Text

Description automatically generated

**MASTER IN DATA SCIENCE AND ADVANCED ANALYTICAL METHODS**

Reformulating Lisbon parishes

Tongjiuzhou Liu M20211012

**Supervisor:** Prof. Alberto Acedo Sanchez

Prof. Mijail Juanovich Naranjo Zolotov

**Introduction:**

In 2012 Lisbon suffered one of the most important administrative changes in its history. Their administrative boundaries changed from 54 units before 2012 to the current 24 parishes. This transformation combined adjoining parishes, alleviating the over-dimensioned administrative positions, and creating a new parish result of the 1998 Lisbon World Exposition. This research wants to understand this change from a data-driven approach.

Because in 2021, Portugal completes its national census. Exactly 10 years have passed since the last census in 2011. We will use data from both censuses to complete Reformulating Lisbon parishes as well as regionalization and clustering analysis.

**Methodological approach:**

We consider clustering techniques and regionalization methods. In this process, we will explore the socioeconomic characteristics of Lisbon. We will extract common patterns from a multidimensional data cloud generated by the Census Bureau on small regions. We first explore the multivariate nature of our dataset, suggesting some ways to examine statistical and spatial distributions before performing any clustering. Focusing on individual variables and their pairwise associations can help guide subsequent clustering or regionalization applications. We then consider clustering methods for geo-demography - applying multivariate clustering to spatially referenced demographic data. We use two popular clustering algorithms: K-means and Ward's hierarchical approach.

**Expected results and contributions:**

As we will see, mapping the spatial distribution of the resulting clusters reveals interesting insights into the socio-economic structure of Lisbon. We also see that in many cases clusters are spatially fragmented. That is, a cluster may actually consist of different regions that are not spatially connected. In fact, some clusters will have members spread throughout the map. This will illustrate why connectivity may be important when building insights into spatial data, as these clusters will not provide comprehensible regions at all. With this in mind, we will continue to regionalize and explore different ways of incorporating geographic constraints into the exploration of Lisbon's social structure. Finally, the most practical and effective model of regionalization for Lisbon is derived.

**Summary of the guidance document:**

In this paper they consider clustering techniques and regionalization methods. In the process, they explored the socioeconomic characteristics of neighborhoods in San Diego. They extracted common patterns from the cloud of multidimensional data that the Census Bureau produces about small areas through the American Community Survey. They began with an exploration of the multivariate nature of the dataset by suggesting some ways to examine the statistical and spatial distribution before carrying out any clustering. Focusing on the individual variables, as well as their pairwise associations, can help guide the subsequent application of clusterings or regionalizations. They then considered geodemographic approaches to clustering—the application of multivariate clustering to spatially referenced demographic data. Two popular clustering algorithms are employed: k-means and Ward’s hierarchical method. After that they moved on to regionalization, exploring different approaches that incorporate geographical constraints into the exploration of the social structure of San Diego.

Oitavo Relatório, June 2017. *Follow-up and Monitoring of the Lisbon Administrative Reform Process Eighth Report, Part 3 Lisbon's municipal decentralization process.* Lisbon Administrative Reform Follow-up and Monitoring Group*.*

The article describes in detail the process of municipal decentralization in Lisbon in this section. It covers the reasons for the start of this program and the final goals that were arrived at after the study. Several important guiding concepts and models of urban governance for the implementation of the program are clearly articulated. The process of forming the final program, guided by these concepts and models, is then also described.

It can be seen from the article that the existing 24 parishes of Lisbon were born entirely out of conference discussions, supported by advanced theories of urban governance but without a great deal of involvement of modern data science theories.

# Chao Wu, Wei Hu, Mengjie Zhou, Sheng Li and Yan Jia, 15 April 2019. *Data-driven regionalization for analyzing the spatiotemporal characteristics of air quality in China.*

This paper focuses on the extraction of local principal components of air quality indicators based on the geographically weighted principal component analysis (GWPCA) method, which is superior to PCA in accounting for spatial heterogeneity. A commonly used data-driven regionalization framework for studying air quality, identifying areas with similar air pollution behavior, and locating emission sources includes combining principal component analysis (PCA) with cluster analysis ( CA) methods in combination. However, traditional PCA does not consider spatial heterogeneity, which is a noteworthy problem in geographic studies. Then, based on the results of GWPCA, spatial cluster analysis (SCA) was used to identify areas with similar air pollution behavior. The results are all visualized, indicating that GWPCA has higher explanatory power than conventional PCA. Our revised framework for assessing air quality based on GWPCA and SCA can effectively guide environmentalists and geographers to assess and improve air quality from a spatial perspective. In addition, the visualization results can be used by urban planners and governments to monitor and manage air pollution. Finally, policy recommendations for mitigating air pollution through regional cooperation are suggested.

**Current research progress:**

We have now done a basic modeling and analysis of the regionalization of Lisbon based on the 2011 census data. We have also tried to introduce some external data such as airbnb's housing data. We are trying to find more data that can be used for geographic data analysis to analyze Lisbon's regionalization in many ways. We have already collected geographic data on transportation, education, health and security of Lisbon in Pordata. The next step will be to organize these data and combine them with census data for overall modeling. The 2021 census data will be used as the evaluation dataset. Finally, the results of our data-driven reform Lisbon regionalization will be compared and analyzed with the existing Lisbon regionalization to identify the strengths and weaknesses of the results.