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 some new parishes result of the 1998 Lisbon World Exposition. Portugal completed a national census in 2011, and the results of this census were an important reference data for the Lisbon parish reform. Portugal completed a new census in 2021, and ten years have passed since the Lisbon parish reform. This research looks at a data-driven approach to reformulate Lisbon's parish boundaries. As well as the regionalization and clustering analysis basic on the data from two census data.

In both censuses, we can find data on building status, housing status, household structure, demographic information, educational information and employment status, each of which is broken down into a number of different indicators These indicators are multidimensional and have far-reaching implications for parish reform in Lisbon. In previous studies, principal component analysis (PCA) has been combined with cluster analysis (CA) as a common data-driven regionalization framework for exploring important information from multidimensional raw data, which can be used to identify multidimensional indicators with similar characteristics and their resulting impacts. Specifically, PCA is a useful dimensionality reduction technique that identifies key variables by analyzing the correlation of indicators in each principal component (PC).CA is used to investigate similar data with similar variables and questions. And usually these studies are limited to substances that are highly mobile geographically, such as air quality analysis (Wu et al., 2019), temperatures (Carvalho et al., 2016), precipitation regimes (Darand & Mansouri Daneshvar, 2014). In this study we will look at population as a material of mobility as well, building a regionalized analysis model based on the use of PCA and CA.

However, traditional PCA does not take into account spatial heterogeneity, which is an important issue and characteristic of cross-spatial relationships and processes (Goodchild & Haining, 2004). The correlation between variables may change with location. Therefore, a model based on geographic weights must be applied. The Geographically Weighted Principal Component Analysis (GWPCA) method, as a localized version of PCA, has stronger explanatory power by considering the spatial non-smoothness of statistical significance compared to global PCA. A geographically weighted variant of PCA was used for the first time to explore the demographic characteristics of Northern Ireland(Lloyd, 2010). GWPCA was applied to study the residential environment and housing market segmentation(Wu et al., 2018). However, the application of GWPCA to urban regionalization remains understudied.

In this study, we will first determine the optimal number of regions based on two different data by using a global PCA approach and appropriate CA techniques. After this geographical weights (GW) will be added to our study by introducing 4 types of GW: QUEEN, KNN , BOLCK , UNITED to determine the most suitable type of GW for this study (J. Rey et al., n.d.). After determining the type of GW we need to add, we will keep tuning the parameters to get the optimal parameters for the GW we need. Finally, we will obtain two new models of the Lisbon parish area based on different generations of census data and reconstructed by GWPCA technique. After this, we need to analyze the two models by comparing their construction process and their results. The most influential elements in this process are explored, and the changes they produce. And analyze whether the five guidelines of regionalization are strictly adhered to: 1. all methods aggregate geographic areas into a predetermined number of regions, while optimizing a specific aggregation criterion. 2. regions within a region must be geographically connected (spatial continuity constraint). 3. the number of regions must be less than or equal to the number of regions. 4. Each region can be assigned to one and only one area. 5. Each region must contain at least one area (Duque et al., 2007).

We will conclude by exploring the implications of a data-driven regional reform of the Diocese of Lisbon based on the practical application of this technology and how it can actually be put into practice. The areas where more in-depth research is needed will be considered.

**References:**

Carvalho, M. J., Melo-Gonçalves, P., Teixeira, J. C., & Rocha, A. (2016). Regionalization of Europe based on a K-Means Cluster Analysis of the climate change of temperatures and precipitation. *Physics and Chemistry of the Earth, Parts A/B/C*, *94*, 22–28. https://doi.org/10.1016/j.pce.2016.05.001

Darand, M., & Mansouri Daneshvar, M. R. (2014). Regionalization of Precipitation Regimes in Iran Using Principal Component Analysis and Hierarchical Clustering Analysis. *Environmental Processes*, *1*(4), 517–532. https://doi.org/10.1007/s40710-014-0039-1

Duque, J. C., Ramos, R., & Suriñach, J. (2007). Supervised Regionalization Methods: A Survey. *International Regional Science Review*, *30*(3), 195–220. https://doi.org/10.1177/0160017607301605

Goodchild, M. F., & Haining, R. P. (2004). GIS and spatial data analysis: Converging perspectives. *Papers in Regional Science*, *83*(1), 363–385. https://doi.org/10.1007/s10110-003-0190-y

J. Rey, S., Arribas-Bel, D., & J. Wolf, L. (n.d.). *Clustering & Regionalization—Geographic Data Science with Python*. Geographic Data Science with Python. Retrieved January 17, 2023, from https://geographicdata.science/book/notebooks/10\_clustering\_and\_regionalization.html

Lloyd, C. D. (2010). Analysing population characteristics using geographically weighted principal components analysis: A case study of Northern Ireland in 2001. *Computers, Environment and Urban Systems*, *34*(5), 389–399. https://doi.org/10.1016/j.compenvurbsys.2010.02.005

Wu, C., Hu, W., Zhou, M., Li, S., & Jia, Y. (2019). Data-driven regionalization for analyzing the spatiotemporal characteristics of air quality in China. *Atmospheric Environment*, *203*, 172–182. https://doi.org/10.1016/j.atmosenv.2019.01.048

Wu, C., ye, X., Ren, F., & Du, Q. (2018). Modified Data-Driven Framework for Housing Market Segmentation. *Journal of Urban Planning and Development*, *144*. https://doi.org/10.1061/(ASCE)UP.1943-5444.0000473