approach to Exploratory Factor Analysis: Marketing Perspective. *Journal of Educational and Social Research*, *February*. https://doi.org/10.5901/jesr.2016.v6n1p215

Wang, F. (2016). Analysis on the Regional Disparity in China and the Influential Factors. *American International Journal of Humanities and Social Science*, *2*(4), 94–104. www.cgrd.org

Zambon, I., Serra, P., Sauri, D., Carlucci, M., & Salvati, L. (2017). Beyond the 'mediterranean city': Socioeconomic disparities and urban sprawl in three Southern European cities. *Geografiska Annaler, Series B: Human Geography*, *99*(3), 319–337. https://doi.org/10.1080/04353684.2017.1294857

11. ANNEXES

	Index
Prop of buildings not exclusively residential	A
Prop of leased or sub-leased classic family accommodation	B
Prop of own housing with charges	C
Prop of overcrowded accommodation	D
Average age of resident population	E
Average age of buildings	F
Prop of resident population working or studying in another municipality	G
Proportion of car use when traveling	H
Prop of resident population with 15 and more years old whose main livelihood is work	I
Prop of resident population that 5 years previously lived outside the municipality	J
Prop of single-person classic families	K
Prop of classic families with 5+	L
Prop of family nuclei of couples with children	M
Prop of buildings needing major repairs or degraded	N
Average households per accommodation	O
Prop of resident population of foreign nationality	P
Prop of socially most valued professionals	Q
Unemployment rate	R
Elderly dependency index	S
Prop of resident population (Who has lived abroad for a continuous period of at least 1 year)	T
Proportion of dwellings with heating	U
Prop. Of population 15+ with no school level completed	V

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
A	1	0.4586	0.0445	-0.2581	0.3705	0.1645	-0.0877	0.0175	0.0485	0.4678	0.5732	-0.3576	-0.4066	-0.142	0.2161	0.273	0.5342	-0.2218	0.2873	0.1044	0.3423	-0.2942
B	0.4586	1	0.265	-0.0197	0.0388	-0.1067	-0.0979	-0.0474	0.0138	0.5051	0.4667	-0.1334	-0.1457	-0.2474	0.3294	0.1806	0.5287	0.1492	-0.0377	-0.0808	0.3117	-0.204
C	0.0445	0.265	1	0.0603	-0.4168	-0.2413	0.0777	0.3742	0.5836	0.2291	-0.0903	0.0624	0.2838	-0.2044	0.0681	-0.1651	0.4286	0.0244	-0.5228	-0.3862	0.405	-0.262
D	-0.2581	-0.0197	0.0603	1	-0.7172	0.0072	0.3084	-0.5664	-0.2288	-0.2254	-0.5337	0.8307	0.7509	0.0267	-0.0671	-0.336	-0.3282	0.7128	-0.56	-0.4616	-0.5346	0.4357
E	0.3705	0.0388	-0.4168	-0.7172	1	0.1826	-0.3338	0.231	-0.2703	0.2186	0.7188	-0.7839	-0.8983	0.0946	0.1354	0.5262	0.2503	-0.5768	0.926	0.5987	0.2736	-0.2856
F	0.1645	-0.1067	-0.2413	0.0072	0.1826	1	-0.0708	-0.1918	-0.2751	-0.0213	0.1797	-0.0909	-0.048	0.1968	0.0585	-0.0511	-0.0488	0.048	0.1934	0.0551	-0.1027	0.1307
G	-0.0877	-0.0979	0.0777	0.3084	-0.3338	-0.0708	1	-0.2288	-0.0142	0.2906	-0.3318	0.2639	0.2917	-0.1116	0.2382	-0.1858	-0.1035	0.182	-0.2525	-0.1195	-0.1503	0.0901
H	0.0175	-0.0474	0.3742	-0.5664	0.231	-0.1918	-0.2288	1	0.6456	0.1424	0.1618	-0.5394	-0.3653	-0.0794	0.0516	0.1532	0.3561	-0.5458	0.0058	-0.0059	0.6749	-0.5246
I	0.0485	0.0138	0.5836	-0.2288	-0.2703	-0.2751	-0.0142	0.6456	1	0.1987	-0.0656	-0.1889	0.0689	-0.1721	0.1201	-0.0393	0.3282	-0.3709	-0.4788	-0.32	0.5212	-0.5055
J	0.4678	0.5051	0.2291	-0.2254	0.2186	-0.0213	0.2906	0.1424	0.1987	1	0.6011	-0.4219	-0.3948	-0.1895	0.5015	0.3749	0.6748	-0.1905	0.1251	0.2526	0.4172	-0.4858
K	0.5732	0.4667	-0.0903	-0.5337	0.7188	0.1797	-0.3318	0.1618	-0.0656	0.6011	1	-0.7489	-0.7377	0.0141	0.3209	0.4912	0.6317	-0.3832	0.6155	0.4251	0.4169	-0.4451
L	-0.3576	-0.1334	0.0624	0.8307	-0.7839	-0.0909	0.2639	-0.5394	-0.1889	-0.4219	-0.7489	1	0.7937	-0.021	-0.3174	-0.413	-0.4658	0.6443	-0.5946	-0.4643	-0.5396	0.5517
M	-0.4066	-0.1457	0.2838	0.7509	-0.8983	-0.048	0.2917	-0.3653	0.0689	-0.3948	-0.7377	0.7937	1	-0.0294	-0.1856	-0.6259	-0.371	0.6195	-0.8056	-0.5699	-0.3795	0.3773
N	-0.142	-0.2474	-0.2044	0.0267	0.0946	0.1968	-0.1116	-0.0794	-0.1721	-0.1895	-0.0141	-0.021	-0.0294	1	-0.1258	0.1641	-0.2337	-0.0717	0.1727	0.1531	-0.2291	0.112
O	0.2161	0.3294	0.0681	-0.0671	0.1354	0.0585	0.2382	0.0516	0.1201	0.5015	0.3209	-0.3174	-0.1856	-0.1258	1	0.214	0.3733	-0.0075	0.0436	0.0427	0.2036	-0.2884
P	0.273	0.1806	-0.1651	-0.336	0.5262	-0.0511	-0.1858	0.1532	-0.0393	0.3749	0.4912	-0.413	-0.6259	0.1641	0.214	1	0.2155	-0.3766	0.4891	0.4657	0.2191	-0.2655
Q	0.5342	0.5287	0.4286	-0.3282	0.2503	-0.0488	-0.1035	0.3561	0.3282	0.6748	0.6317	-0.4658	-0.371	-0.2337	0.3733	0.2155	1	-0.2301	0.0876	0.0488	0.6418	-0.5826
R	-0.2218	0.1492	0.0244	0.7128	-0.5768	0.048	0.182	-0.5458	-0.3709	-0.1905	-0.3832	0.6443	0.6195	-0.0717	-0.0075	-0.3766	-0.2301	1	-0.4767	-0.3758	-0.4981	0.4095
S	0.2873	-0.0377	-0.5228	-0.56	0.926	0.1934	-0.2525	0.0058	-0.4788	0.1251	0.6155	-0.5946	-0.8056	0.1727	0.0436	0.4891	0.0876	-0.4767	1	0.6416	0.0358	-0.0753
T	0.1044	-0.0808	-0.3862	-0.4616	0.5987	0.0551	-0.1195	-0.0059	-0.32	0.2526	0.4251	-0.4643	-0.5699	0.1531	0.0427	0.4657	0.0488	-0.3758	0.6416	1	0.0295	-0.0541
U	0.3423	0.3117	0.405	-0.5346	0.2736	-0.1027	-0.1503	0.6749	0.5212	0.4172	0.4169	-0.5396	-0.3795	-0.2291	0.2036	0.2191	0.6418	-0.4981	0.0358	0.0295	1	-0.632
V	-0.2942	-0.204	-0.262	0.4357	-0.2856	0.1307	0.0901	-0.5246	-0.5055	-0.4858	-0.4451	0.5517	0.3773	0.112	-0.2884	-0.2655	-0.5826	0.4095	-0.0753	-0.0541	-0.632	1

Table 20 Correlation Matrix and Labels

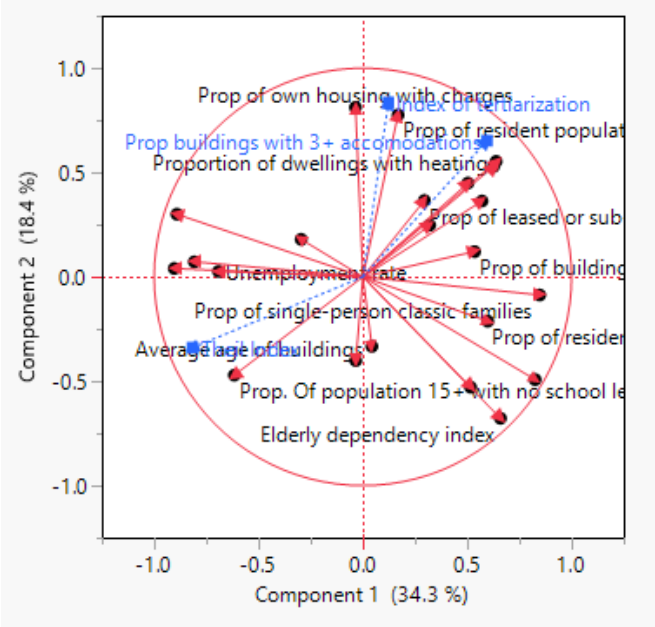


Figure 27 *Representation of Component 1 (Demography) and Component 2 (Socio-Economic)*

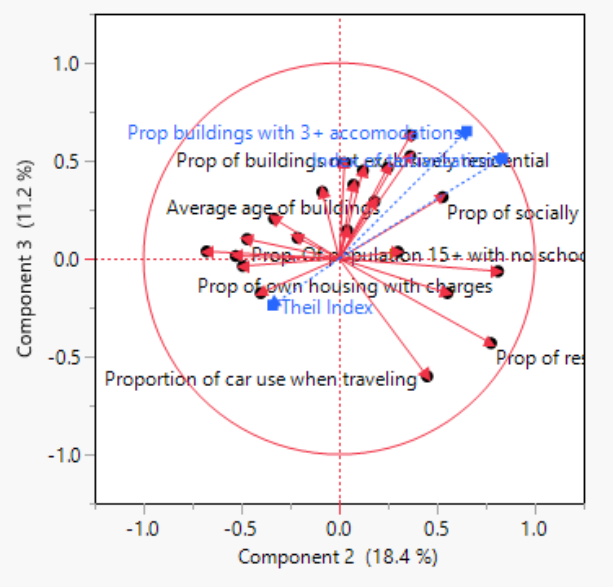


Figure 28 *Representation of Component 2 (Socio-Economic) and Component 3 (Residential Attractiveness)*

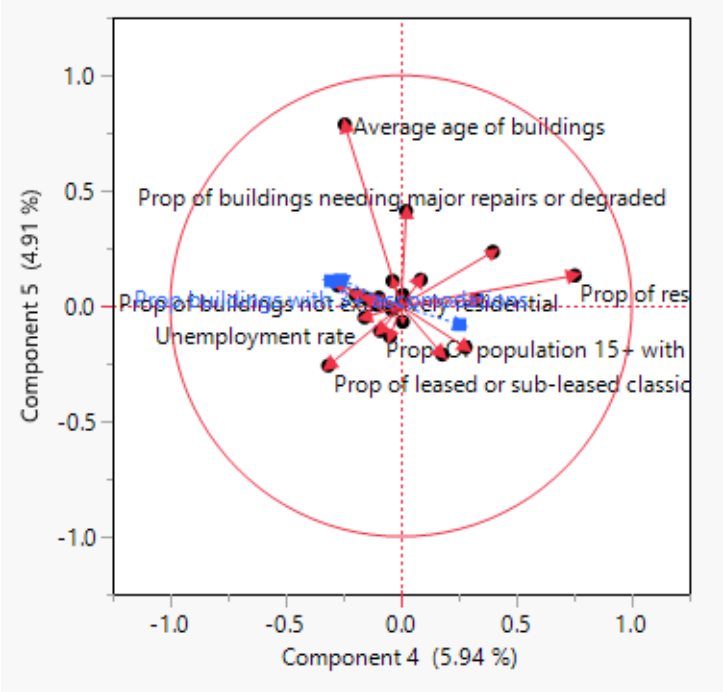
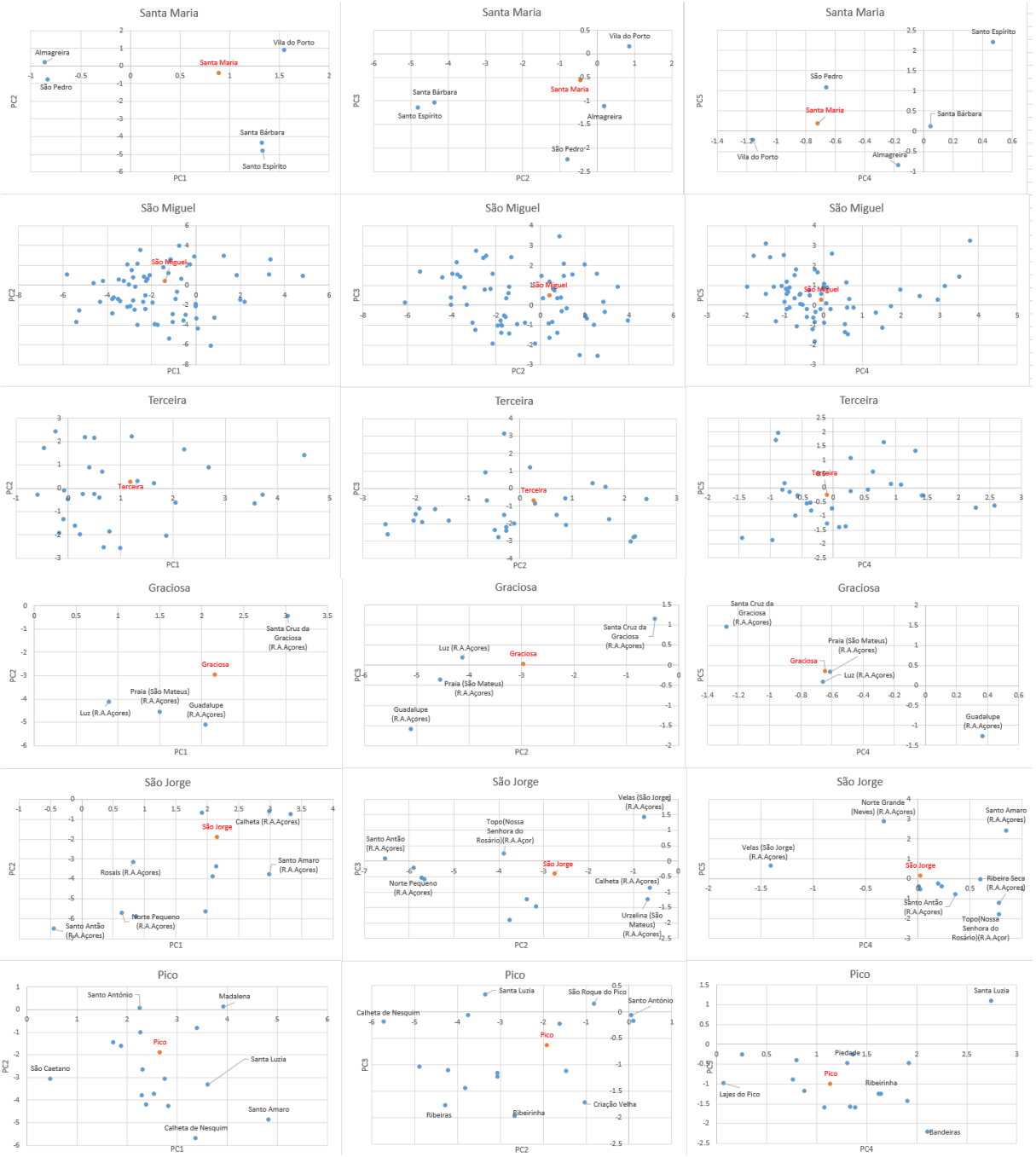


Figure 29 *Representation of Component 4 (Mobility) and Component 5 (Building Condition)*


Santa Maria
Vila do Porto
Almagreira
Santa Maria
São Pedro
Santa Bárbara
Santo Espírito
PC1
PC2
PC3
PC4
PC5
São Miguel
Terceira
Graciosa
Santa Cruz da Graciosa (R.A.Açores)
Praia (São Mateus) (R.A.Açores)
Luz (R.A.Açores)
Guadalupe (R.A.Açores)
São Jorge
Calheta (R.A.Açores)
Rosais (R.A.Açores)
Santo Amaro (R.A.Açores)
Norte Pequeno (R.A.Açores)
Santo Antão (R.A.Açores)
Velas (São Jorge) (R.A.Açores)
Topo(Nossa Senhora do Rosário)(R.A.Açor)
Urzelina (São Mateus) (R.A.Açores)
Norte Grande (Neves) (R.A.Açores)
Ribeira Seca (R.A.Açores)
Pico
Madalena
São Caetano
Santa Luzia
Calheta de Nesquim
São Roque do Pico
Ribeiras
Ribeirinha
Criação Velha
Piedade
Lajes do Pico
Bandeiras

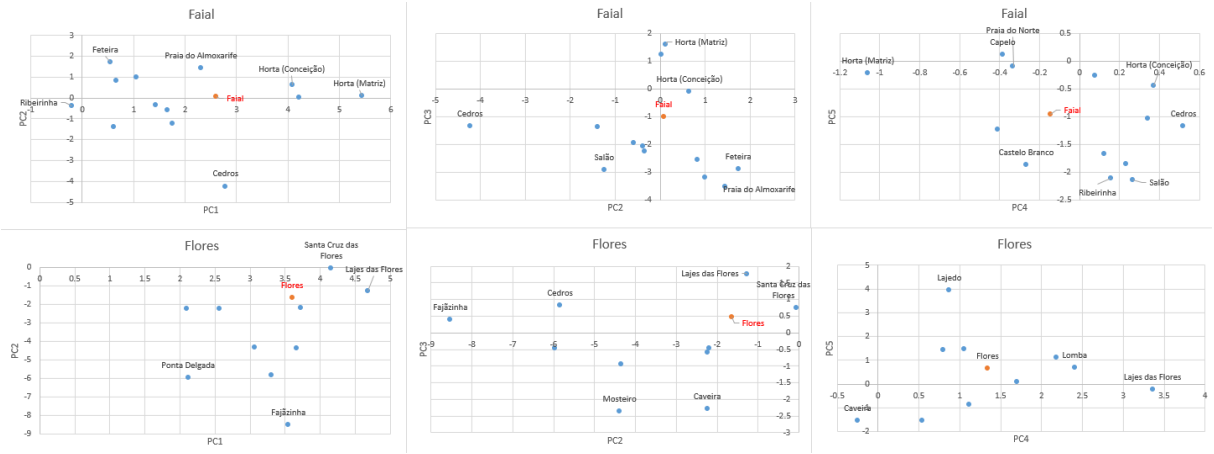


Figure 30 Representation of municipalities by island and principal components

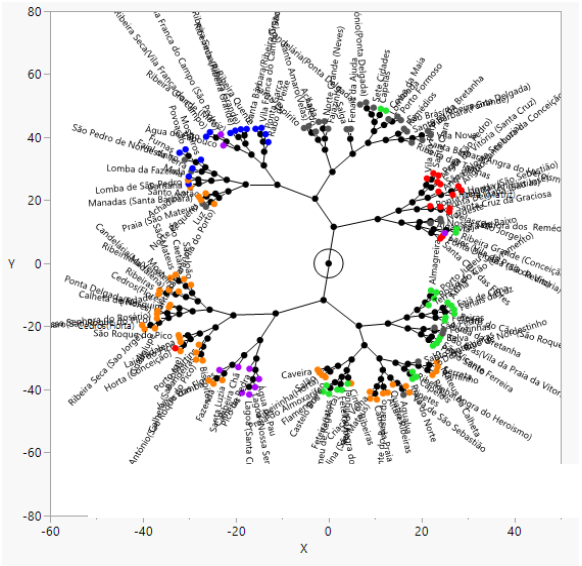


Figure 31 Constellation Plot

Cluster Standard Deviations Lowest->more orange

Cluster	Count	Prin1	Prin2	Prin3	Prin4	Prin5
1	10	1.18895	1.04259	0.57658	0.53662	0.6625
2	7	0.64527	1.38317	0.35136	0.63346	0.63186
3	17	0.85879	1.11001	0.50214	0.66522	0.36216
4	11	0.76141	1.52851	0.58428	0.24557	0.43427
5	5	1.44861	2.06029	1.24899	0.59365	0.45342
6	5	1.76361	0.89094	0.9767	0.66358	1.24082
7	9	0.86727	0.86368	0.70246	0.95105	0.5962
8	16	0.92877	1.01108	0.84855	0.51464	0.58384
9	6	0.6773	0.97499	0.43908	0.40101	0.44671
10	10	1.12548	1.21702	0.57604	0.75786	0.55351
11	12	1.3965	1.59992	0.99517	0.76365	0.81303
12	6	1.11596	1.69127	0.843	0.71506	0.85281
13	2	2.48153	0.1273	0.42141	1.43077	0.07174
14	14	1.51764	1.04936	0.45185	0.46021	0.75534
15	13	1.23823	0.79123	1.03316	0.53677	0.93448
16	8	1.44841	2.36546	0.79854	0.5698	0.48389

Table 21 Cluster Std Deviations

Cluster Comparison

Method	NCluster	CCC	Best
K Means Cluster	3	-1.8498	
K Means Cluster	4	-5.5976	
K Means Cluster	5	-6.1604	
K Means Cluster	6	-1.9932	
K Means Cluster	7	-2.725	
K Means Cluster	8	-3.3065	
K Means Cluster	9	-1.8401	
K Means Cluster	10	-3.9728	
K Means Cluster	11	-4.6335	
K Means Cluster	12	-1.5241	Optimal CCC
K Means Cluster	13	-5.3214	
K Means Cluster	14	-3.4774	
K Means Cluster	15	-3.0111	
K Means Cluster	16	-4.7142	

Table 22 K-means Optimal Solution

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
• Vila do Porto • Ponta Delgada (São Sebastião) • Ponta Delgada (São José) • Ponta Delgada (São Pedro) • Santa Clara • Angra (Nossa Senhora da Conceição) • Angra (Santa Luzia) • Angra (São Pedro) • São Bento • Praia da Vitória (Santa Cruz) • Santa Cruz da Graciosa • Velas (São Jorge) • Horta (Angústias) • Horta (Conceição) • Horta (Matriz) • Santa Cruz das Flores	• Santo Espírito • São Pedro • Achada • Lomba da Fazenda • Nordeste • Salga • Algarvia • Santo António de Nordestinho • São Pedro de Nordestinho • Candelária(Ponta Delgada) • Remédios • Santa Bárbara(Ponta Delgada) • Santo António(Ponta Delgada) • Sete Cidades • Pilar da Bretanha • Nossa Senhora dos Remédios • Fenais da Ajuda • Lomba da Maia • Porto Formoso • São Brás(Ribeira Grande) • Ribeira das Tainhas • Santa Bárbara(Angra do Heroismo) • Agualva • Fontinhas • São Brás(Vila da Praia da Vitoria) • Vila Nova • Luz • Praia (São Mateus) • Norte Grande (Neves) • Santo Amaro(Velas) • Fajãzinha • Lajedo	• Santana • Mosteiros • Faial da Terra • Furnas • Povoação • Ribeira Quente • Maia • Rabo de Peixe • Ribeira Grande (Matriz) • Ribeira Seca (Ribeira Grande) • Ribeirinha(Ribeira Grande) • Santa Bárbara(Ribeira Grande) • Água de Alto • Ponta Garça • Vila Franca do Campo (São Miguel) • Ribeira Seca(Vila Franca do Campo)	• Santa Bárbara(Vila do Porto) • Achadinha • Ginetes • Lomba de São Pedro • Vila Franca do Campo (São Pedro) • Altares • Cinco Ribeiras • Doze Ribeiras • Raminho • Serreta • Vila de São Sebastião • Biscoitos • Cabo da Praia • Fonte do Bastardo • Quatro Ribeiras • Porto Martins • Guadalupe • Calheta (Sao Jorge) • Norte Pequeno • Ribeira Seca (Sao Jorge) • Santo Antão • Topo(Nossa Senhora do Rosário) • Manadas (Santa Bárbara) • Rosais • Urzelina (São Mateus) • Calheta de Nesquim • Lajes do Pico • Piedade • Ribeiras • Ribeirinha(Velas) • São João • Bandeiras • Candelária(Madalena) • Criação Velha • Madalena • São Caetano • São Mateus • Santa Luzia • Santo Amaro(Sao Roque do Pico) • Santo António(Sao Roque do Pico) • São Roque do Pico • Castelo Branco	• Almagreira • Arrifes • Capelas • Covoada • Fajã de Baixo • Fajã de Cima • Fenais da Luz • Feteiras • Relva • Rosto do Cão (Livramento) • Rosto do Cão (São Roque) • São Vicente Ferreira • Ajuda da Bretanha • Ribeira Grande (Conceição) • Feteira(Angra do Heroismo) • Porto Judeu • Posto Santo • Ribeirinha(Angra do Heroismo) • São Bartolomeu de Regatos • São Mateus da Calheta • Terra Chã • Capelo • Feteira(Horta) • Flamengos • Pedro Miguel • Praia do Almoxarife • Praia do Norte	• Água de Pau • Cabouco • Lagoa (Nossa Senhora do Rosário) • Lagoa (Santa Cruz) • Ribeira Chã • Calhetas • Pico da Pedra • Lajes (Vila da Praia da Vitoria) • Lajes das Flores

			• Cedros(Horta) • Ribeirinha(Horta) • Salão • Fazenda • Lomba • Mosteiro • Caveira • Cedros(Flores) • Ponta Delgada(Flores)		

Table 23 Municipalities in each cluster