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 result of this census was 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 regarding on the data from two census data.Regarding the administrative division of Lisbon, in the two census data we can see that the metadata unit in the data is the subsection. The subsections form a higher unit section, and the parish unit is made up of several sections. All Lisbon parishes form the Lisbon city level unit. Lisbon's administrative reform process, The reports prepared by the scientific team were presented and publicly discussed in November 2010, in a session organized and promoted by the Municipal Assembly. The reconfiguration of the administrative design of the Parishes, bringing them closer either to the contemporary identity configurations of the city, or to its number of inhabitants, or to a scale of critical mass that is more equitable and in line with the needs of empowerment of efficient and effective proximity urban governments - these, therefore, going from a number of 53 to 24 (Seixas & Rosa do Egipto, 2017). We will use the subsection data to be the unit as the basis for analysis of clustering regionalization, etc., for the ultimate purpose of Reformulating Lisbon parishes.

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(Granato et al., 2018; Gu et al., 2016; Penkova, 2017; Thorpe et al., 2016) which can be used to identify multidimensional indicators with similar characteristics and their resulting impacts (Genga et al., 2012; Hearty & Gibney, 2008). 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. Usually, these studies are limited to the geographical research of air quality analysis (Wu et al., 2019), temperatures (Carvalho et al., 2016), rainfall regimes (Darand & Mansouri Daneshvar, 2014). In this study, we will also consider the census results as high-dimensional self-similarity data and build 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, GWPCA has not been very widely implemented and applied in urban regionalization.

We will attempt 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.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Publication (APA style)** | **Country** | **Main topic/focus of study** | **Keywords** | **Independent variables** | **Dependent (target) variables** | **Theories used** | **Methods used** | **Technologies examined** | **Data (primary/secondary)** | **Main findings/contributions** | **Limitations** | **Conclusions** |
| 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 | Portugal | To study climate change on a regional scale using Earth system models, the spatial domain is divided into regions based on climate change | Climate change; Surface temperatures; K-Means Clustering; Precipitation; Europe | daily total precipitation, minimum and maximum temperatures | regions |  | MPI-ESM-LR model, K-Means Clustering |  | data used was provided by the Coupled Model Intercomparison Project Phase 5(CMIP5) | Regions can be identified based on multivariate climate change | data span is limited, and a portion of the data is generated by simulation. | The results of the univariate application of the method are consistent with those found in the literature, showing overall similar regions of variability. The regions obtained for the multivariate version are mainly defined by latitudes on European land, with some characteristics of land-sea interactions. Moreover, all regions have at least one variable with a statistically different distribution, thus providing confidence in the regions obtained. |
| 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 | Iran | Daily gridded precipitation data obtained from APHRODITE database between 1951 and 2007 were analyzed to regionalize the precipitation situation in Iranian countries. | Precipitation Regimes Principal Component Analysis (PCA) Hierarchical Clustering Analysis (HCA) GIS Iran | Precipitation Data | regions |  | Principal Component Analysis (PCA),Hierarchical clustering | APHRODITE's water resources website | APHRODITE daily gridded precipitation data for the Middle East from 1951 to 2007 | Regionalization of precipitation in Iran using principal component analysis and hierarchical cluster analysis | Data from third-party institutions, data reliability to be assessed | This study suggests the potential use of the APHRODITE dataset for the regionalization of precipitation conditions in Iran. Based on the results, it is recommended to use this dataset for grouping precipitation conditions in arid and semi-arid regions of mid-latitudes |
| Genga, A., Baglivi, F., Siciliano, M., Siciliano, T., Tepore, M., Micocci, G., Tortorella, C., & Aiello, D. (2012). SEM-EDS investigation on PM10 data collected in Central Italy: Principal Component Analysis and Hierarchical Cluster Analysis. Chemistry Central Journal, 6(2), S3. https://doi.org/10.1186/1752-153X-6-S2-S3 | Italy | Identify clusters of particles that can be distinguished on the basis of their chemical composition and morphology, study the relationship between chemical and morphological parameters and assess the differences between sampling sites. | Fractal Dimension Hierarchical Cluster Analysis Morphological Parameter Urban Site Soot Particle | Chemical elements of PM10 | clusters |  | Principal Component Analysis (PCA),Hierarchical clustering | Scanning electron microscope analysis | PM10 was collected at 3 different sites in central Italy with different conditions: yard, urban and rural sites | The identification of PM10 particle populations that can be distinguished based on chemical composition and morphology revealed differences between sampling sites. | Single-particle studies provide useful information for understanding the formation of particulate matter, but they are not exhaustive because the information they provide is provided over a limited period of time and therefore they are complementary to other PM studies. | Multivariate statistical analysis of particle composition and morphology data allows us to provide information on PM sources and highlight differences and similarities between sites. |
| 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 | USA | Some important developments in GIS and spatial data analysis since the early 1950s | GIS,spatial data analysis |  |  |  |  |  |  | A critical review of the past 50 years is presented. We then reflect on current challenges and speculate on the future. |  | The two fields of GIS and spatial data analysis meet in the field of geographic information science to support and add value to each other. |
| Granato, D., Santos, J. S., Escher, G. B., Ferreira, B. L., & Maggio, R. M. (2018). Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: A critical perspective. *Trends in Food Science & Technology*, *72*, 83–90. https://doi.org/10.1016/j.tifs.2017.12.006 | Brazil | Analysis of multivariate associations between bioactive compounds and functional properties in foods using principal component analysis (PCA) and hierarchical cluster analysis (HCA) | Principal component analysis,hierarchical cluster analysis,bioactive compounds,functional properties, food | Bioactive compounds and functional properties | clusters |  | Principal Component Analysis (PCA),Hierarchical clustering | NMR spectra, HPLC chromatograms | Chemical experimental data | When appropriate, it should be remembered that correlations between compound content and biological activity can be appropriately discussed using correlation coefficients. |  | The use of PCA and HCA in food chemistry research has increased in the last few years because the results are easy to interpret and discuss, especially when analyzing large data sets. |
| Gu, F., Hall, P., & Miles, N. J. (2016). Performance evaluation for composites based on recycled polypropylene using principal component analysis and cluster analysis. *Journal of Cleaner Production*, *115*, 343–353. https://doi.org/10.1016/j.jclepro.2015.12.062 | China | Evaluation of the properties of recycled polypropylene based composites using principal component analysis and cluster analysis | Principal component analysis,hierarchical cluster, recycled polypropylene | Properties of composite materials | clusters |  | Principal Component Analysis (PCA),Hierarchical clustering |  | Experimental test data | In this study, principal component analysis (PCA) was used as a decision support tool to evaluate the performance of plastic composites based on multiple properties. | Limitation of the number of experiments | This study focuses on the application of CA and PCA on recycled plastics and their talc or GF composites |
| Hearty, Á. P., & Gibney, M. J. (2008). Comparison of cluster and principal component analysis techniques to derive dietary patterns in Irish adults. *British Journal of Nutrition*, *101*(4), 598–608. https://doi.org/10.1017/S0007114508014128 | UK | Comparing cluster analysis and principal component analysis techniques to derive dietary patterns of Irish adults | Dietary patterns Cluster analysis Principal component analysis | Food intake data | clusters |  | Principal Component Analysis (PCA),Hierarchical clustering |  | Northern/Southern Ireland Food Consumption Survey 1997-9 | For cluster analysis, the most appropriate food group variable format was found to be the percentage contribution to energy intake, which yielded six clusters: 'Traditional Irish'; 'Continental'; 'Unhealthy foods'; 'Simple foods and low-fat milk'; 'Healthy foods' ; 'wholemeal breads and desserts'. For PCA, the g/d format of food groups was found to be the most appropriate format, which revealed four dietary patterns: 'unhealthy foods and high alcohol'; 'traditional Irish'; 'healthy foods'; 'sweet convenience foods and low alcohol'. | Data sources are old | Clustering and PCA identify similar dietary patterns when presenting the same data set |
| 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 | UK | Analysis of demographic characteristics using geographically weighted principal component analysis | Northern Ireland,Census data,Principal components analysis,Spatial statistics, Geographical weighting | census data | clusters |  | Principal Component Analysis (PCA),Hierarchical clustering,geographically weighted variant of PCA (GWPCA) |  | Northern Ireland as represented in data released from the 2001 Census of Population | This paper focuses on two particular issues that are rarely considered in multivariate demographic data analysis. First, it makes use of the correct treatment of component data (i.e., sets of values where the sum is a constant, such as 1 (proportion) or 100 (percentage)). | Limitations of the data | Key substantive findings include the fact that the characteristics that best distinguish members of a population are geographically variable. |
| Penkova, T. G. (2017). Principal component analysis and cluster analysis for evaluating the natural and anthropogenic territory safety. *Procedia Computer Science*, *112*, 99–108. https://doi.org/10.1016/j.procs.2017.08.179 | Russia | Principal component analysis and cluster analysis for assessing natural and man-made territorial security | comprehensive multidimensional analysis, principal component analysis, cluster analysis, data mining, terrritory safety evaluation, prevention of emergencies, territorial management | data of the Territory Safety Passports of the Krasnoyarsk region | clusters |  | Principal Component Analysis (PCA),Hierarchical clustering |  | 2014 collected in Center of Emergency Monitoring and Prediction (CEMP) | Construction of bicluster and tricluster structures in multidimensional data spaces; main cluster characteristics studied. The results of this analysis allow the identification of high-risk areas and ranking them according to the degree of danger with which natural and technological emergencies occur. This assessment informs decision-making, enabling authorities to more effectively deploy forces and means for territorial protection, and develop a system of measures to prevent and mitigate the consequences of large-area emergencies. | Principal component analysis without geographic weights | As part of the emergency risk assessment work, the results of this study made it possible for CEMP experts to develop a system of measures to prevent and mitigate the consequences of emergencies in the Krasnoyarsk region. |
| Seixas, J., & Rosa do Egipto, J. (2017). *Follow-up and Monitoring of the Lisbon Administrative Reform Process Eighth Report*. Lisbon Administrative Reform Monitoring Group. | Portugal | As the last report on the period in which the city's new municipal and governance structures were established, a more analytical and strategic structure for the entire process is presented. | Administrative Reform of Lisbon |  |  |  |  |  | Lisbon Municipality | It recalls the key aspects as well as the principles of urban governance that originated this broad process of reform of Lisbon's political spaces. | The data used come from the unilateral, Lack of third-party data | The report also includes, in its final chapter, a global appreciation of the whole process, as well as a systematization of the elements that this working group considers essential for a consolidated evolution of the proximity governance in Lisbon. |
| Thorpe, M. G., Milte, C. M., Crawford, D., & McNaughton, S. A. (2016). A comparison of the dietary patterns derived by principal component analysis and cluster analysis in older Australians. *International Journal of Behavioral Nutrition and Physical Activity*, *13*(1), 30. https://doi.org/10.1186/s12966-016-0353-2 | Australia | This study aimed to compare dietary patterns derived from principal component analysis (PCA) and cluster analysis (CA) in older adults and examine their associations with sociodemographics and health behaviors. | Principal component analysis Cluster analysis Dietary patterns Comparison Older adults Retirement | Health and Diet Data for the Elderly | clusters |  | Principal Component Analysis (PCA),Hierarchical clustering |  | Men (n = 1888) and women (n = 2071) aged 55–65 years completed a 111-item food frequency questionnaire in 2010. across the clusters using one-way ANOVA. | Both PCA and CA identified two major dietary patterns in pre-retirement elderly men and women. These results add to the limited literature on dietary patterns in older adults. | Capacity limitations of survey data, and reliability needs to be improved | Overall, the dietary patterns identified by PCA were more interpretable than those by CA. Initiatives targeting these collective health behaviors (risk factors for chronic disease) may help improve the health of older adults, the study suggests . |
| 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 | China | Spatial-temporal characteristics of air quality in China based on data-driven regional analysis | Principal Component Analysis (PCA) Hierarchical Clustering Analysis (HCA) China | Ambient air quality monitoring indicators | clusters |  | Principal Component Analysis (PCA),Hierarchical clustering,geographically weighted variant of PCA (GWPCA) |  | China National Meteorological Administration | The improved air quality assessment framework based on GWPCA and SCA can effectively guide environmentalists and geographers to assess and improve air quality from a spatial perspective. |  | The results show that GWPCA has higher explanatory power than traditional PCA. Policy recommendations for air pollution mitigation through regional cooperation are presented.. |
| 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 | China | A data-driven framework for delineating housing submarkets by integrating Geographically Weighted Principal Component Analysis (GWPCA), Spatial Heterogeneity Tests, Density-Based Spatial Clustering (DBSC) Algorithms, and Hedonic Validation | Residential factor; Principal component analysis (PCA); Geographically weighted principal component analysis (GWPCA); Spatial clustering; Submarket. | geographic data, housing prices | clusters |  | Principal Component Analysis (PCA),Hierarchical clustering,geographically weighted variant of PCA (GWPCA) |  | 2016 statistical yearbook of Shenzhen | Apply the modified framework to the housing market segmentation in Shenzhen, China. The results show that the modified framework works well in the Shenzhen market segment. This framework has important implications and great potential for identifying housing submarkets. Statistically, it can be generalized and applied to housing markets in other cities. | Some sources of data are unknown | The main contribution of this paper is the integration of GWPCA and DBSC into a framework that integrates housing markets from a statistical perspective. Analyzing the loadings of the PCs with the highest variable extracted from GWPCA proved important to draw some interesting conclusions about the relationship between the variable and housing desirability. |