Testing a Methodologic Approach to Territory Categorization Using Census Data

Margarida Miguel Costeira e Pereira

Research Centre for Anthropology and Health; Department of Geography and Tourism of the University of Coimbra

Helena Guilhermina da Silva Marques Nogueira

Research Centre for Anthropology and Health; Department of Geography and Tourism of the University of Coimbra

Abstract

The more traditional perspective in Geography views territory as a dichotomous reality of sociodemographic and biophysical characteristics yet, literature states that both territorial dimensions coexist and interact. The number of indicators for each territorial dimension is vast which makes the territory categorization a complex task. Despite it, territory categorization generates knowledge about the differences between territories that improve the planning process and make it more directed and strategic. This work main goal is to test a methodologic approach to categorize Lisbon Municipality area at the statistical section level, integrating information on the physical and social aspects. For this purpose, multivariate and complementary statistical analysis techniques will be used: Principal Component Analysis (applied to the physical and social aspects of the territory) and Hierarchical Classification Analysis (based in the factors extracted in the Principal Component Analysis). The resulting clusters will then be mapped and its distribution will be tested for spatial autocorrelation.

Keywords: Territory; Principal Components Analysis; Hierarchical Classification Analysis; I of Moran.

Introduction

Territory can be defined as a delimited and precise space, with multiple dimensions, in which several power relations interact (Albagli, 2004; Faria and Bortolozzi, 2009).

Territory results from the integration of culture in the environment (Braga, 2007), however, it is not only a social construction resulting from historical power relations, but also a natural space (Haesbaert and Limonad, 2007).

The definition of territory varies according to different sciences, which makes this concept ambiguous (Faria and Bortolozzi, 2009). From a political point of view, territory is closely related to the borders of the State/Nation as in the environmental perspective the territorial biophysical characteristics are emphasized (Faria and Bortolozzi, 2009). Once territory integrates several dimensions, namely geographical, anthropological, cultural, social, economical and bioecological (Albagli, 2004), one can state that the choice of the aspects to be highlighted depends on the research objective and the researchers own conceptualization of territory (Faria and Bortolozzi, 2009).

This study is based on a geographical interpretation of the territory as a combined space, resulting from a set of processes in which the material, physical and social aspects of human action are inseparable and interdependent (Haesbaert and Limonad, 2007).

Albagli (2004: 27) states that: "Territorial differences and inequalities lie both in their own physical and social characteristics and in the way they fit into larger structures. Each territory is thus shaped by the combination of internal and external conditions and forces and must be understood as part of a spatial totality."

This means that each territorial unit is determined by endogenous and exogenous factors and also by the way they interact: they carry intrinsic characteristics, which give each territory a peculiar structure, and in turn determines the way each territory is integrated in wider spaces conditioning the evolution of its own internal structure.

The categorization and differentiation of each territorial unit by its level of development, or living conditions, for example, is an attempt of researchers to understand and explain the differences between territories (Martín e Barros et al., 2015). In addition, analyzing how territories relate to each other, observing the spatial networks formed according to certain aspects, allows strategic and more targeted planning (Martín and Barros, 2015).

For example, studies in the public health field often use deprivation indices at the area level rather than individual information to identify areas of greater vulnerability, thus less healthy, which are priority areas in the fight against health inequalities (Allik et al., 2016; Knighton et al., 2016). These indices, also called Geographic Deprivation Indices, generally use census data and are validated in Western Europe (Knighton et

al., 2016). Since deprivation is a multidimensional concept, composite indices in opposition to a single indicator better reflect the true level of deprivation at different geographic scales, as the existing evidence suggests (Lian et al., 2016).

As mentioned previously, each territory is the product of the interaction between physical and social aspects, and both dimensions of the territory are composed of a large number of indicators. The objective of this study is to classify the area of the Municipality of Lisbon, at the statistical section level, integrating information on the physical and social aspects, simultaneously, and to test the existence of spatial autocorrelation.

Methodology

This study was carried out in the Municipality of Lisbon, capital of Portugal, which is composed of 1054 sections. However, the analyses performed refer to 1053 sections due to lack of information for one section.

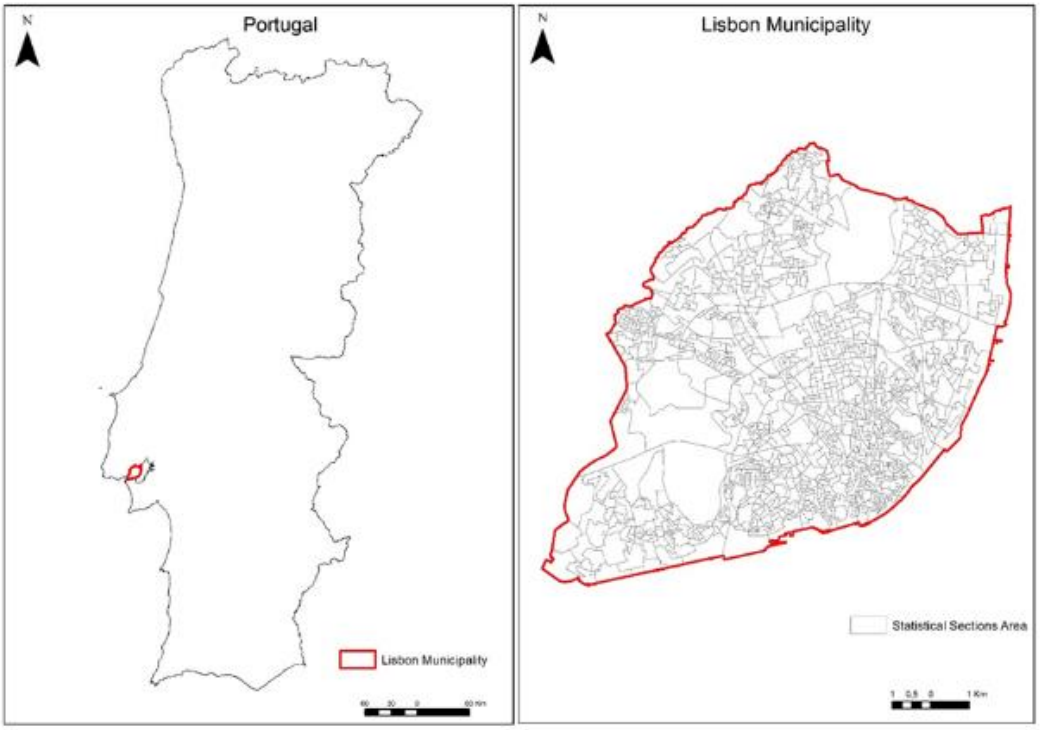

Fig. 1 Study Area: Lisbon Municipality by Statistical Sections

Data collection and organization

Data used in this study was collected at the statistical section level and was collected from the 2011 census. Sixty-one sociodemographic, economic and building variables were freely available and collected on the National Statistics Institute (INE) website. Information on land use (at level 2) was collected on the Territory General Directorate website and comprised 11 variables, dated from 2007.

Data were grouped in two large dimensions: Physical Aspects (PA) that include information on buildings and land use, and Social Aspects (SA) integrating demographic and socioeconomic information. Afterwards, the information was converted into percentage resulting in 34 indicators in the PA dimension and 24 indicators in the SA dimension.

Data Analysis

All statistical analysis was performed using SPSS (v.22) except the Moran I that was calculated in ArcGIS (v.10.4.1).

In order to reduce the number of indicators in each of the dimensions (PA and SA), the statistical method used was the Principal Components Factor Analysis (PCA). The PCA was performed with a Varimax orthogonal rotation, which facilitates the interpretation of the components by maximizing intra and intercomponent variation, thus seeking that in each component only some indicators present high loading values, appearing in the other factors with values close to zero.

PCA on PA initially resulted in 12 factors according to the Kaiser method (value of eigenvalues above 1). To reduce the number of factors, some of the original information was excluded (indicators that presented loadings between -0.49 and 0.49 in all the components initially extracted), which resulted in a new extraction of 6 factors. No indicators were excluded from the SA dimension because the results of the first analysis were more parsimonious: 5 factors. In addition to considering the loadings, the scores or coordinates resulting from the PCA were analysed and mapped.

PCA were followed and complemented by an Ascending Hierarchical Classification (AHC), commonly known as cluster analysis, performed with the 6 PA factors and the 5 SA factors. AHC was computed using the Ward method and the Euclidean quadratic distance as methods of similarities and distances between elements and between groups of elements.

Moran's I measures the spatial autocorrelation by analysing the degree of dependence between the values of the sections estimating how much of the value of each section depends on the values of the neighbouring sections. This index varies between -1 and 1 - the closer it is to 1 the stronger the spatial autocorrelation, the closer it approaches -1, the less similar are the neighbouring areas. When the value is equal to or close to zero it means that there is no spatial autocorrelation, that is, the study areas are spatially independent. In order to calculate the statistical significance of Moran's I, the pseudosignificance test was used.

Moran's I was calculated using CHA results (clusters), previously calculated in SPSS and imported into ArcMap. CHA was computed choosing the contiguity of boundaries and corners as the conceptualization of spatial relations, that is, taking into account the values of all the sections that share a border, a nodule or that overlap.

Results

Principal Components Factor Analysis

Physical Aspects

Table I shows the value of the highest loadings per factor and the respective percentage of variance explained. The six resulting factors account for almost 75% of the variance in this group of indicators.

Table I Loadings e % of variance explained per factor

Factor (% Explained Variance)	Indicator	*Loadings*	Designation
1 (28,76%)	% houses with water	0,953	Housing Conditions
	% houses with toilet	0,941	
	% houses with sewage	0,937	
	% houses with bath	0,731	
2 (14.49%)	% houses with 50m^2	0,855	House Dimension
	% houses with 1 or 2 divisions	0,696	
	% buildings without plate	0,668	
	% houses with 3 or 4 divisions	0,625	
3 (10.24%)	% empty houses	0,857	Non Residential
	% mainly non residential buildings	0,669	

	% family houses	-0,858	
4 (8.30%)	% buildings with 3 or more floors	0,838	Urban Area
	% urban land use	0,583	
	% buildings with 1 or 2 floors	-0,838	
5 (6.76%)	% buildings with plate	0,915	Construction Material
	% buildings with concrete	-0,723	
6 (5.60%)	% temporary crops	0,964	Agricultural Area

The first factor is related to housing conditions, such as the percentage of houses with water, toilet and sewage. The variables that most contribute to the second factor are the percentage of houses with 50m^2 followed by the houses with one or two divisions (Housing Dimension). The third factor is related to the percentage of empty houses and mainly non-residential buildings, as opposed to family houses (Non Residential). The percentage of buildings with 3 or more floors and the predominantly urban land use, as opposed to the percentage of buildings with only 1 or 2 floors are highlighted in the fourth factor (Urban Area). In the fifth factor, the indicator with the highest loading value is the percentage of buildings with plate (Construction Material), and finally in factor 6 (Agricultural Area), the percentage of land uses of temporary crops is the only indicator with a high loading value.

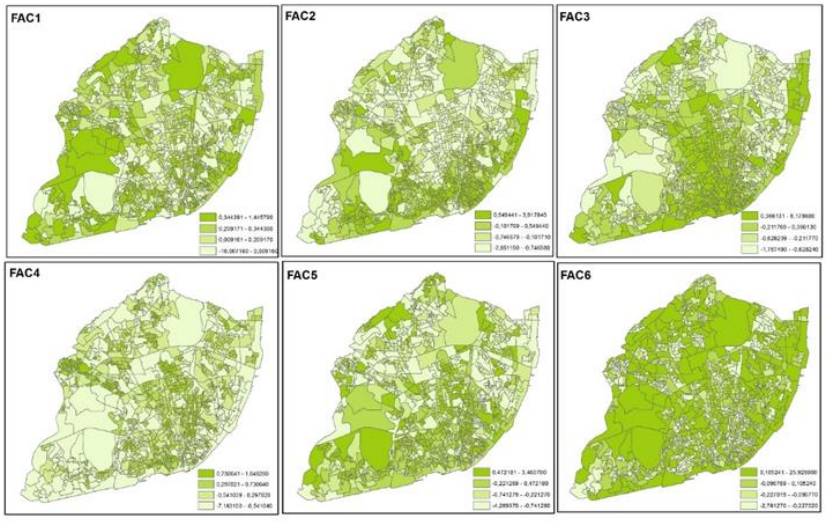

Fig 2 Spatial distribution of scores/coordinates of the factors categorized in quartiles

Figure 2 represents the spatial distribution of the scores/coordinates by statistical section, categorized in quartiles. This figure show that factors 1 (Housing Conditions), 2 (Housing Dimension) and 5 (Construction Material) do not appear to have any pattern in their distribution. Factor 3 (Non Residential) has higher values in the central sections as well as in the sections near the river. Most of the sections of the municipality of Lisbon have score values above 0 (mean value) in factor 4 (Urban Area), and the sections with larger area and more peripheral have the highest score values in factor 6 (Agricultural Areas).

Social Aspects

Table II shows the highest loadings per factor and the percentage of variance explained by the SA dimension. In this case, the five resulting factors account for almost 82% of the variance of this group of indicators.

Table II Loadings e % of variance explained per factor

Factor (% Explained Variance)	Indicator	*Loadings*	Designation
1 (35.68%)	% houses occupied by renters	0,889	Low Socioeconomic Level
	% residents looking for job	0,833	
	% classic families with unemployed	0,782	
	% residents who don´t read or write	0,733	
	% residents with 6 years of schooling	0,725	
	% residents with 4 years of schooling	0,714	
	% residents with academic degree	-0,765	
	% classic families without unemployed	-0,782	
	% houses occupied by owners	-0,867	
2 (26.39%)	% employed residents	0,779	Socioeconomic Stable
	% residents employed in the tertiary sector	0,765	
	% residents between 15 and 64 years old	0,759	
	% residents with 9 years of schooling	-0,456	
	% residents 65 years old	-0,723	
	% retired residents	-0,769	
3 (8.79%)	% classic families with 3 or 4 persons	0,805	Families with Under 15 y/o Children
	% nuclear families with children less than 15 years old	0,72	
	% residents employed in the secondary sector	0,444	
	% residents with 12 years of schooling	-0,677	

	% classic families with 1 or 2 persons	-0,829	
4 (5.59%)	% male residents	0,904	Residents Gender
	% female residents	-0,904	
5 (5.09%)	% residents looking for first job	0,539	Families with Over 15 y/o Children
	% nuclear families with children above 15 years old	0,807	

Indicators in factor 1 are associated with the labour situation, educational level and type of residence housing (Low Socioeconomic Level). The second factor highlights individuals employed in the tertiary sector, in working age residents as opposed to residents over 65 and retired (Socioeconomic Stable). In factor 3 the percentage of families with 3 or 4 persons and with children under 15 years of age have the highest positive loadings and smaller families with a high negative value (Families with Under 15 y/o Children). The fourth factor is related to the percentage of male and female residents (Residents Gender) and, finally, the percentage of families with children above 15 years of age in search of the first job have higher loadings in factor 5 (Families with Over 15 y/o Children).

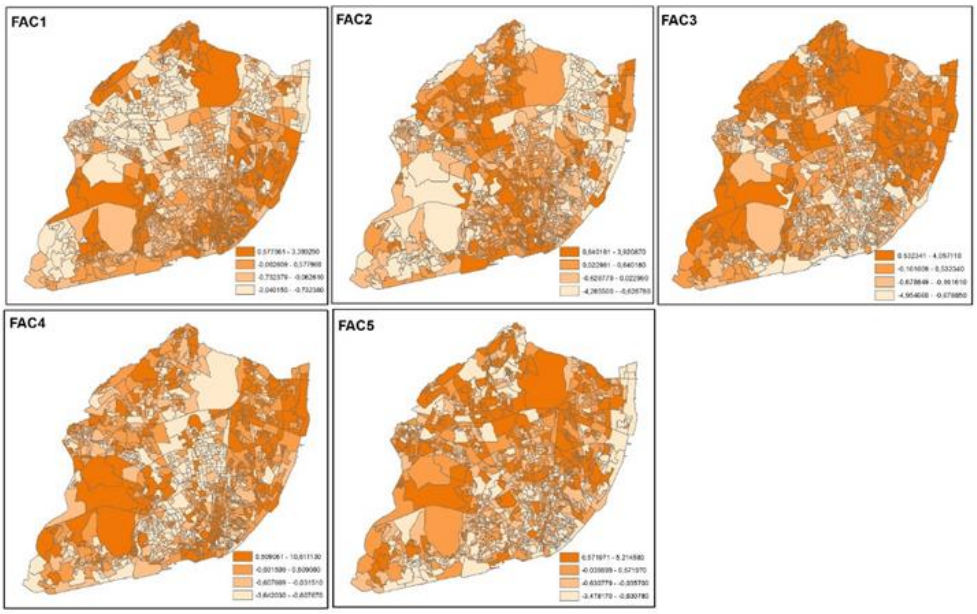

Fig. 3 Spatial distribution of scores/coordinates of the factors categorized in quartiles

The spatial distribution of factor 1 (Low Socioeconomic Level) shows that the sections with the highest scores are located in the centre of Lisbon and near the river. Factor 2 (Socioeconomic Stable) and 4 (Residents Gender) do not present a clear distribution pattern of the scores by section. Sections with the highest values in factor 3 (Families with Under 15 y/o Children), are located in the peripheral areas as well as in the northeast area of the municipality. Sections with scores higher in factor 5 (Families with Over 15 y/o Children) appear to be concentrated in the peripheral areas and to the north of the municipality.

Ascending Hierarchical Classification (Clusters Analysis)

The performed classification suggested the formation of four distinct clusters. Figure 4 and Table III synthesize information by cluster allowing a better understanding of each cluster meaning. In the graph of Figure 4, which presents the means of the factors per cluster, one can verify that each cluster represent a different conjugation of factors' values.

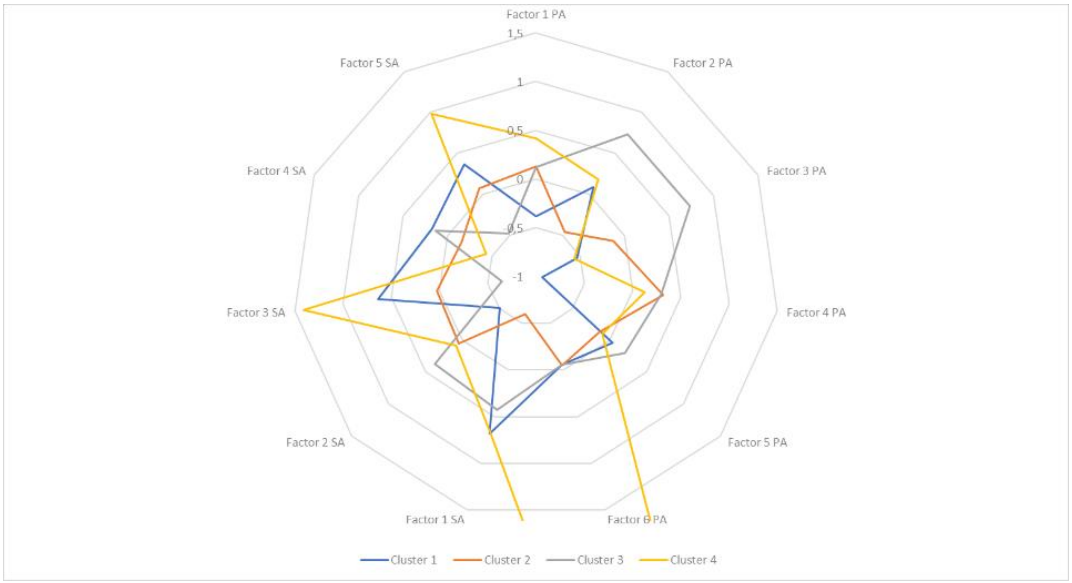

Fig 4 Mean factor score per cluster

Note: The average value of factor 6 PA (Agricultural Area) is too high to be represented in this graph (13,79).

In summary, cluster 1 integrates sections that are similar in terms of the size of housing (small housing), with a low socioeconomic level and with families with under and over 15 years old individuals. Cluster 2 predominantly consists of urban sections

with good housing conditions, families with over 15 years old individuals and employees in the tertiary sector. In cluster 3, sections are characterized by the small size of housing and mainly non-residential buildings. In this cluster, there are low socioeconomic level sections as well as sections with employees in the tertiary sector. Cluster 4 is uniquely related to the agricultural area, regarding land use typology.

Table III Synthetic Description of Cluster Factors

Clusters	Factors with highest score mean	Factor description	Designation
1	Factor 2 PA Factor 1 SA Factor 3 SA Factor 5 SA	House Dimensions Low Socioeconomic Level Families with Under 15 y/o Children Families with Over 15 y/o Children	Areas of Socioeconomic Vulnerability
2	Factor 4 PA Factor 1 PA Factor 5 SA Factor 2 SA	Urban Area Housing Conditions Families with Over 15 y/o Children Socioeconomic Stable	Areas of Greater Urbanity
3	Factor 2 PA Factor 3 PA Factor 1 SA Factor 2 SA	House Dimensions Non Residential Low Socioeconomic Level Socioeconomic Stable	Essentially Non-Residential Urban Areas
4	Factor 6 PA	Agricultural Area	Agricultural Areas

Cluster 1 (Areas of Socioeconomic Vulnerability) encompasses 24.79% of the population living in the study area, which underlines the precariousness conditions in which almost a quarter of the residents of this municipality live in. On the other hand, about 48% of the statistical sections, where almost half of the population of the municipality of Lisbon live in, are Areas of Greater Urbanity (cluster 2) - with good housing conditions and jobs in the tertiary sector.

In Cluster 3 (Essentially Non-Residential Urban Areas), there is a little more than 21% of the population of the whole municipality, in about 27% of its statistical sections. And in cluster 4 (Agricultural Areas) that encompasses only 4 sections with 2226 (0.38%) individuals.

Table IV Sections and population per cluster

Cluster	Sections		Population	
	N	%	N	%
1	261	24,79	153528	28,03
2	505	47,96	271641	49,59
3	283	26,88	120338	21,97
4	4	0,38	2226	0,41

Figure 5 reveals that, the spatial distribution of clusters seems to have a pattern. The value of Moran's I confirms that the distribution of clusters is spatially autocorrelated and significant once Moran's I equals 0.49 with a z-score=28.37 and p=0.000.

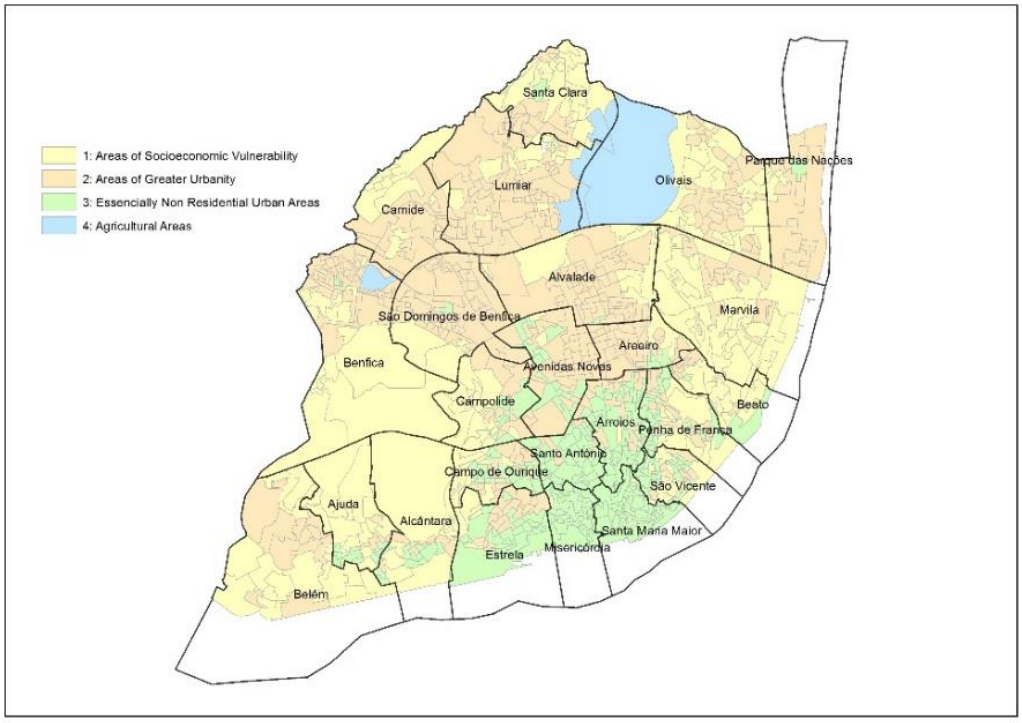

Fig. 5 Clusters spatial distribution

The map presented in Fig. 5 overlaps the administrative boundaries of each parish of the Municipality of Lisbon and the distribution of clusters by statistical section. In this map one can verify that there is internal homogeneity in the parishes with respect to the clusters of the sections that compose them. For example, the parishes of Misericórdia and Santa Maria Maior are almost exclusively composed by sections classified as cluster 3, Essentially Non Residential Urban Areas, that is, parishes where services and commerce predominate.

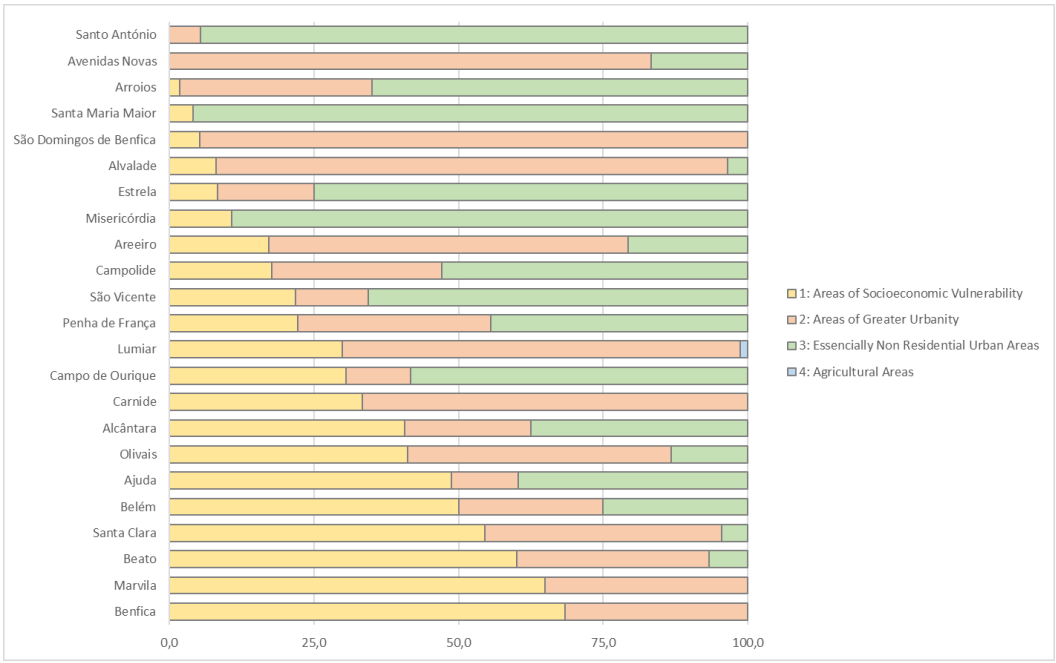

Fig. 6 Percentage of statistical sections per cluster and parish

Analysing the percentage of statistical sections by cluster and parish, we verified that:

Benfica, Marvila and Beato are the parishes with the highest percentage of sections classified as Socioeconomic Vulnerability Areas (cluster 1)

Cluster 2 (Areas of Greater Urbanity) is in a higher percentage in the parishes of Avenidas Novas, São Domingos de Benfica and Alvalade.

All parishes have sections classified as Essentially Non-Residential Urban Areas (cluster 3).

Only 4 sections are classified as Agricultural Areas (cluster 4).

Conclusions

The categorization of the territory is a complex task essentially due to the multidimensionality of the concept "territory" and to the large number of territorial indicators available. However, it is essential to know the territory in depth, to understand its differences and to interpret them. Only this way one can intervene to attenuate these differences, in particular through strategic and targeted planning.

PCA is the most appropriate statistical analysis to reduce the number of indicators needed to describe territorial units, without prejudice of losing relevant information. However, the spatial representation of the factor loadings distribution by sections is difficult to read and interpret.

The application of HCA to PA and SA factors together facilitates the interpretation of the distribution of these characteristics by cluster and section after mapping.

Moran's I proves the spatial autocorrelation of the mapping of the distribution of clusters by sections. Sections with similar PA and SA are geographically closer and spatially pooled significantly.

The grouping of sections by clusters is relatively uniform per parish and there are also parishes whose sections belong almost exclusively to a single cluster.

The methodology of factorial analysis, followed by hierarchical analysis and its mapping and spatial autocorrelation test through the Moran I, proved to be useful in the analysis of the statistical sections of the Municipality of Lisbon. The aggregation of sections by sections classification with clusters is not due to chance and the interpretation of the information of each cluster seems to be theoretically valid. However, this methodology needs replication in other territories and at different scales in order to prove its broader applicability.

References

[1] Albagli, S. (2004). Território e Territorialidade. In R. D. Editora (Ed.), Territórios em Movimento: Cultura e Identidade como Estratégia de Inserção Competitiva (pp. 23-70).

[2] Allik, M., Brown, D., Dundas, R., & Leyland, A. H. (2016). Developing a new small-area measure of deprivation using 2001 and 2011 census data from Scotland. Health Place, 39, 122-130. doi:10.1016/j.healthplace.2016.03.006

[3] Braga, R. M. (2007). O Espaço Geográfico: Um Esforço de Definição. GEOUSP - Espaço e Tempo, 22, 65-72.

[4] Faria, R. M. d., & Bortolozzi, A. (2009). Space, territory and health: contributions of Milton Santos for the theme of the geography of health in Brazil. Raega - O Espaço Geográfico em Análise, 17, 31-41.

[5] Haesbaert, R., & Limonad, E. (2007). O território em tempos de globalização. etc, espaço, tempo e crítica Revista Eletrônica de Ciências Sociais Aplicadas e outras coisas, 1(2 (4)), 39-52.

[6] Ianoş, I., Petrişor, A.-I., Zamfir, D., Cercleux, A.-L., Stoica, I.-V., & Tălângă, C. (2013). In search of a relevant index measuring territorial disparities in a transition country. Romania as a case study. Die Erde - Journal of the Geographical Society of Berlin, 144(1), 69-81. doi: 10.12854/erde-144-5

[7] Knighton, A. J., Savitz, L., Belnap, T., Stephenson, B., & VanDerslice, J. (2016). Introduction of an Area Deprivation Index Measuring Patient Socioeconomic

Status in an Integrated Health System: Implications for Population Health. EGEMS (Wash DC), 4(3), 1238. doi:10.13063/2327-9214.1238

[8] Lian, M., Struthers, J., & Liu, Y. (2016). Statistical Assessment of Neighborhood Socioeconomic Deprivation Environment in Spatial Epidemiologic Studies. Open J Stat, 6(3), 436-442. doi:10.4236/ojs.2016.63039

[9] Martín, A. C., & Barros, C. M. d. C. P. (2015). Designing a Living Conditions Index and Classification of the National Territory. Revista Cubana de Medicina General Integral, 31(3), 323-332.

[10] Nogueira, H. (2006). os Lugares e a saúde: uma abordagem da Geografia às variações em saúde na Área Metropolitana de Lisboa. (Doutoramento), Universidade de Coimbra.