NOVA IMS Information Management School

MMAA

Mestrado em Métodos Analíticos Avançados

Master Program in Advanced Analytics

REFORMULATING LISBON PARISHES

Data-Driven Redistricting Plan for Lisbon's Administrative Area: A Comprehensive Analysis Based on the Census Data

Tongjiuzhou Liu

Dissertation presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão da Informação

Universidade Nova de Lisboa

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

REFORMULATING LISBON PARISHES

by

Tongjiuzhou Liu

Dissertation presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics

Adviser: Mijail Juanovich Naranjo Zolotov

November, 2023

Reformulating Lisbon parishes

Data-Driven Redistricting Plan for Lisbon's Administrative Area: A Comprehensive Analysis Based on the Census Data

This document was created with the (pdf/Xe/Lua)LaTeX processor and the NOVAthesis template (v7.1.5) (Lourenço 2021).

Acknowledgements

We extend our deepest gratitude to all those who contributed to the success of this study. Our heartfelt thanks go to the members of the Lisbon Geographic Society for providing access to historical data and invaluable cartographic resources. We are also indebted to the Statistical Department of Lisbon for their assistance with the latest census data and for their ongoing commitment to data transparency.

The completion of this research was made possible through the dedicated guidance and insightful contributions of our esteemed Professor Alberto Acedo Sanchez. His expertise and scholarly advice have been instrumental in steering this study towards its fruitful culmination.

We extend our heartfelt gratitude to our adviser, Professor Mijail Juanovich Naranjo Zolotov, whose expertise and thoughtful mentorship have greatly enriched our analytical perspectives and academic endeavors.

Their combined support, both academically and intellectually, has not only enhanced our research but also fostered our growth as scholars in the field. We are deeply appreciative of their unwavering commitment and invaluable assistance throughout this academic journey.

This research was a collaborative effort, and its findings are a testament to the power of shared knowledge and collective endeavor.

”

“The path not fully taken by predecessors will be continued by successors.”

— Xu Xiake, Xu Xiake’s Travel Diaries
(Travel writer and geographer)

Abstract

The problem addressed in this study revolves around the inefficiency of current administrative regionalization methods in capturing the dynamic nature of urban population distributions. Traditional models often fail to accommodate the rapid changes and complexity inherent in urban expansion and demographic shifts. This issue not only poses significant challenges in urban planning and resource allocation but also has profound implications for socio-economic development and environmental management. The intricacies of urban development necessitate a solution capable of handling high-dimensional data and providing actionable insights that are crucial for informed decision-making and policy formulation. We help on this matter by a novel regionalization technique based on a data-driven approach. By utilizing Non-negative Matrix Factorization (NMF) and MaxP optimization, our method leverages the strengths of NMF in dealing with non-negativity data to decompose high-dimensional census datasets, while MaxP captures the spatial continuity and ensures statistically significant districts. The results from applying this method indicate a data-driven definition of urban districts that align more closely with actual population distributions and geographic continuities. The implications of this solution are far-reaching, offering a new paradigm in urban regionalization that can significantly influence urban policy and governance, facilitate resource optimization.

Keywords: Urban Regionalization, Urban Planning, Spatial Analysis, Data-driven methods

Resumo

O problema abordado neste estudo gira em torno da ineficiência dos métodos atuais de regionalização administrativa em capturar a natureza dinâmica das distribuições populacionais urbanas. Modelos tradicionais frequentemente falham em acomodar as rápidas mudanças e complexidades inerentes à expansão urbana e aos deslocamentos demográficos. Essa questão não só apresenta desafios significativos no planejamento urbano e na alocação de recursos, mas também tem implicações profundas para o desenvolvimento socioeconômico e a gestão ambiental. As complexidades do desenvolvimento urbano necessitam de uma solução capaz de lidar com dados de alta dimensão e fornecer insights acionáveis que são cruciais para a tomada de decisões informadas e formulação de políticas. Nós contribuímos para este assunto com uma técnica de regionalização inovadora baseada em uma abordagem orientada por dados. Ao utilizar a Fatoração de Matriz Não Negativa (NMF) e a otimização MaxP, nosso método aproveita as forças da NMF no tratamento de dados de não negatividade para decompor conjuntos de dados censitários de alta dimensão, enquanto o MaxP captura a continuidade espacial e garante distritos estatisticamente significativos. Os resultados da aplicação deste método indicam uma definição orientada por dados de distritos urbanos que se alinham mais estreitamente com as distribuições populacionais reais e as continuidades geográficas. As implicações desta solução são abrangentes, oferecendo um novo paradigma na regionalização urbana que pode influenciar significativamente a política e a governança urbanas, facilitando a otimização de recursos.

Palavras-chave: Regionalização Urbana, Planejamento Urbano, Análise Espacial, Métodos Baseados em Dados

Contents

List of Figures

List of Tables

1

Introduction

In recent years, significant changes are occurring in urban centers and internal structures due to shifting demographics and evolving urban dynamics. However, administrative boundaries have remained largely unchanged, highlighting the necessity for a reevaluation considering these transformations. The city of Lisbon is no exception, and in response to these challenges, the Portuguese government implemented administrative reforms in July 2012. The Parliament approved a proposal, including Chapter Two outlining the plan to reconfigure Lisbon's parish map. This chapter stipulated that the reconfiguration would be based on the principles of rationalization and adjustment of territorial organization, with the aim of creating larger and more balanced parishes. The original 53 parishes were subjected to modifications such as merging or extending, resulting in the current 24 parishes that came into effect in January 2013. The implementation of Lisbon's administrative reform was a complex and challenging process, but it resulted in a more modern and efficient administrative structure for the city. However, as Lisbon's population continues to diversify and urbanize, the current administrative framework may no longer be sufficient to meet the needs of the city's residents and future planning challenges. There is an urgent need for alternative methodologies to redefine and redistrict cities that can accomodate better the diverse needs of the city and plan that can adapt to evolving circumstances.

From the best of our knowledge, the Lisbon redistriction process effectuated in 2012 resulting into the current 24 parishes was based on trying to maintain similar number of dwellers for each parish and developed through extensive consultation and discussions to better meet the administrative and governance needs of Lisbon. This study aims to propose a redistricting plan for Lisbon parishes based on a data-driven approach, utilizing the 2011 census data to inform the decision-making process. By analysing key demographic variables, such as population density, age distribution, education, employment, housing, and migration patterns, we seek to identify trends, disparities, and opportunities that can inform the redistricting process. In doing so, we are proposing and parametrizing alternative methodologies to acquire more homogeneous and equitable parishes in Lisbon.

This study initiates with a detailed examination of Lisbon's current administrative structure, including an analysis of the 2011 census data to highlight patterns and trends essential for informing the proposed redistricting plan. Advanced methods, such as spatial cluster analysis from data science, regionalization studies, and other relevant tools, will be employed to identify specific challenges and opportunities in Lisbon. This approach could effectively define alternative goals and objectives for the redistricting process. We will develop a comprehensive methodology, analysing and commenting on the results, to propose recommendations for redefining administrative boundaries in cities. The aim is to resdistrict Lisbon administrative boundaries following a data-driven approach to potentailly encompass alternatives of reallocating resources, and implementing policies in new boundaries that may better meet citizens' needs. The paper is structured to provide a thorough analysis of previous studies in Chapter 2, outline the methodology in Chapter 3, detail the step-by-step process from cluster analysis to the final regional division plan in Chapter 4, discuss the results and the limitations of the research in Chapter 5, and conclude with future research suggestions in Chapter 6.

2

Literature review

In the classic paper on re-scaling cities(Friedmann and Wolff 1982), the following important points are expressed viz: The importance of understanding the changing territorial organization of cities and regions in the context of global capitalism, as this can impact the spatial distribution of economic activities, populations, and resources. The need for a comprehensive and nuanced approach to studying world cities that considers multiple factors, such as political economy, culture, and social processes. The potential for world cities to be sites of both inequality and innovation, where the concentration of economic and social activities can create both opportunities and challenges for different segments of the population.

2.1 Case studies of transformation in Regional Urban Systems

Regional Studies. In approaching the concept of the regional urban system, attention is initially drawn to the better-known types of economic region. The distinctive nature of the regional economy is next examined, and it is argued that its spatial structure represents an important dimension. Spatial structure can be characterized in a variety of ways, the most comprehensive of which employs the perspective of an urban system. This is examined firstly in terms of models from location theory, which provide important points of reference, and then within the setting of the present-day city-region (J. B. Parr 2014).

Following more than 25 years of animated debate among French elected officials and civil servants about territorial reform, the government have finally decided to reduce the number of regions from 22 to 13 in a process dubbed 'le big bang des régions' by the French media.The previous structure was notoriously complicated, with 36,700 communes, 2,600 inter-communal groupings, 101 departments and 22 regions, leading to waste, duplication, and a lack of transparency. The aim of the reform is to make the public sector in France more dynamic, responsive, and better adapted to the geography of the modern economy, while freeing up the regions to focus on economic development by devolving power to local levels on more everyday issues (Mcnally 2016).

In 1999, Poland underwent administrative reforms that reduced the number of top-level administrative districts from 49 to 16. The focus of the responsibilities of these districts was also restructured to prioritize overall regional development, higher education, regional infrastructure, and the management of EU funds. This resulted in the creation of 16 voivodeships, 308 counties, 66 county towns, and 2,478 municipalities. When considering changes to the territorial division, it is important to consider combining urban and rural areas in sparsely populated regions or areas with high concentrations of service infrastructure in central towns. Any proposed changes should be reviewed by experts and the public before implementation. Changes that have sufficient justification and social acceptance can be introduced gradually or through administrative pilots, which typically test more complex reforms in selected administrative units or in more administratively efficient regions (Kaczmarek 2016).

Denmark also underwent administrative reforms in early 2007. The Danish territorial reform aimed to create municipalities and regions that were financially and professionally sustainable. Before the reform, municipalities in some areas of the country had already started merging, with referendums leading to the formation of single municipalities on islands like Bornholm, Ærø, and Langeland. The structural agreement recommended a minimum size of 30,000 inhabitants for new municipalities, with a minimum of 20,000 inhabitants required for the formation of new municipalities or binding cooperation with neighboring municipalities. As a result, after the reform, the average municipality size increased from around 20,000 inhabitants to about 55,000 inhabitants, and the number of municipalities with less than 20,000 inhabitants decreased significantly. The regions were also created to have an improved professional and financial basis to perform their tasks in providing high-quality healthcare to their populations. The regions have a population of between approximately 0.6 million to 1.6 million inhabitants, significantly larger than the counties before the reform (*Evaluation of the Local Government Reform* 2013).

The network partitioning method of area division has been studied (Hamilton and Rae 2020). The author used a network partitioning algorithm called Combo to analyze commuting data from the 2011 UK Census. The goal was to improve the methodology of regional partitioning and produce practical results. By using this approach, the author was able to create 17 new regions for Scotland, which is a reduction from the existing 32 regions.

2.2 A brief commentary on the evolution of Urban Planning in Europe

The difference between a city and other human settlements lies not only in its relatively large size but also in its functions and special symbolic status, which may be granted by central authorities. The term can also refer to the streets and physical structures

of the urban area, as well as the group of people living there, and can be used in a general sense to denote urban as opposed to rural areas (Lynch 2008) . The history of urban planning can be traced back to some of the earliest known cities, particularly in the Indus Valley and Mesoamerican civilizations, who built their cities on grids and divided them into different zones apparently for different purposes. The influence of planning is ubiquitou in today's world and can be seen most clearly in the layout of planned communities, which are designed comprehensively prior to construction, often considering interrelated physical, economic and cultural systems (Smith 2002). Hippodamus of Miletus (498-408 B.C.), an ancient Greek architect and town planner, is widely recognized as the "father of European town planning." He is best known for his influential "Hippodamian Plan" (grid plan), which revolutionized the layout of cities (Glaeser 2011). During the Second French Empire, under the leadership of Napoleon III, Baron Georges-Eugene Haussmann redesigned the city of Paris into a more modern capital, featuring long, straight, and wide boulevards(Jordan 1992). In the second half of the 20th century, urban planners gradually shifted their focus towards individualism and diversity in the city center (Routley 2018). Urban planning guides the development of cities, suburbs, and rural areas in an organized manner by addressing questions related to how people will live, work, and entertain in specific regions (Caves 2005). Zoning is an urban planning method that divides land into areas with specific regulations for new development. The rules for each zone determine whether planning permission for a given development may be granted. The guidelines set for zoning can include the types of land use allowed, size and dimensions of lots, and the form and scale of buildings. This helps guide urban growth and development in a municipality or other tier of government (*Urban Stormwater Management in the United States* 2009).

Urban regionalization analysis through spatial data analysis already exists in several studies, such as the delineation of policy-relevant urban area boundaries in England through the analysis of spatial economic data (Coombes 2014). One study suggests that urban planning needs to be reconsidered, requiring a new approach to define the "scale" of spatial practices to keep up with the evolving territorial organization of global capitalism by the late 20th century (Brenner 1999). A theory of urban reform that was developed a hundred years ago: mobile data sets will provide the basis for defining the boundaries of urban areas (**1915**). This work has had a profound influence on urban zoning, for example, and later researchers have drawn a very general principle from it that most of the time it is more appropriate to define regional sets first before zoning within a region, and some urban areas are polycentric. (J. Parr 2005). Previous studies have also shown that New urban systems are emerging with increasingly polycentric geometries that challenge the traditional urban center model (Keil 1994). The degree of urbanization is a modern indicator that helps define the composition of a city: "a population of at least 50,000 inhabitants in continuous dense grid cells (> 1,500 inhabitants per square kilometer)" (DIJKSTRA et al. 2020).

In many European countries, cities, particularly the larger ones, are further subdivided into administrative units for more granular governance and management. In the UK, cities may be divided into boroughs or districts, which are further broken down into wards and, in some cases, parishes. France's cities, like Paris, are categorized into 'communes', and larger cities have 'arrondissements'. Germany's cities, known as 'Städte', can have subdivisions called 'Stadtbezirke', with smaller neighborhoods termed as 'Ortsteile'. Italian cities, or 'Comuni', have city sectors referred to as 'Quartieri or Circoscrizioni'. In Spain, cities called 'Municipios' have divisions known as 'Distritos'. Dutch cities, termed 'Gemeenten', might be segmented into 'Stadsdelen' or 'Wijken', while Poland's 'Miasta' can be divided into 'Dzielnice'. These categorizations are general, and exact divisions may vary based on the specific country or region.

2.3 Redistricting methods and its spatial structure

In a recent study on Spatial Redistricting (Biswas 2022), a range of models were explored and investigated. These models included the SPATIAL Method, which uses spatially aware search operators for contextually relevant solutions with a potential trade-off in computational demand due to spatial constraints' complexity. The Sampling-based Approach employs Flip-walk based Markov Chains for efficient sampling from a large solution space, though it may not always converge to the optimal solution. MCMC is suited for sampling from complex distributions but may be slow to converge. The Flip-based Proposal maintains district contiguity but can get trapped in local optima. Sampling-based Models—BAA, BCAA, and AIO—prioritize balance but may neglect other aspects like compactness. Exact Methods offer precise solutions adhering to population balance constraints but are computationally intensive for large-scale problems. These models collectively provide a comprehensive toolkit for addressing the multifaceted redistricting problem, balancing computational time, solution quality, and the complexity of constraints.

Simulated redistricting based on post-census boundary redrawing has been utilized in a study (McCartan et al. 2022) encompassing several meticulous steps and considerations. This includes a transparent simulation process incorporating state-specific redistricting criteria into alternative plan simulations. The Sequential Monte Carlo Algorithm generates plans adhering to contiguity and population equality hard constraints, along with soft constraints for compactness. Data assembly merges precinct-level shapefiles with demographic data. The simulation ensures convergence and stability of the alternative plan sample. Diagnostics confirm sample diversity and representativeness, including checks for population deviation and minority voting age population, ensuring compliance with legislative requirements and reflecting demographic considerations.

In a recent publication(Walker 2023), a thorough exploration of spatial analysis and modeling techniques using US Census data is provided. The book dives deep into spatial analysis, addressing key topics such as spatial overlays, joins, small area

time series analysis, and distance/proximity analysis. Additionally, it examines the dynamics of spatial neighborhoods, the role of spatial weights matrices, and the nuances of both global and local spatial autocorrelation. Shifting its focus to modeling, the book discusses a range of concepts including isolation and diversity indices, regression modeling using US Census data, spatial regression, and Geographically Weighted Regression (GWR). Concluding chapters offer insights into the classification and clustering of ACS data, supplemented by practical exercises. These chapters collectively serve as a comprehensive guide, demonstrating the application of advanced statistical methods and spatial analysis techniques in R, which are instrumental for demographic research and urban planning based on Census data.

The third section of "Geographic Data Science with Python" (J. Rey, Arribas-Bel, and Wolf 2023), a detailed explanation is provided on the practical application of clustering techniques and regionalization methods. Utilizing data from the American Community Survey (2017), the study explores clustering, which categorizes observations based on similarity, and regionalization, a clustering method with additional geographical constraints. These approaches aid in understanding the socio-economic characteristics of areas such as neighbourhoods in San Diego. The research employs clustering algorithms like k-means and Ward's hierarchical method. An exploration of the spatial distribution of clusters offers deeper insights into the socio-economic structure of the San Diego metropolitan area. Statistical analysis of the clusters reveals their characteristics and profiles. Geodemographic analysis, applied to the San Diego census tracts, visualizes the spatial distribution of socio-demographic traits through maps. This method compresses statistical variation into a single categorical dimension, facilitating easier visualization and interpretation. Regionalization methods impose spatial constraints on clusters, producing geographically coherent clusters. Different spatial weights, like Queen contiguity and k-nearest neighbours, influence the final regional structure. The study compares various regionalizations based on geographical coherence and feature fit. Measures like the isoperimetric quotient and Calinski-Harabasz score evaluate the compactness and fit of different solutions. Tools such as the Adjusted Rand Score and Mutual Information Score assess the similarity of labelings generated by different clustering algorithms. In summary, the study comprehensively and practically explores clustering and regionalization in geographic data science, emphasizing the importance of multivariate analysis, spatial constraints, and comparison of different methods for understanding and interpreting complex socio-economic data.

The application of the models discussed in "Geographic Data Science with Python" (J. Rey, Arribas-Bel, and Wolf 2023) presents several limitations when dealing with large-scale datasets, such as our Lisbon census data. Firstly, the clustering analysis algorithms employed, including k-means and hierarchical clustering, do not consider geographical spatial data. This leads to highly fragmented results, which, while not meaningless, are of limited utility for our purpose of regional re-planning. Secondly, the hierarchical clustering with spatial constraints, also referred to as the regionalization

model in the text, produces outcomes excessively dependent on spatial constraints. Such regionally constrained results may not accurately reflect the actual demographic and socio-economic distribution. Therefore, our research aims to draw upon the concepts and ideas from this book rather than directly applying the models it uses. This approach allows us to adapt and develop methodologies better suited to our data's specificities and the objectives of our study.

2.4 The case of Lisbon

The administrative zoning reform in Lisbon can be traced back to the 1880s. The Law of July 18, 1885, consisting of 230 articles, was meticulously crafted through extensive consultations and legislative processes. This law provided a comprehensive legal framework for the administrative division of the city of Lisbon. Recognizing the unique needs of this important city, the government and legislative institutions of that era realized the necessity for a specialized management model. Consequently, they developed this law, which divided Lisbon into four administrative boroughs.

The most recent administrative zoning reform occurred in 2012. Portuguese Law No. 56/2012, enacted on November 8, introduced a new administrative reform for the City of Lisbon. This reform led to the decentralization of power from the Lisbon City Hall to each of the 24 "Junta de Freguesia" (parish councils) that make up the municipality. The motivation behind this administrative reorganization stemmed from the need to modernize and adapt the governance model of the city of Lisbon. This need was further compounded by the fact that Lisbon serves as the national capital and hosts various national government institutions. Additionally, it was influenced by the size and administrative disparities among the existing parishes within the county. Between 2009 and 2012, the Lisbon City Council, with the crucial involvement of Seixas, actively prepared and approved an administrative reform. This reform addressed both vertical changes at the parish level and horizontal restructuring within municipal departments (Marcuse 2010). Thus, the reform plan was developed through extensive consultation and discussions to better meet the administrative and governance needs of Lisbon (Santos 2015; Seixas 2019a,b).

From this historical perspective, it's evident that most of Lisbon's administrative reforms have been developed based on negotiation and discussion, often in response to various objective factors. Therefore, we have decided to reformulate this reform to explore how a redistricting of Lisbon freguesias based on from a data-driven approach looks like and its advantages and cons.

3

Methodology

3.1 Overview of demographic data and information in Lisbon

This study focuses on developing a comprehensive redistricting plan for the Lisbon administrative area. This study employs a mixed-methods approach, combining quantitative and spatial data analysis techniques. In 2011, the Área Metropolitana de Lisboa had a population of 26.7% of the national total, with an annual growth rate of 0.6%. The population density was 940.0 individuals per km², and the index of aging was 117, indicating that there were more elderly people than young people. Most families consisted of two or more people, with a growing number of people living alone. The percentage of foreign residents was 7.2%, and the population with higher education was 19.6%, From Pordata (a database of certified statistics about Portugal).

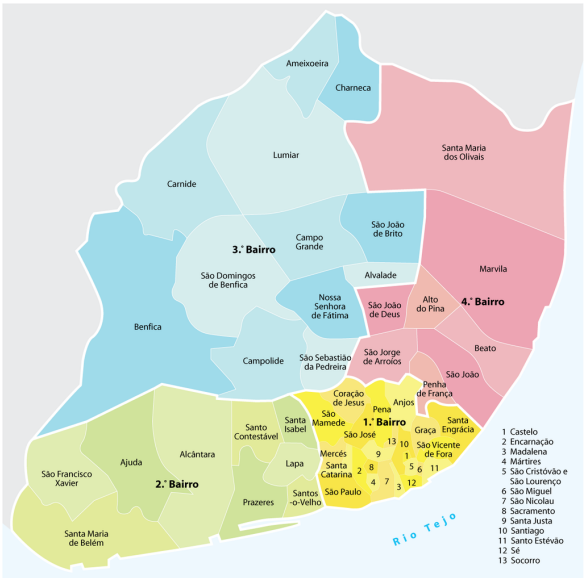
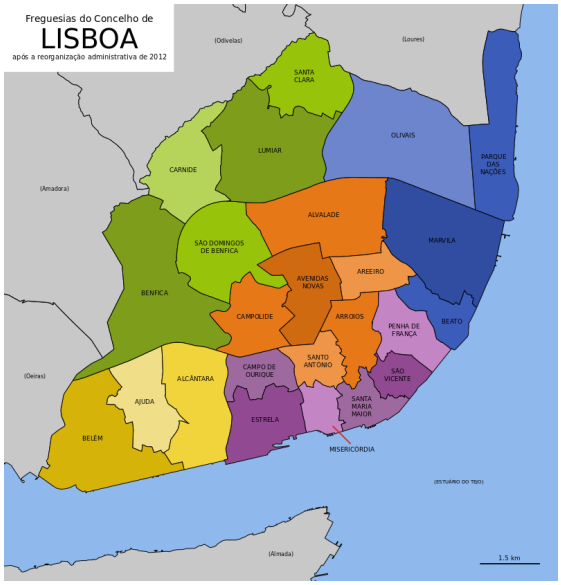


Figure 3.1: Lisbon freguesias distribution before 2012 (JPG)

In 2021, the population of the Área Metropolitana de Lisboa increased to 27.8%, with a lower annual growth rate of 0.2%. The population density also increased to 951.9 individuals per km². The index of aging also increased to 151, indicating a more

significant number of elderly people compared to young people. The percentage of foreign residents increased to 8.9%, and the population with higher education increased to 26.6%. The number of households living alone and using transportation colectivo also increased, From Pordata.



(JPG)

Figure 3.2: Lisbon freguesias distribution after 2012

3.2 Data Preparation

The primary data source for this research is the 2011 Portuguese National Census conducted by the Portuguese National Institute of Statistics (INE) in 2011. Census enumerators collected data from households and individuals across the country. The INE was responsible for organizing and implementing the census, ensuring the accuracy of data collection and analysis. The comprehensive results and data from the census were gradually released starting on July 29, 2011. According to the census, Lisbon's population in 2011 was 547,733, a decrease from 564,477 in 2001, with a population density of approximately 6,378 people per square kilometer. Lisbon's population was aging, with a higher proportion of elderly people and a median age of around 42.2 years. Compared to other regions in Portugal, Lisbon had smaller average household sizes, with 2.4 people per household. Lisbon had a relatively high number of vacant housing units, accounting for about 12.2% of the total housing stock. The educational level of Lisbon's population was higher than that of other regions in Portugal, with a higher proportion of individuals completing higher education. The unemployment rate was higher than the national average, but the city also had a higher proportion of individuals engaged in professional, scientific, and technical activities.

Before conducting any analysis, the raw census data will be preprocessed, cleaned, and transformed as needed to ensure its usability for the study. In this study, we used features derived from census data. Since all the data values are non-negative, starting from zero and upwards, we employed the Non-negative Matrix Factorization (NMF) technique for dimensionality reduction. NMF is particularly suitable for datasets with non-negative values as it factorizes the original data into two lower-dimensional matrices, ensuring that all values remain non-negative. This property of NMF makes it an ideal choice for preserving the inherent structures and patterns present in datasets like census data, where negative values wouldn't have a meaningful interpretation (Hedjam, Abderrahmane, and Cheriet 2022).

3.3 Redistricting Lisbon parishes using a data science approach

This section begins with an in-depth descriptive analysis of the 2011 census data, focusing on a selection of key variables across six categories to uncover significant spatial patterns within Lisbon. The section then progresses to explore the regionalization process. Here, we critically evaluate various techniques, ultimately selecting the Max-P Regionalization method for its adaptability, ability to maintain spatial contiguity, and proficiency in optimizing objective functions within specific constraints. This selected method, applied with multi-dimensional constraints that consider demographic, economic, and geographic factors, facilitates the formation of spatially contiguous and diverse regions. This approach underlines the potential of data science in urban planning and policy-making, adeptly capturing the multifaceted nature of urban spatial structures.

3.3.1 Processing Redistricting plan

The redistricting plan development will involve integrating the findings from the descriptive analysis and spatial cluster analysis to propose new administrative boundaries. Here is the flow chart to process the redistricting plan.

3.3.2 Descriptive Analysis

First, we conducted a preliminary descriptive analysis of the 2011 census data. This data includes 134 variables, consisting of one geographical information variable, nine qualitative variables such as administrative codes, names, and time, and 124 quantitative variables. These 124 quantitative variables are divided into six main categories: buildings('EDIFICIOS'), individuals('INDIVIDUOS'), families('FAMILIAS'), residences('ALOJAMENTOS'), education('EDUCAÇÕES'), and employees('EMPREGADOS'). Given the complexity and multitude of quantitative variables, we plan to select the most representative variable from each of these six main categories for visual analysis.

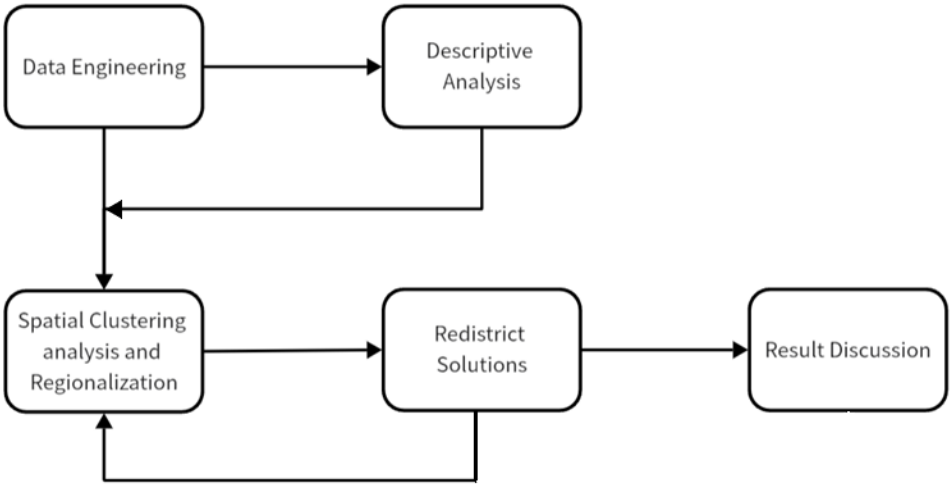


Figure 3.3: Processing plan

This figure 3.4 provides a wealth of information. Firstly, the light to dark color gradient represents increasing data amounts, and different colors indicate different categorical variables. In terms of buildings, the city center and areas near the river have relatively higher numbers of buildings, while the west to north parts of the city generally have fewer buildings. Notably, there's a sparse area right in the city center. In terms of individuals, representing both residents and non-residents, populous areas are still centered around the city center, with significantly populated areas also in the northwest and north. The distribution of families and residences mostly aligns with population density. Regarding education, indicated by the number of individuals with a college degree, there are nearly vacant areas in the north, southwest, and east. Lastly, the distribution of employees appears to be closely related to population density.

3.3.3 Regionalization Process

In the realm of spatial analysis, regionalization techniques offer vital tools for clustering and delineating territories based on specific attributes or criteria. One crucial decision faced by researchers in this domain is selecting the optimal method for a given study or application. In our exploration, we critically examined multiple methods, including the hierarchical clustering technique with varying spatial weights. This method involves forming a hierarchy of clusters based on certain criteria, with the flexibility of incorporating different spatial weights to emphasize or diminish the role of proximity in the clustering process. Another method we evaluated was the Spatial 'K'luster Analysis by Tree Edge Removal (SKATER). The SKATER method focuses on the partitioning of geographical space using a tree-based approach and seeks to identify areas that are homogeneous with respect to specified criteria. After a comprehensive comparison of these methodologies, we concluded that the Max-P Regionalization technique is

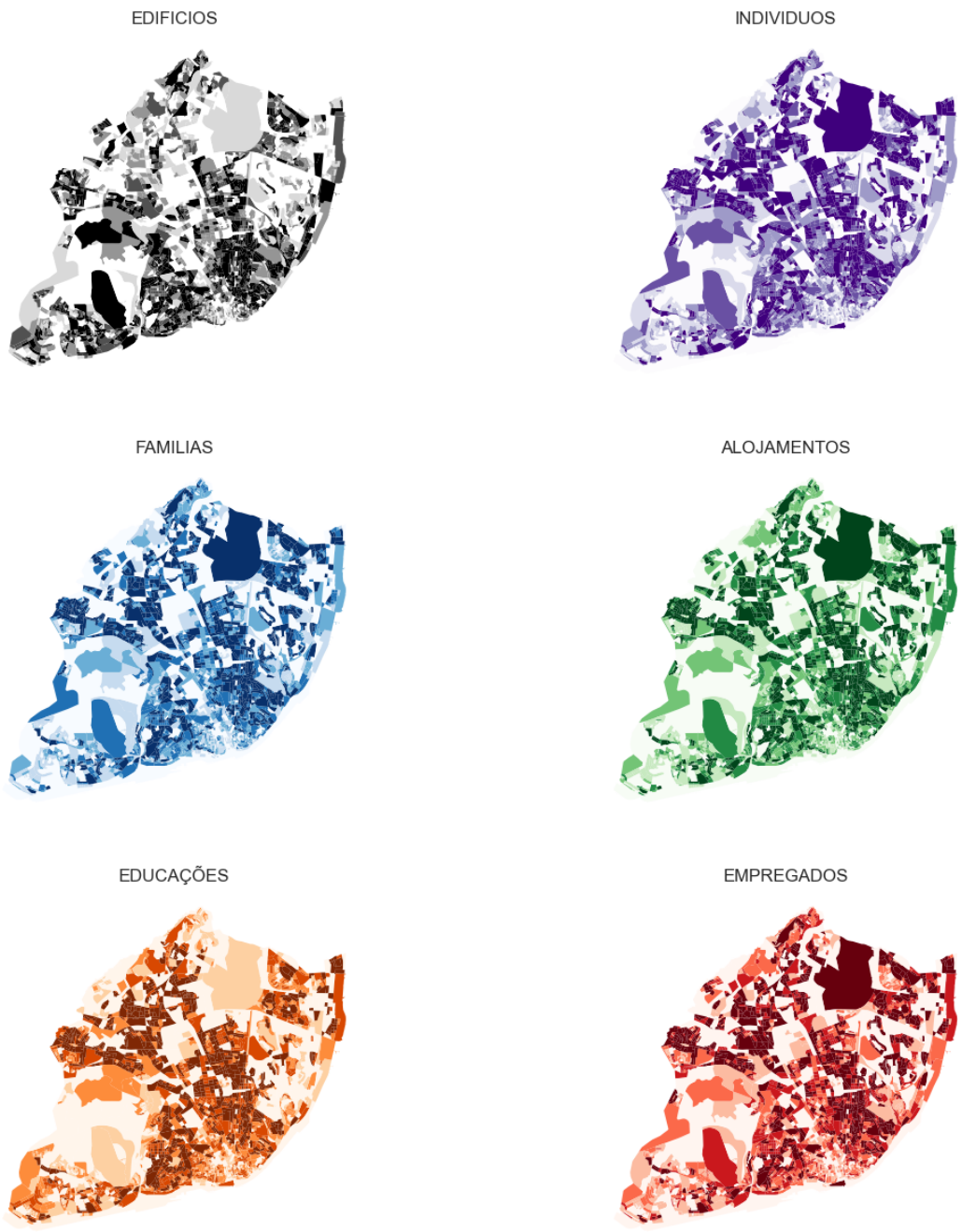


Figure 3.4: Main variables spatial visualization. The light to dark color gradient represents increasing data amounts, and different colors indicate different categorical variables

superior for our specific objectives due to the following reasons:

Flexibility and Generalization: The Max-P regionalization technique does not require a predetermined number of regions, making this method more adaptable and data-oriented than other techniques. This makes our data even more insightful in the study.

Spatial Contiguity: The one of most requested of our results is the technique ensures that the resulting regions are spatially connected, yielding a more intuitive and geographically coherent clustering. In our research, this is crucial to obtaining connecting areas for creating new parishes.

Optimization of Objective Function: Max-P regionalization maximizes a given objective function, ensuring that the resulting districts match the main goal of the study. We can utilize this property to continually modify the parameters to ultimately generate 24 districts.

Efficient Handling of Constraints: It efficiently incorporates constraints, such as minimum population thresholds or specific attribute sums, ensuring that districts are both statistically and practically meaningful.This allows us to produce different regionalized results by setting different thresholds for getting the optimal parishes .

In our study, we leveraged the distinctive attributes of the Max-P Regionalization technique, applying multi-dimensional constraints to the data. Specifically, we established varying categories of thresholds, encompassing demographic, economic, and geographic parameters, to ensure a robust and comprehensive regionalization process that aligns with the nuanced complexities of urban spatial structures. This approach allowed us to account for diverse urban dynamics and to tailor the regional clusters to reflect the multifaceted nature of the areas under study. Max-P Regionalization is a cutting-edge approach designed for the delineation of spatially contiguous and heterogeneous districts. Its primary aim is to aggregate various geographical zones into as many homogeneous districts as possible by maximizing a specific objective function. This is subject to constraints, such as the minimum number of observations or the minimum sum of a particular attribute within each district. One significant advantage is its ability to balance the number of districts with their internal homogeneity, offering a detailed representation of the geographical space being analyzed. Unlike methods that require a predetermined number of districts, the count of districts in the max-p problem emerges naturally from the data. This feature allows us to obtain the number of districts more accurately by constantly adjusting parameters. Initially conceptualized as a mixed-integer problem by (Duque, Anselin, and S. J. Rey 2012) , the max-p is recognized as an NP-hard challenge. Precise solutions are typically practical only for smaller datasets. Due to this complexity, numerous heuristic methods have been developed to tackle the max-p problem, including the strategy detailed in (Wei, S. Rey, and Knaap 2021) study, which has been incorporated into PySAL.

3.3.4 Analysis and Comparison of the Obtained Results

To preface the detailed analysis and comparison of the obtained results, this section initially outlines the methodology and principles behind the statistical measures used. The measures, including the Isoperimetric Quotient (IPQ) and the Jaccard Index, are essential tools in assessing geographical coherence and spatial similarity, respectively. They enable a quantitative evaluation of the compactness of geographical shapes and the overlap between different spatial sets. Additionally, the chapter discusses the Intersection and Symmetric Difference methods, which are instrumental in visualizing and quantifying the interaction or divergence between geographic features. The following sections delve into each of these measures in detail, elucidating their calculation principles and applications in the context of redistricting and spatial analysis. This foundational understanding sets the stage for a comprehensive comparison of the newly generated area with the original area, providing insights into the effectiveness of redistricting efforts and the spatial characteristics of the districts under study.

3.3.4.1 Isoperimetric quotient (IPQ)

Isoperimetric quotient (IPQ) is the most common of measure of geographical coherence involves the "compactness" of a given shape. This compares the area of the district to the area of a circle with the same perimeter as the district. To obtain the statistic, we can recognize that the circumference of the circle is the same as the perimeter of the district i, so $\Pi = 2\pi r$. Then, the area of the isoperimetric circle is $A_c = \pi r^2 = \pi \left(\frac{\Pi}{2\pi}\right)^2$. Simplifying, we get: $IPQ = \frac{A_i}{A_c} = \frac{4\pi A_i}{\Pi^2}$ For this measure, more compact shapes have an IPQ closer to 1, whereas very elongated or spindly shapes will have IPQs closer to zero. For the clustering solutions, we would expect the IPQ to be very small indeed, since the perimeter of a cluster/district gets smaller the more boundaries that members share. Computing this, then, can be done directly from the area and perimeter of a district.

The IPQ is commonly used in political science and geography to study the relationship between the shape of a political unit and its political, economic, or social characteristics. For example, some researchers have found that states with higher IPQs tend to have more efficient and effective governments, while states with lower IQs tend to have more political instability and economic inequality.

3.3.4.2 Jaccard Index

Comparison of Visualized Results Using the Jaccard Index and its Principle.The Jaccard Index, also known as the Jaccard similarity coefficient, is a statistical measure employed to gauge the similarity between two sample sets. The formula for its computation is:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Where A and B are two sample sets. In geographical imagery or Geographic Information System (GIS) data, these sets might denote specific districts of geographical space, like

land use types, land cover, or other geographic features. The value of the Jaccard Index lies between 0 and 1. A value of 1 signifies complete overlap of the two spatial sets, while 0 indicates no overlap whatsoever. For instance, consider two land use images, A and B. If one wishes to discern the similarity in land use types between these images, the Jaccard Index can be employed. The overlap and union of land-use pixels in A and B provide the Jaccard Index value. In essence, the Jaccard Index serves as an effective tool for quantifying and comparing similarity between geographical images or spatial datasets.

Comparison of the Newly Generated Area with the Original Area Using the Jaccard Index and its Principle. For two polygons A and B, the Jaccard Index J is defined as:

$$J(A, B) = \frac{\text{Area of Union}(A, B)}{\text{Area of Intersection}(A, B)}$$

The numerator represents the area of overlap between the two polygons. The denominator is the combined area of the two polygons minus their overlapping district (to avoid double-counting). A Jaccard index nearing 1 indicates high similarity between the polygons, while a value nearing 0 indicates low similarity. For example, the area has a high Jaccard index, suggesting a significant overlap between the old and new delineations.

3.3.4.3 Intersection and Symmetric Difference

Visualization of Intersection and Symmetric Difference and its Principle. Intersection Area: Represents the overlapping district between the old and new delineations. Symmetric Difference Area: Represents the areas unique to each polygon, i.e., areas without overlap. In Geographic Information Systems (GIS), these tools are routinely used to ascertain how multiple geographic features interact or diverge.

Calculation Principle of Intersection and Symmetric Difference Area and Their Proportions. To delve deeper into the similarities or differences between two districts, we can determine the ratio of their Intersection Area to the Symmetric Difference Area. The ratio is defined as:

$$Ratio = \frac{\text{Intersection Area}}{\text{Symmetric Difference Area}}$$

A ratio exceeding 1 signifies that the area of overlap (intersection) is more extensive than the area of non-overlap (symmetric difference), whereas a ratio under 1 indicates the opposite.

4

Results

This chapter presents results from an analysis utilizing Non-negative Matrix Factorization (NMF) and Max-P Regionalization techniques on 2011 census data. Initially, an optimal number of NMF components is determined using the elbow method, identifying six as ideal. These components are then visualized to show their geographical distribution across Lisbon, highlighting distinct patterns. The Max-P Regionalization approach is employed to cluster sub-sections into districts, initially yielding 306 districts. Adjustments in the model parameters eventually achieve a desired outcome of 24 districts. The study systematically explores and visualizes data across six variable categories - Buildings, Dwellings, Individuals, Households, Education, and Employment. Each category's influence on regionalization is analyzed separately, providing insights into urban dynamics and form.

4.1 Dimensionality reduction: Non-negative Matrix Factorization (NMF)

4.1.1 Numbers of components selection

Firstly, we developed a program to determine the optimal number of components for the NMF model. The program developed identifies the optimal number of components for the Non-negative Matrix Factorization (NMF) model. It consists of two functions: the first calculates reconstruction errors for each component from 1 to a specified maximum using NMF to factorize and then reconstruct the data, measuring error with the Frobenius norm. The second function identifies the elbow point in the reconstruction error curve, crucial for determining the optimal number of NMF components. The script computes errors for up to 20 components, locates the elbow point, and plots the reconstruction errors against the number of components, highlighting the elbow point with a red dashed line. This visualization aids in balancing data representation and complexity in choosing the most suitable number of components for NMF.

From the visualization Figure 4.1, we can determine that the optimal number of components is 6. This determination is based on the identification of the elbow point

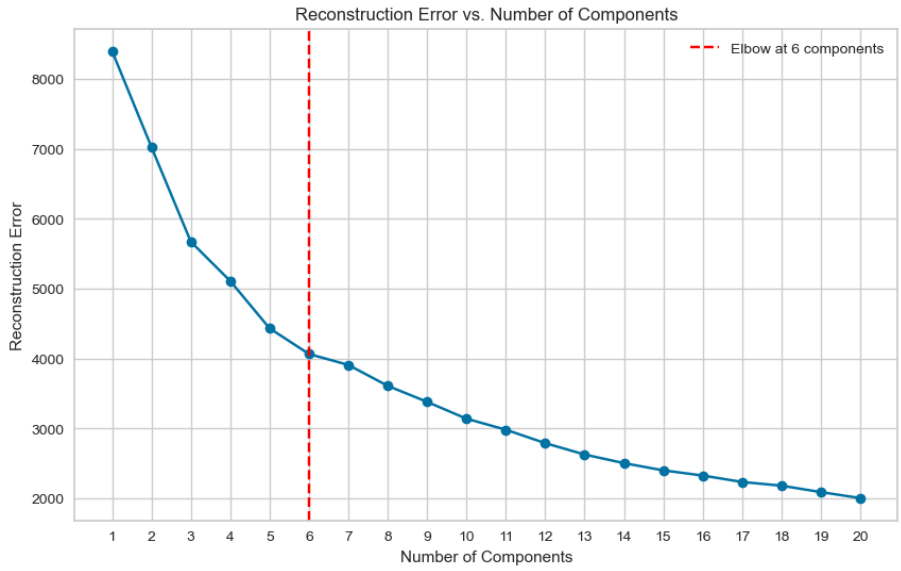


Figure 4.1: NMF component number selection for dimensionality reduction of 2011 census data

in the reconstruction error curve by the second function. Next, we will transform the original variables using NMF, resulting in a matrix composed of these 6 components. These 6 components can replace the original dataset and variables in subsequent analyses.

4.1.2 NMF components visualization

This figure 4.2 represents the distribution of the first component obtained after applying Non-negative Matrix Factorization (NMF) with a reduction to six components. The distribution visualizations of other five components are in the appendix A. The map is color-coded using a quintile classification, which divides the data into five equal parts, ensuring an even distribution of data points within each category. This type of classification is useful for identifying patterns in the data that are not immediately obvious and can be particularly effective in highlighting areas of high and low values.

Looking at the map, the shades of color vary from light to dark, indicating the intensity or value of the first component in each area: The lightest color, which likely represents the lowest values (ranging from 0.000 to 0.723), covers the largest area. These districts may be characterized by lower activity or presence of whatever factor the first component represents. The intermediate colors (ranging from 0.723 to 5.942) cover smaller, more scattered areas. These could represent transitional zones or areas with moderate levels of the factor in question. The darkest color (ranging from 5.942 to 34.630) highlights the areas with the highest values of the first component. These areas are significantly fewer and more concentrated, suggesting districts with the highest intensity of the factor measured by this component.

The spatial distribution suggests that the factor or features captured by the first

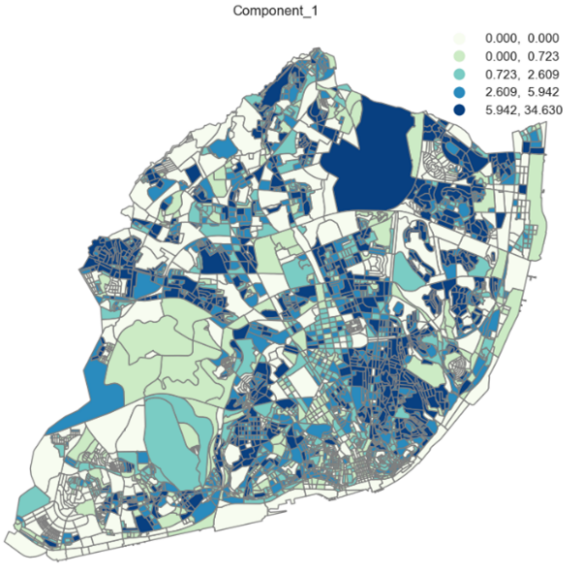


Figure 4.2: Quintile Distribution Visualization of the First NMF Component. The figure showcases a quintile-based color gradation of the first NMF component across data, revealing the most significant areas of activity in the darkest shades.

component are most prevalent in these darkly shaded areas. These could be areas of specific interest, such as high-density residential districts, commercial hotspots, or other significant urban features, depending on the nature of the data used in the NMF model. The quintile approach ensures that outliers or extreme values don't skew the visualization, allowing for a balanced representation across the spectrum of values.

4.2 Max-P Regionalization: Redistricting Lisbon Parishes

To develop our holistic view, we can consider the six components as a multi-dimensional array and aim to cluster the 3623 sub-sections into the maximum number of districts such that each district contains at least $150 \approx \frac{3623}{24}$ sub-sections. We first define variables in the data frame to measure regional homogeneity. Next, we specify numerous parameters to input into the max-p model: a spatial weight object denoting the spatial connectivity of the sub-sections, the minimum number of sub-sections each district must have (threshold), and the number of top candidate districts to consider when allocating enclaves (top_n). We create an attribute 'count' that will serve as the threshold attribute added to the data frame. We can then instantiate the model, solve it, and obtain visualization results on Figure 4.3.

By continuously adjusting and experimenting with the parameters, we eventually achieved the desired regionalization outcome of 24 districts. Figure 4.4 shows the result of the 24 districts. This Figure was generated using the max-p model with a threshold set to the count of subsections, utilizing six components derived from NMF dimensionality reduction. The spatial weights are based on the queen contiguity of the

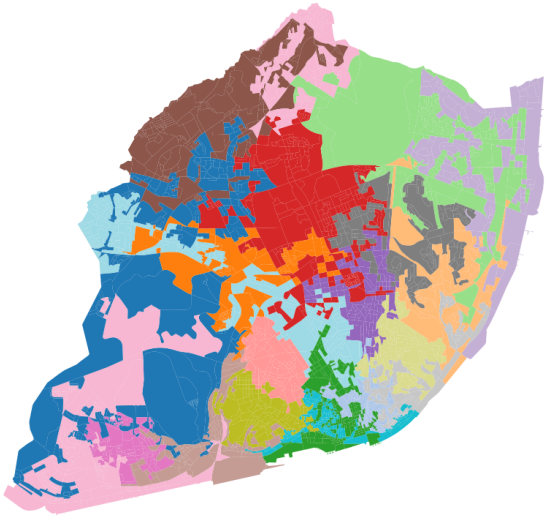


Figure 4.3: Max-P initialization visualization results

original data, and the Varying State Enclave Assignment was set to 5. Afterward, we matched the Freguesias from the 2011 census data with their corresponding current Freguesias name, effectively updating our records to reflect the current nomenclature. Subsequently, we applied the Max-P Regionalization result to assign new labels to all subsections. In the end, These new labels were named based on the highest proportion of present-day Freguesia names within each new label, ensuring that the Max-P-derived classification resonates with the contemporary administrative divisions.

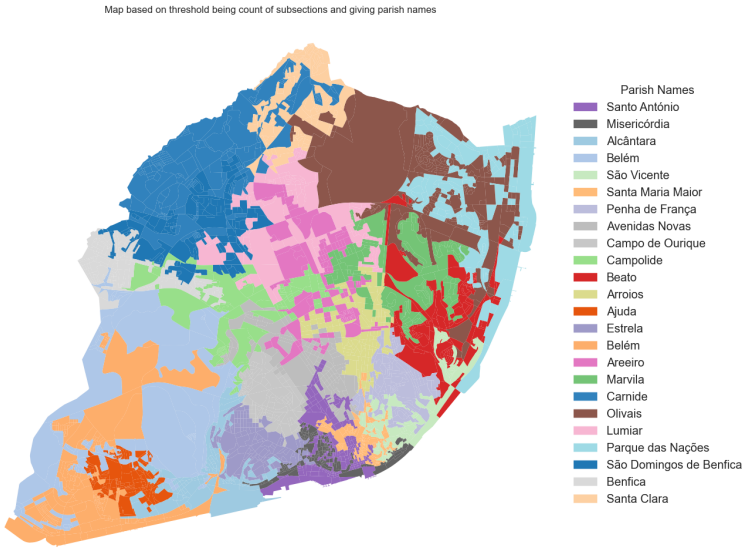
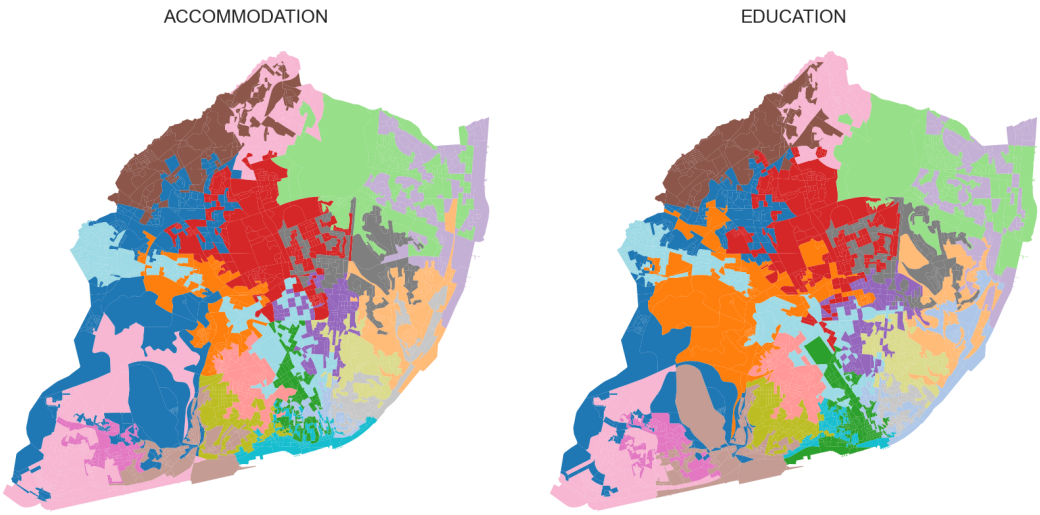


Figure 4.4: Map based on the six components with the threshold being count of subsections and giving parish names

The original dataset comprised variables that were broadly categorized into six

distinct types: Buildings, Dwellings, Individuals, Households, Education, and Employment. Considering this categorization, we resolved to operate our Max-P model separately for each variable type, maintaining a consistent threshold equal to the number of subsections. This deliberate and methodical approach enabled us to generate six distinct outcomes, each reflective of the spatial distribution and regional characteristics pertaining to the respective variable type. Figure B.1 shows two examples of the classification data. The others are in the appendix B.



Results based on accommodation data

Results based on education data

Figure 4.5: Max-P with classification data for the case of accommodation and education

Upon analyzing the outcomes, we observed that while the segmentation of areas on each map differed, the overarching trends were markedly consistent. This observation prompted an investigation into the influence of primary variables from six distinct categories on these recurring patterns. Employing the Max-P model, which is versatile with varying data sets, we delved into the influence of principal variables from categories such as heritage buildings, households, accommodation facilities, employed populace, and educated populace. An individual analysis of each central variable was conducted to discern its impact on the regional demarcation process. Initially, we continued with the Max-P model, integrating six components extrapolated from NMF for dimensional reduction. The spatial weighting was determined by the queen contiguity from the original dataset. However, we varied the threshold across different variables: the count of buildings, families, accommodation units, current residents, employed individuals, and educated individuals. This methodology illuminated the unique contribution of each element in sculpting the urban terrain, capitalizing on the Max-P model's adaptability for an all-encompassing examination. Subsequently, we delineated the urban expanse, categorizing it based on these predominant features, and each segment was labeled appropriately. The figure 4.6 set the threshold at the

number of families, with additional details provided in the appendix C.

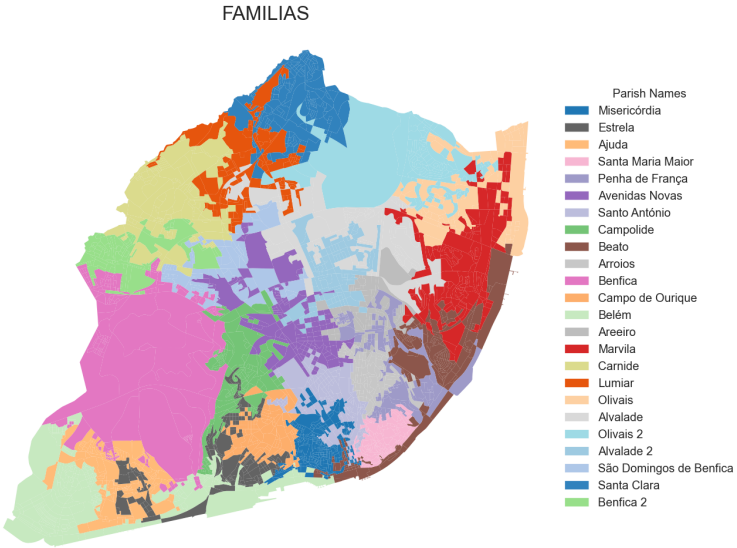


Figure 4.6: Result based on the six components with the threshold being the number of families and giving parish names

This figure 4.6 presents a color-coded visualization with the threshold variable being the count of families. Each color represents one of the parishes, as listed in the legend to the right. The variation in color intensity across different parishes suggests the count of families differs from one parish to another. Analyzing the map, we can observe: Parishes such as Misericórdia, Estrela, and Ajuda, indicated by distinct colors, stand out, suggesting unique family count characteristics that set them apart from other parishes.There is a diversity of family counts across the city, as no single color dominates the map. This diversity implies a heterogeneous distribution of family sizes or family-dominated households throughout Lisbon.The spatial arrangement of colors shows that neighboring parishes can have vastly different family count profiles. For instance, areas with darker shades may indicate higher family counts, whereas lighter shades suggest fewer families.

4.2.1 Assessment of Parishes' Compactness and Spatial Organization

In our analytical process, we computed the Isoperimetric quotient (IPQ) for all generated results to quantitatively assess the compactness of result's delineations. Subsequently, we aggregated these IPQ values to derive their respective sums, means, and medians. These aggregated metrics were systematically catalogued across several categories, each demarcated by a unique threshold criterion, and were designated as follows: `'New_label'` for the threshold of counts of subsections, `'New_Edificio_label'` for the number of buildings, `'New_Individuo_label'` for the count of individuals, `'New_Familia_label'` for the number of families, `'New_Alojamento_label'` for the

total accommodations, `'New_Education_label'` for the population with higher education, and `'New_Employee_label'` for the employed residents. Each category represents a distinct demographic or structural characteristic of the urban landscape, providing a comprehensive insight into the spatial organization of the study area. The results are in the table 4.1.

Table 4.1: Aggregated IPQ Metrics

Labels name	Sum	Mean	Median
New_label	1.019065	0.042461	0.030513
New_Edificio_label	0.875162	0.036465	0.027078
New_Individuo_label	0.916508	0.038188	0.034583
New_Familia_label	1.142353	0.047598	0.039392
New_Alojamento_label	1.106191	0.046091	0.032161
New_Education_label	0.898992	0.037458	0.035358
New_Employee_label	1.367089	0.056962	0.037122

The table 4.1 shows that districts demarcated by thresholds of family count and employment status exhibit a greater degree of compactness, suggesting an association with more centralized or clustered spatial arrangements. Conversely, thresholds based on the number of buildings and individuals correlate with less compact regional shapes, potentially reflecting a more dispersed spatial distribution of these elements. Furthermore, the generally low levels of the IPQ indicators across all categories illuminate a tendency toward fragmentation within the regionalization results.

4.2.2 Assessing Parishes' Congruence in Lisbon's Parishes: A Jaccard Index-Based Approach

4.2.2.1 A general overview based on Jaccard Index

Initially, we computed the Jaccard similarity coefficients to assess the geographical concordance between the outcomes of our study and the existing parish plans from 2011 (FR11) as well as the post-reform parish plans of 2012 (FR21). The results are documented in the subsequent table, where 'FR11' denotes the parish planning of 2011, and 'FR21' signifies the restructured parish planning following the 2012 reform. Furthermore, the outcomes were classified according to varying thresholds: 'New_label' corresponds to the count of subsections, 'New_Edi' pertains to the threshold of buildings, 'New_Ind' relates to the individual count, 'New_Fam' designates the family count, 'New_Alo' denotes the accommodation count, 'New_Edu' signifies the count of residents with higher education, and 'New_Emp' refers to the count of employed residents (Table 4.2).

Table 4.2: Jaccard Similarity Coefficients

index	FR11	FR21	New_label	New_Edi	New_Ind	New_Fam	New_Alo	New_Edu	New_Emp
FR11	–	0.965	0.901	0.909	0.891	0.917	0.907	0.920	0.908
FR21	0.965	–	0.911	0.909	0.901	0.934	0.917	0.921	0.928
New_label	0.901	0.911	–	0.923	0.912	0.926	0.906	0.915	0.911
New_Edi	0.909	0.909	0.923	–	0.900	0.935	0.890	0.937	0.915
New_Ind	0.891	0.901	0.912	0.900	–	0.918	0.929	0.927	0.938
New_Fam	0.917	0.934	0.926	0.935	0.918	–	0.895	0.938	0.939
New_Alo	0.907	0.917	0.906	0.890	0.929	0.895	–	0.896	0.913
New_Edu	0.920	0.921	0.915	0.937	0.927	0.938	0.896	–	0.948
New_Emp	0.908	0.928	0.911	0.915	0.938	0.939	0.913	0.948	–

High Similarity Between Years: The Jaccard similarity between 'FR11' and 'FR21' is very high (0.965), indicating that the administrative divisions before and after the reform in 2012 are quite similar. This suggests that while reforms were implemented, the overall spatial configuration of Lisbon's districts maintained a high degree of continuity.

Comparison with New Classifications: When comparing the 2011 divisions ('FR11') with new classifications ('New_label', 'New_Edi', etc.), the similarity scores are generally high (ranging from 0.891 to 0.920), implying that the new classifications retain a significant resemblance to the 2011 divisions. The lowest similarity score in this comparison is with 'New_Ind', suggesting that individual-based classifications diverge the most from the 2011 divisions.

Among New Classifications: The similarity scores among the new classifications are also high, with 'New_Fam' and 'New_Edu' showing particularly high similarity to 'FR21' (0.934 and 0.921, respectively). This could mean that family and education-based classifications are more aligned with the post-reform divisions of 2012.

Highest and Lowest Similarities: The highest similarity within the new classifications is between 'New_Fam' and 'New_Edi' (0.937), indicating that family count and educational attainment are closely related in terms of spatial distribution. Conversely, the lowest similarity within the new classifications is between, 'New_Edi' and 'New_Alo' (0.890), suggesting that these categories differ more in their spatial distributions.

Employment-Based Classification: The 'New_Emp' has high similarity scores with both pre-reform and post-reform divisions, with the highest score being with the 'FR21' (0.928). This indicates that employment-based classifications are closely aligned with the current administrative structure.

4.2.2.2 A particular overview based on Jaccard Index

In the process of generating names for the newly resulted parishes, not all original parish names were retained across the new results. This phenomenon can be attributed to our computational methodology, wherein we named the new labels based on the most frequent original parish name within each label's subsections.To streamline the

analysis and derive efficient results, we strategically selected four parishes – 'Olivais', 'Santa Maria Maior', 'Belém', and 'Benfica' – deemed to be the most representative for the purpose of our study. We will calculate the Jaccard index for these parishes to quantify their spatial congruence across each newly delineated parish division in comparison to the current existing parish divisions. The results are in the Table 4.3.

Table 4.3: Transposed Jaccard Similarity Coefficients for Various Parishes

Metric	Olivais	Santa Maria Maior	Belém	Benfica
New Parish	0.513104	0.285778	0.372327	0.188043
New Edificio Parish	0.609581	0.209859	0.708507	0.208663
New Individuo Parish	0.627708	0.131105	0.423643	0.301520
New Familia Parish	0.693732	0.202210	0.537416	0.499737
New Alojamento Parish	0.230351	0.483995	0.387098	0.201364
New Education Parish	0.253499	0.184237	0.431263	0.405565
New Employee Parish	0.292082	0.151855	0.675245	0.524663

Olivais consistently shows moderate to high similarities in new divisions, with the strongest overlap in 'New_Familia_Parish' (0.6937), highlighting family distribution's importance. Santa Maria Maior's lower Jaccard indices suggest more variation from current divisions, with housing factors being influential as seen in its highest similarity with 'New_Alojamento_Parish' (0.4840). Belém's high match with 'New_Edificio_Parish' (0.7085) indicates building distribution's significant impact. In contrast, Benfica displays the least similarity, particularly with 'New Parish' (0.1880), suggesting substantial regional boundary changes.

Higher Jaccard indices in 'New_Edificio_Parish' and 'New_Familia_Parish' across parishes imply that buildings and family counts align closely with existing divisions, indicating stable regional traits. Moderate similarities in 'New_Individuo_Parish' and 'New_Alojamento_Parish', and lower scores in 'New_Education_Parish' and 'New_Employee_Parish', suggest significant shifts or different distributions in these variables.

4.2.3 Visualization of Intersection and Symmetric Difference

To conduct an in-depth analysis, we meticulously compare the district of Olivais, a significant parish, across various new planning results with its current distribution, focusing on intersections and differences. This involves using visualizations to compare Olivais' configuration within different divisions, such as those based on buildings, individuals, families, accommodations, education, and employment.

In each map, Olivais is distinctly outlined against various division criteria, with the 'Current Parish' overlaid in sky blue for consistency, and 'New Parish' criteria in light coral for clarity in spatial changes. This approach allows for a nuanced comparative analysis of Olivais across different planning scenarios. Figure 4.7 presents these results,

including an overlay method from geopandas to highlight intersected areas in purple, emphasizing where current and specific divisions coincide.

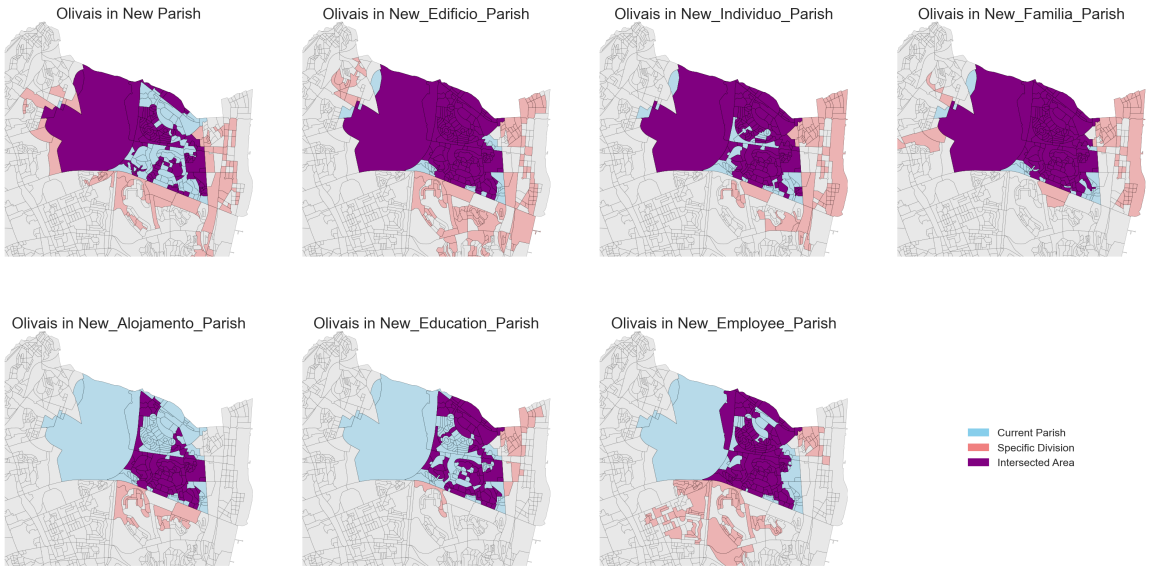


Figure 4.7: Comparison of Visualized Results of Olivais

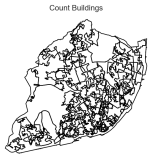
4.3 GeoSpatial Insights: Visualizing Lisbon's parishes with all regionalization results

To delve deeper than mere visual comparisons in analyzing regionalization plans, we utilize the Jaccard index, as detailed in Section 4.2. For each geographical unit in our dataset, we compute its Jaccard similarity across all combinations of planning schemes, and then average these values to derive each unit's Jaccard index. This includes comparisons with Lisbon's current parish distribution. The methodology encompasses classifying each unit under all plans, calculating Jaccard indices for each plan pair, and averaging these to obtain a comprehensive index for each unit. These indices, once computed for all units, are integrated into the dataset, offering a quantitative measure of classification similarities and differences under various regionalization plans. Figure 4.8a represents current parish boundaries, the other figuers 4.8b, 4.8c, 4.8d, 4.8e, 4.8f, 4.8g and 4.8i represents the new boundaries of our results

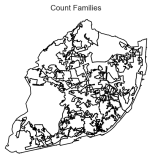
Prior to visualization, we normalize each unit's Jaccard index within a 0-1 range, enhancing the effectiveness of color mapping in Figure 4.8k. A score of 1 indicates high similarity across all plans, while 0 denotes no similarity. The resulting map portrays colored polygons against Lisbon's 24 parishes, focusing on aggregated boundaries formed through a 'dissolving' process that merges polygons sharing common parish values. This multi-layered map not only reveals the spatial distribution and intensity of these categorizations but also accentuates the broader groupings defined by the 24 parish

variable, providing a layered perspective on urban spatial organization.represents the new boundaries of our results

(a) Current Parishes Boundary



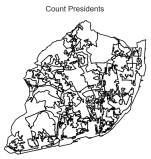
(b) Count Buildings Boundary



(c) Count Families Boundary

(d) Count Families Boundary

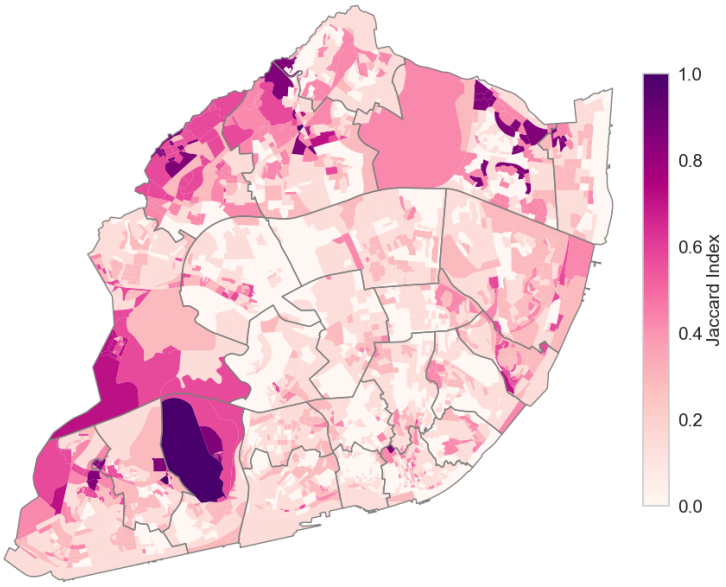
(e) Count Presidents Boundary



(f) Count Presidents Boundary

(g) Count Educated Residents Boundary

(i) Count Employed Residents Boundary



(k) Jaccard Index for all disturbiutions

Figure 4.8: Visualizing Lisbon's parishes with all regionalization results. The colors of the polygons are calculated as the average of the Jaccard indices for each unit across seven regionalization plans and current parishes distribution.

5

Discussion and Limitations

5.1 Discussion

The essence of a neighborhood is relationality and diversity (Shelton and Poorthuis 2019).The research conducted provides a comprehensive analysis of several regionalization schemes, juxtaposed with Lisbon's existing parish distribution. This juxtaposition reveals certain visual parallels; however, a deep dive into the data-driven analysis suggests a stark contrast in the actual similarities between these districts. The minimal similarity observed underscores the critical role played by the Max-P model in regional delineation, highlighting its efficacy in creating distinct and unique regional classifications. The model's application reveals significant insights, particularly in districts demarcated by family count and employment thresholds. These areas exhibit a notable degree of compactness, suggesting a correlation with more concentrated or densely arranged urban spaces. This finding points to the possibility of these areas being central hubs or densely populated urban locales.

In our analysis of regionalization outcomes, a salient complication emerged, namely the recurrence of parish names in numerous instances of the results. Take, for example, the regionalization framed by family counts as a threshold: the parishes named Olivais, Alvalade, and Benfica are each mentioned three times. This repetition is a consequence of our methodology for designating labels, which involves naming each label according to the prevailing proportion of present-day parish names contained within it. As a result of this method, the most representative units within contiguous labels may both derive from a singular, currently existing parish.

This predicament suggests that, depending on the thresholds applied—whether based on demographics such as family counts, housing, or employment status—certain parishes manifest a propensity to be subdivided or consolidated in the regionalization framework. It highlights an underlying complexity in spatial demarcations and the need to refine our labeling approach to accommodate the nuances of spatial continuity and administrative boundaries. It also underscores the dynamic nature of urban geographies (Blunt and Sheringham 2019), where the demarcation of neighborhoods

can shift in response to the specific criteria or variables under consideration in regional planning.

On the other hand, regional delineations grounded in the thresholds of building and individual counts are associated with less compact regional shapes. This pattern likely mirrors a more dispersed and expansive spatial distribution of these elements, possibly indicating suburban or less densely populated areas. Such a distribution pattern can be crucial in urban planning and policy-making, as it reflects the underlying urban sprawl and the distribution of infrastructure and resources.

The comparison between the six outcomes derived from distinct classification variable thresholds and the previous results based on 2011 census reveals substantial differences, which hold significant implications for our research. These differences are not merely statistical variances but are indicative of the profound influence that the choice of threshold variables exerts on the delineation of districts. The sensitivity of regionalization to the selected threshold variables highlights the pivotal role these indicators play in shaping the spatial narrative of an area. It compels researchers to exercise judicious selection of these thresholds, ensuring that they are reflective of the study's specific aims and the inherent characteristics of the population and infrastructure being examined. They suggest that changes in data collection methods or classification criteria over time can lead to fundamentally different outcomes, challenging the notion of static or universal urban boundaries (Wineman, Alia, and Anderson 2020). This underscores the importance of contextualizing regionalization results within the temporal and methodological framework in which they were generated.

Furthermore, the study's use of the Isoperimetric quotient (IPQ) indices across various categories reveals a pervasive trend towards fragmentation in the regionalization results. The low IPQ scores across the board are indicative of a theoretical predilection for highly fragmented outcomes in regional delineation. This trend points to the need for a more integrated approach in urban planning that considers the interplay of various socio-economic and structural elements.

The application of Jaccard similarity analysis sheds light on the impact of Lisbon's 2012 administrative reform on the city's spatial organization. Remarkably, this analysis reveals that the reform did not bring about substantial changes in the spatial arrangement of the city's districts. The new regional classifications, based on a plethora of demographic and structural factors such as family size, educational attainment, and employment status, mirror the divisions existing before and after the reform. This high degree of similarity underscores the enduring nature of these factors as critical components of Lisbon's administrative geography. The consistency observed in these classifications likely mirrors the deep-rooted socio-economic structures that have long shaped Lisbon's urban fabric. This revelation is significant as it highlights the stability and resilience of the city's socio-economic landscape amidst administrative changes, offering valuable insights for future urban development and policy planning.

5.2 Limitations

First, the census data, while collected at high standards, are subject to potential obsolescence given their decadal interval, raising concerns about their timeliness. Despite stringent collection protocols, there remains a possibility of errors, omissions, or inaccuracies. For specialized applications requiring finer spatial resolution, the data may seem generalized. Given the inherently non-negative characteristic of census data, employing Non-negative Matrix Factorization (NMF) techniques is appropriate. However, one challenge with NMF is its propensity to converge on local optima. As with all dimensionality reduction methods, there is a risk of losing detailed patterns, particularly when minimizing the number of components. In the realm of regionalization using the MaxP method, its advantages are evident. MaxP adeptly captures the spatial aspect of geographical data, ensuring that statistically significant districts represent authentic geographic patterns. Nevertheless, it is not without pitfalls. Like many optimization techniques, MaxP is susceptible to local optima. Selecting appropriate parameters, especially thresholds, may require iterative tuning, especially for large datasets. Moreover, this method tends to merge smaller areas into larger ones, potentially sacrificing finer geographic details.

Executing the Max-P model presents substantial computational difficulties, particularly when larger threshold values are set. The runtime of the program can exponentially increase, with the duration stretching up to eleven hours. This extended processing time significantly restricts the number of debugging iterations we can perform, necessitating a reduction in trial runs to expedite the achievement of viable results. The need to minimize iterations imposes a stringent constraint on our analytical process, compelling us to make calculated decisions about which model configurations to prioritize in pursuit of our research objectives.

These computational challenges highlight the delicate balance between data detail and processing capacity in spatial analysis. To mitigate these issues, we are investigating methods to optimize the Max-P algorithm, such as employing parallel computing and refining data structures. The integration of high-performance computing resources also stands as a potential solution to enhance the efficiency of our modeling endeavors. Such improvements are essential to enable a more comprehensive and iterative examination of the patterns within our scope of study, thus expanding the capabilities and applicability of our regionalization analyses.

Our strategy, aligned with the existing administrative or statistical boundaries of 24 districts, ensures practical applicability and operational simplicity, aiding policymaking, planning, and comparative research. However, this simplified approach may not encapsulate the intrinsic structure of the data, potentially impeding the revelation of core data patterns. Adhering strictly to a preset number of areas may limit MaxP's flexibility, stifling the search for adaptive partitioning and potentially overlooking certain detailed patterns. Furthermore, the discussions in the latter part of this chapter

only analyze the Jaccard index for selected areas, while the analysis of Intersection Area and Symmetric Difference Area is confined to a single area. This is due to the limitations in the length of content, preventing an exhaustive discussion, which results in the potential oversight of key information when partitioning certain areas.

The potential obsolescence of census data and the risk of NMF to converge on local optima, are important considerations for future research. To advance this work, future studies could employ more recent or real-time data sources to capture the most current urban dynamics. Additionally, they could explore alternative dimensionality reduction techniques that might better preserve detailed patterns or use ensemble methods to mitigate the risk of converging on local optima. When applying MaxP, researchers could also consider adaptive algorithms that dynamically adjust thresholds to better capture the granularity of urban structures. By proposing these avenues for future research, the paper can provide a roadmap for addressing its current limitations and improving upon the methodologies used. And policymakers could use the findings to understand the shifting demographics within districts, helping them to allocate resources more effectively. Residents might be interested in how redistricting could affect their daily commutes, access to services, and property values. Urban planners could leverage the detailed demographic breakdowns to design infrastructure that better serves the needs of the community. By illustrating these potential impacts, the paper can demonstrate the real-world relevance of its findings.

During our research, we identified an intriguing aspect worth delving into in 2011, Lisbon was divided into four districts, namely '1.º Bairro', '2.º Bairro', '3.º Bairro' and '4.º Bairro'. After the new parish planning in 2012, all parishes were reclassified into five zones: 'Centro Histórico', 'Ocidental', 'Centro', 'Oriental', 'Norte'. The figures are in appendix D. However, we found no documentation detailing the rationale or utility behind this reclassification. Notably, neither the former nor the latter divisions held administrative functions. Yet, this method of division seems to offer greater administrative efficiency and convenience compared to both the previous and current parish divisions. This undoubtedly presents another direction for future exploration.

6

Conclusion

This research provides a data-driven conceptual method for urban regional delineation, illustrating how city boundaries can be redefined from the vantage point of data science. Our findings and methodologies offer valuable insights for future urban planners. Although these outcomes might predominantly serve as a reference framework rather than a direct implementation guide, they lay the groundwork for a new direction rooted in scientific rationality and objectivity.

Nevertheless, our work merely marks the beginning in this expansive field of exploration. Comprehensive data-driven urban regional partitioning necessitates deeper, more extensive research. This encompasses augmenting data richness, deepening the understanding of its structure, employing a varied set of techniques for integrative analysis, and enhancing the rigor and multidimensionality of result evaluations. Continually validating and refining the model is pivotal for its maturity. Solely relying on quantifiable data for urban redistricting is insufficient. Urban planning involves factors challenging to quantify, like cultural heritage, historical significance, and community traditions. Quantifying these aspects is the next challenge to face.

As we gather more data, we must also be cautious of falling into the "data infatuation" trap. At times, more data doesn't necessarily translate to more meaningful or accurate conclusions. Irrespective of technological advancements, we must always recognize that all endeavors aim for a specific purpose. Perhaps, after successive trials and studies, we might unearth a streamlined and efficient method for urban regional partitioning, city redistriciting, presenting a potent tool for urban planners globally. Because as human civilization advances, we persistently grapple with the task of designing and structuring more efficient urban constructs to enhance our living experiences.

Bibliography

Biswas, S. (2022). *Spatial Optimization Techniques for School Redistricting*. Accessed: 2023-11-21. URL: https://vtechworks.lib.vt.edu/handle/10919/110433 (cit. on p. 6).

Blunt, A. and O. Sheringham (2019). "Home-city geographies: Urban dwelling and mobility". In: *Progress in Human Geography* 43.5, pp. 815–834. DOI: 10.1177/0309132518786590. URL: https://doi.org/10.1177/0309132518786590 (cit. on p. 29).

Brenner, N. (1999). "Globalisation as Reterritorialisation: The Re-scaling of Urban Governance in the European Union". In: *Urban Studies* 36.3, pp. 431–451. DOI: 10.1080/0042098993466 (cit. on p. 5).

Caves, R. W., ed. (2005). *Encyclopedia of the City*. Routledge. DOI: 10.4324/9780203484234 (cit. on p. 5).

Coombes, M. (2014). "From City-region Concept to Boundaries for Governance: The English Case". In: *Urban Studies* 51.11, pp. 2426–2443. DOI: 10.1177/0042098013493482 (cit. on p. 5).

DIJKSTRA, L. et al. (2020-03). *How do we define cities, towns, and rural areas?* URL: https://blogs.worldbank.org/sustainablecities/how-do-we-define-cities-towns-and-rural-areas (cit. on p. 5).

Duque, J. C., L. Anselin, and S. J. Rey (2012). "The MAX-P-REGIONS Problem". In: *Journal of Regional Science* 52.3, pp. 397–419. DOI: 10.1111/j.1467-9787.2011.00743.x. URL: https://doi.org/10.1111/j.1467-9787.2011.00743.x (cit. on p. 14).

Evaluation of the Local Government Reform (2013). URL: https://english.im.dk/media/22356/evaluation-of-the-local-government-reform-2013.pdf (cit. on p. 4).

Friedmann, J. and G. Wolff (1982). "World city formation: An agenda for research and action". In: *International Journal of Urban and Regional Research* 6.3, pp. 309–344. DOI: 10.1111/j.1468-2427.1982.tb00384.x (cit. on p. 3).

Glaeser, E. (2011). *Triumph of the City: How Our Greatest Invention Makes Us Richer, Smarter, Greener, Healthier, and Happier*. Penguin (cit. on p. 5).

Hamilton, R. and A. Rae (2020). "Regions from the ground up: A network partitioning approach to regional delineation". In: *Environment and Planning B: Urban Analytics and City Science* 47.5, pp. 775–789. DOI: 10.1177/2399808318804226 (cit. on p. 4).

Hedjam, R., R. Abderrahmane, and M. Cheriet (2022). "Non-Negative Matrix Factorization with Scale Data Structure Preservation". In: *ar5iv*. URL: https://ar5iv.org/html/2209.10881 (cit. on p. 11).

Jordan, D. P. (1992). "The City: Baron Haussmann and Modern Paris". In: *The American Scholar* 61.1, pp. 99–106 (cit. on p. 5).

Kaczmarek, T. (2016). "Administrative division of Poland—25 years of experience during the systemic transformation". In: *EchoGéo* 35. DOI: 10.4000/echogeo.14514 (cit. on p. 4).

Keil, R. (1994). "Global Sprawl: Urban form after Fordism?" In: *Environment and Planning D: Society and Space* 12.2, pp. 131–136. DOI: 10.1068/d120131 (cit. on p. 5).

Lourenço, J. M. (2021). *The NOVAthesis LaTeX Template User's Manual*. NOVA University Lisbon. URL: https://github.com/joaomlourenco/novathesis/raw/main/template.pdf (cit. on p. i).

Lynch, K. A. (2008). "What Is the Form of a City, and How Is It Made?" In: *Urban Ecology: An International Perspective on the Interaction Between Humans and Nature*. Ed. by J. M. Marzluff et al. Springer US, pp. 677–690. DOI: 10.1007/978-0-387-73412-5_44 (cit. on p. 5).

Marcuse, P. (2010). "In Defense of Theory in Practice". In: *City: Analysis of Urban Trends* 14.1-2, pp. 4–12 (cit. on p. 8).

McCartan, C. et al. (2022). "Simulated redistricting plans for the analysis and evaluation of redistricting in the United States". In: *Scientific Data* 9.1, Article 1. DOI: 10.1038/s41597-022-01808-2. URL: https://doi.org/10.1038/s41597-022-01808-2 (cit. on p. 6).

Mcnally, K. (2016-05). *Why did France change its regions?* URL: https://www.completefrance.com/news/why-did-france-change-its-regions-6261616/ (cit. on p. 3).

Parr, J. (2005). "Perspectives on the City-Region". In: *Regional Studies* 39, pp. 555–566. DOI: 10.1080/00343400500151798 (cit. on p. 5).

Parr, J. B. (2014). "The Regional Economy, Spatial Structure and Regional Urban Systems". In: *Regional Studies* 48.12, pp. 1926–1938. DOI: 10.1080/00343404.2013.799759 (cit. on p. 3).

Rey, J., D. Arribas-Bel, and L. J. Wolf (2023). *Clustering Regionalization—Geographic Data Science with Python*. Retrieved January 17, 2023. URL: https://geographicdata.science/book/notebooks/10_clustering_and_regionalization.html (cit. on p. 7).

Routley, N. (2018-01). *The Evolution of Urban Planning*. URL: https://www.visualcapitalist.com/evolution-urban-planning/ (cit. on p. 5).

Santos, J. L. L. d. (2015). "A Reforma Administrativa de Lisboa—A nova gestão de equipamentos coletivos urbanos de proximidade". PhD thesis. PhD Thesis (cit. on p. 8).

Seixas, J. (2019a). "A Reforma Político-Administrativa de Lisboa: Substância, evolução e reflexão sobre os processos de descentralização em Portugal". In: *IV Conferência P3DT*, pp. 14–21 (cit. on p. 8).

— (2019b). "The political-administrative reform of Lisbon: Substance, evolution and reflection on decentralization processes in Portugal". In: *CICS.NOVA, Faculty of Social Sciences and Humanities, New University of Lisbon* (cit. on p. 8).

Shelton, T. and A. Poorthuis (2019). "The Nature of Neighborhoods: Using Big Data to Rethink the Geographies of Atlanta's Neighborhood Planning Unit System". In: *Annals of the American Association of Geographers* 109.5, pp. 1341–1361. DOI: 10.1080/24694452.2019.1571895. URL: https://doi.org/10.1080/24694452.2019.1571895 (cit. on p. 29).

Smith, M. E. (2002). *The Earliest Cities*. URL: https://www.academia.edu/2976000/The_Earliest_Cities_2002_ (cit. on p. 5).

Urban Stormwater Management in the United States (2009). National Academies Press. DOI: 10.17226/12465 (cit. on p. 5).

Walker, K. (2023). *Analyzing US Census Data: Methods, Maps, and Models in R*. CRC Press (cit. on p. 6).

Wei, R., S. Rey, and E. Knaap (2021). "Efficient regionalization for spatially explicit neighborhood delineation". In: *International Journal of Geographical Information Science* 35.1, pp. 135–151. DOI: 10.1080/13658816.2020.1759806. URL: https://doi.org/10.1080/13658816.2020.1759806 (cit. on p. 14).

Wineman, A., D. Y. Alia, and C. L. Anderson (2020). "Definitions of "rural" and "urban" and understandings of economic transformation: Evidence from Tanzania". In: *Journal of Rural Studies* 79, pp. 254–268. DOI: 10.1016/j.jrurstud.2020.08.014. URL: https://doi.org/10.1016/j.jrurstud.2020.08.014 (cit. on p. 30).

A

Appendix 1

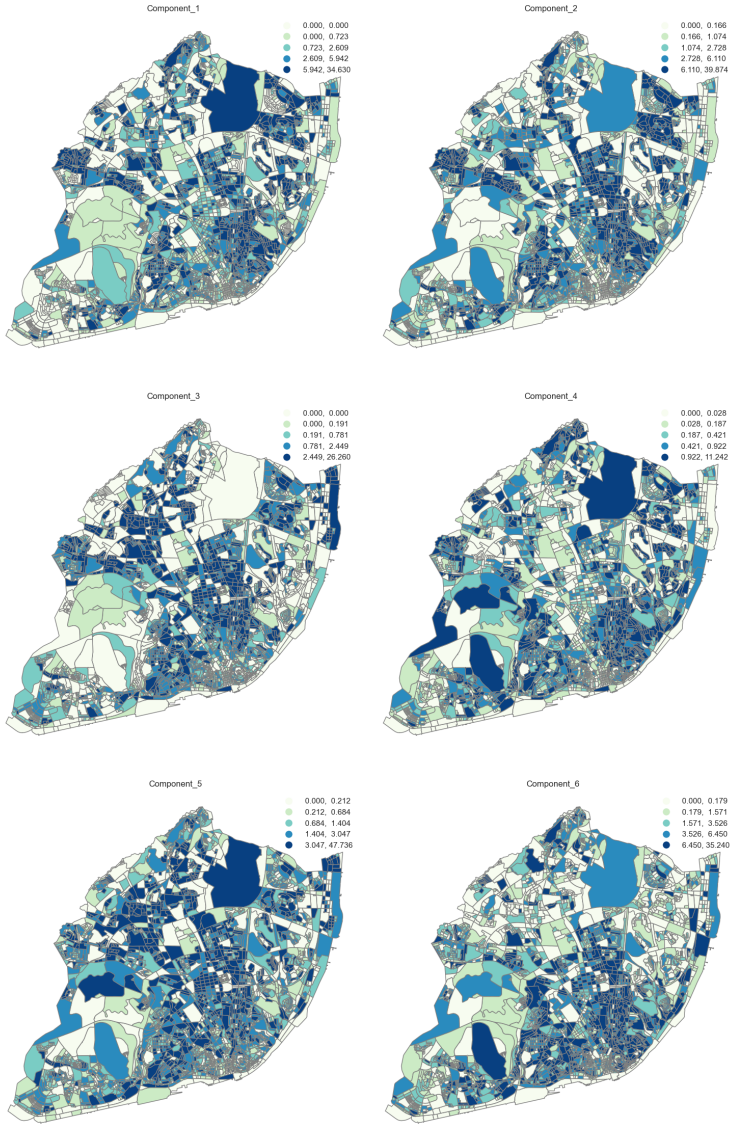
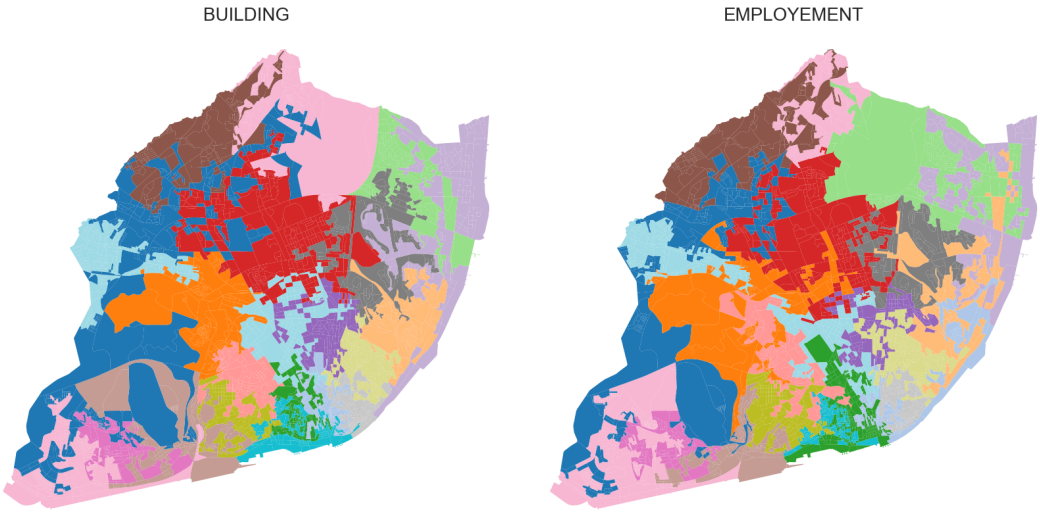


Figure A.1: Quintile Distribution Visualization of the 6 NMF Components. The figure showcases a quintile-based color gradation of the first NMF component across data, revealing the most significant areas of activity in the darkest shades.

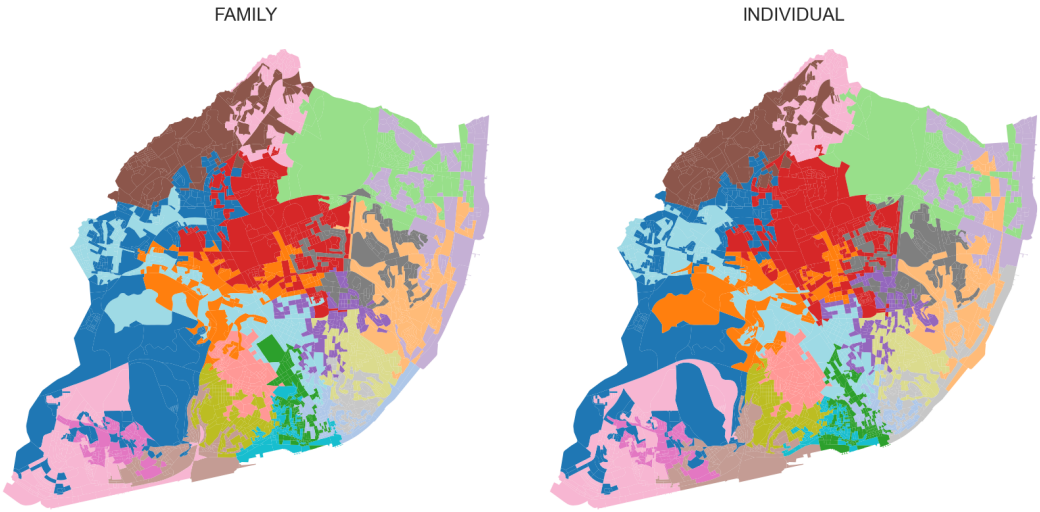
B

Appendix 2



Results based on buildings data

Results based on employement data



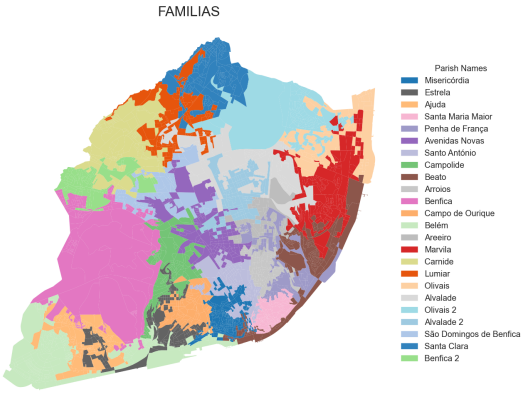
Results based on families data

Results based on individuals data

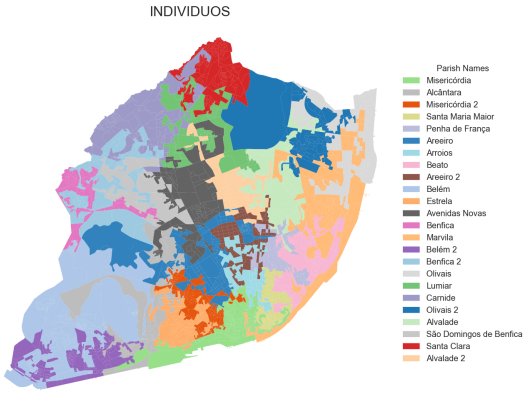
Figure B.1: Max-P with classification data for the case of buildings, employement, families and individuals

C

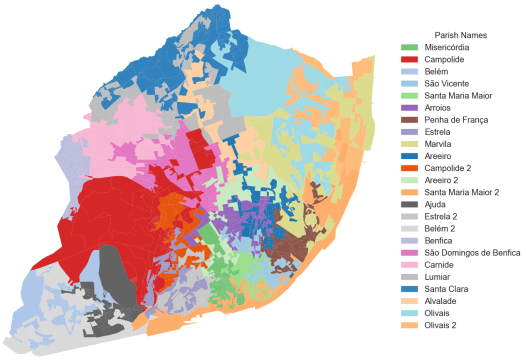
Appendix 3



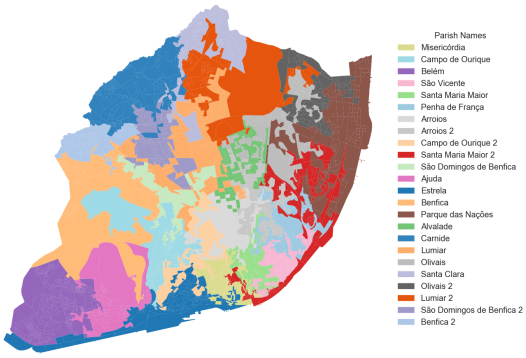
(a) Threshold being the number of families



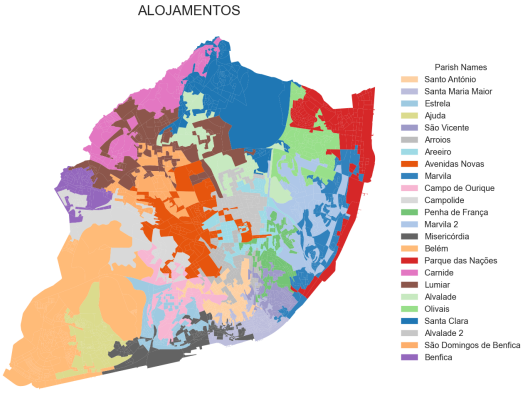
(b) Threshold being the number of presents



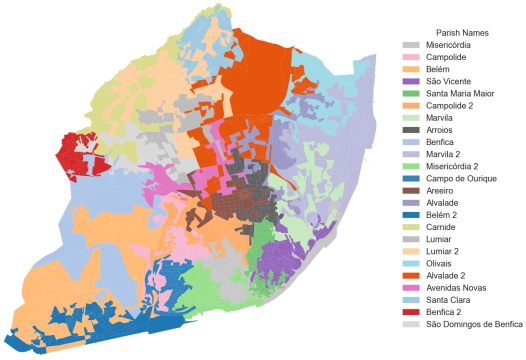
(c) Threshold being the number of buildings



(d) Threshold being number of employed residents



(e) Threshold being the number of accommodations

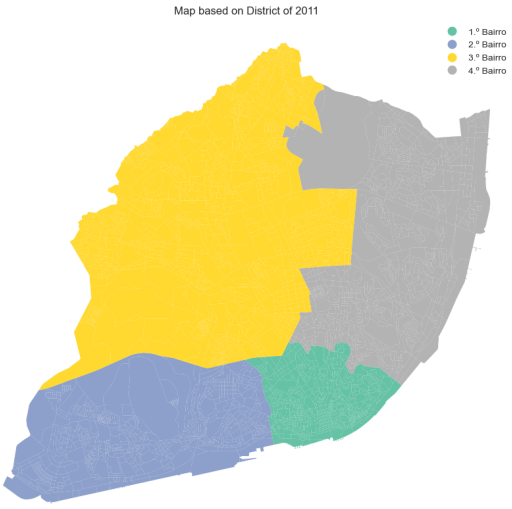


(f) Threshold being number of residents with higher education

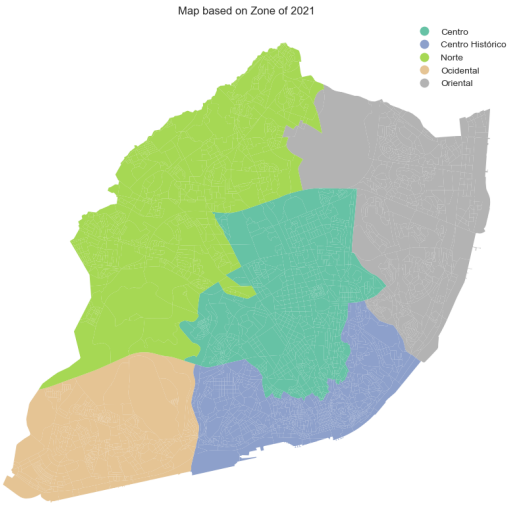
Figure C.1: Result based on the six NMF components with the threshold being the numbers of diffrent variables and giving parish names

D

Appendix 4



Map based on District of 2011



Map based on Zone of 2021

Figure D.1: Map of the two non-administrative areas plan of Lisbon

E

Appendix 5

For more information:

CENSOS 2011 (INE),

CENSOS 2021 (INE),

Lisbon City Council geographic data sets(Geodados),

List of parishes in Lisbon(Wikipedia),

Max-P Regionalization(Pysal),

Non-negative matrix factorization(Wikipedia)(scikit-learn),

Law no. 56/2012, of November 8th Administrative reorganization of Lisbon(PGDL).