R Shiny Application for Geospatial Clustering Analysis

Megan Sim Tze Yen
Singapore Management University
megansim.2018@smu.edu.sg

Darryl Kwok Bing Heng
Singapore Management University
darryl.kwok.2018@smu.edu.sg

Pengtai Xu
Singapore Management University
pengtai.xu.2021@smu.edu.sg

1. ABSTRACT

With Geographical Information being readily available everywhere, users now have a diverse range of data to perform Geospatial Analysis. The steps taken to prepare the data for in-depth analysis usually differ across datasets. The subsequent actions executed after data preprocessing are strikingly similar across Geospatial projects that falls under the same category i.e. Geographic Segmentation. To meet the rising demand for quick but accurate generation of the project results, we designed and developed a Geographical Segmentation Application that utilises different clustering algorithms such as, Hierarchical Clustering, Clustgeo and SKATER.

2. INTRODUCTION

Comprehensive plans are required to set goals and guidelines for future growth and development. These plans are required to improve the welfare of the people and induce the creation of better social, economic and physical environments. Government bodies often create untargeted and uniform policies or strategies that hope to address the majority of the issues in the country. They often fail to consider the specific needs of each group that are across the country. With spatial information being affected by many factors, there could be many different factors why and how occurrences of events are clustered or segregated in certain locations. We attempt to create a generalised Geographical Segmentation tool to identify clusters within any datasets to perform in-depth analysis. Geographic Segmentation divides and separates a target market into different segments by using geographical location, to better serve and target each segment specifically. This is often done based on geographic information and also various other factors such as climate, cultural preferences, populations and more. If government policies are untargeted and irrelevant, it will be costly and damaging. Geographical segmentation is an effective approach for government bodies to identify the specific needs of each segment to better serve them.

3. MOTIVATION AND OBJECTIVES

Our research and development were spurred by the increasing need and demand for quick but accurate execution of geographic segmentation. It aims to provide the users with the ability to perform parameter tuning and the flexibility of uploading different datasets to get a quick and in-depth analysis of the different segments of their data. More specifically, we attempted to create an application that supports the following objectives: 1. To be able to upload varying types of datasets that include geographical and non-geographical data. 2. To create an interactive visualization that encompasses the different segments, utilising a dataset of their choice. 3. To provide the users with the ability to choose an algorithm of their choice from a pool of frequently used algorithms. 4. To provide the users with the ability to tune and define the parameters that are required to generate an accurate and in-depth analysis of 5. their datasets. 5. To reduce the bottleneck caused by the repetitive acts of generating results from similar geographical segmentation projects.

4. LITERATURE REVIEW

We chose the study on Spanish Employment, Normative versus analytical regionalisation procedures (Juan Carlos et al, 2004) as our reference literature because we felt that the approaches taken were the most similar to our project. One method of study that stood out to us was the Two-Stage strategy method. In the first stage, the conventional clustering method was used. Hierarchical, Partitional and k-means clustering can be chosen as a method of choice for the first stage. At the second stage, cluster revision in terms of geographical contiguity will be executed. This is to ensure the homogeneity of regions that were generated in the first stage. Additionally, it is to obtain evidence of spatial dependence among the elements. However, it was also found that the number of groups actually depended on the degree of spatial dependence rather than the researcher's criteria of study.

In the study, Normative and Analytical Regionalisation were 2 types of regionalisation used. Although normative regionalisation is a good method to use, it might not always be appropriate due to some underlying factors that we might not be able to discover. This will definitely affect the credibility of the results. On the other hand, analytical regionalisation takes into account functional zones, geographical contiguity, equality as well as the interaction between the regions. We feel that with the consideration of a wide range of factors, analytical regionalisation is a far more superior

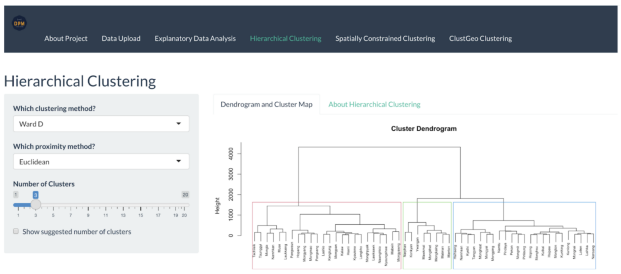

Figure 1: A Sample Page for Layout Demonstration

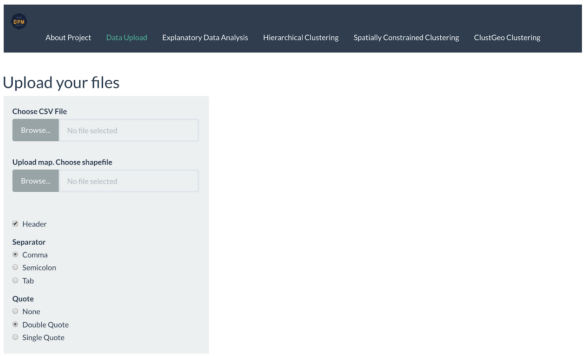

Figure 2: Data Upload Page

approach than normative regionalisation.

5. DESIGN FRAMEWORK

We designed our data visualization according to Schneiderman's mantra [1]: zoom and filter, details on demand. The types of views available consist of maps and charts, which correspondingly deal with geospatial and aspatial visualisations.

We have a sequential flow in terms of user journey with the application. The overall sequence includes: data upload, exploratory data analysis and in-depth clustering analysis. Multiple clustering analysis methods have been made available to the user, which covers the main clustering methods used in both academia and industry. These methods include hierarchical clustering, clusterGeo and SKATER.

We have a consistent design for each section of the dashboard. For each page, we have a control panel on the left, which allows users to select analysis methods according to their own needs. On the right, we show the detailed results of the analysis in intuitive visualizations.

6. DEMONSTRATION

We allow users to upload their own datasets for clustering analysis (Figure 2).

After data uploading, users are able to conduct exploratory data analysis to decide the clustering method suitable for their case (Figure 3).

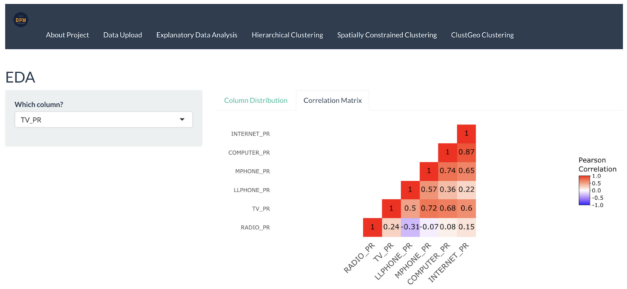

Figure 3: Exploratory Data Analysis Page

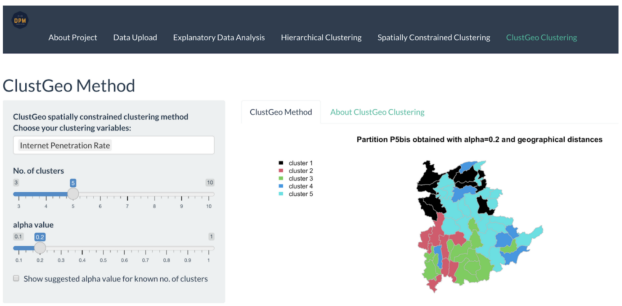

Figure 4: ClustGeo Clustering Analysis Page

Users then go to the most suitable clustering analysis method to conduct detailed analysis in order to understand the clustering pattern of their geospatial dataset (Figure 4).

7. DISCUSSION

With our regionalisation & geographical segmentation tool, users have a hassle-free way of obtaining fast visualisations with tweakable parameters and algorithms. In addition, users who might be new to spatial analysis get to work with a set of algorithms that are commonly used in the field, as well as read the explanations on how said algorithms work and a breakdown of the parameters (such as different proximity methods).

To use our tool, we require users do a preliminary clean-up of the data - while we have pre-processing functions to address things like the coordinate system and missing values, responsibility is on the user to ensure that their geo spatial and aspatial data are in the appropriate formats (which are widely used within the field of spatial analysis), as well as ensure that both datasets have similar features to 'join' on. This serves as a good practice for users to carry with them, not just for using this tool but also for future geographical analyses they might explore, where data cleaning and manipulation serves as a good foundation to build their analyses upon.

Our tool is simple by design: it is meant to set a foundational view or serve as a preliminary visualisation for analyses requiring geographical segmentation. However, users who want greater levels of customisation or deeper insights might want to turn to other tools on the market, or attempt

such regionalisation analysis themselves. Of course, there are features we would have liked to add to our current tool to make it more usable for all types of users, which is detailed in our next section.

8. FUTURE WORK

While we have attained the primary objective of creating our regionalisation tool, there are a few improvements that we would have liked to incorporate given more time and resources. These improvements are mostly centered on improving user experience by making the tool more reactive and intuitive, such as: Less stringent data requirements: our data input requires the aspatial and geospatial data to be in .csv and .shp format respectively, and to join them, the identifying feature to join on has to be of the same name. This puts responsibility on the user to clean up/manipulate their data in order to use our tool, which might dissuade users from trying the tool, or might even make the tool unusable for those who have issues with data cleaning + manipulation. Ideally, an improved version of our tool would be able to take in a variety of data (in typical formats). Additional pre-processing features: our preprocessing is the default preprocessing for data meant for geographical segmentation. This might fulfil the basic requirements of the user, but they might want to manipulate specific variables or omit certain sections of the pre-processing. Auto-detect & suggest creation of new variables: In our test dataset, we have preemptively created a new set of variables for gauging the penetration rate (for each ICT measure), which gives a more accurate sensing of distribution over a counting-based distribution. However, it would be ideal if the creation of such new variables is intuitively derived from the dataset: the user should have some input as to what new variable they want to derive, and for users who are less sure about which variables are appropriate for geographical segmentation, the tool should suggest the recommendation of new variables. Visualisation of segmentation over time: certain applications of regionalisation, such as a study on the cases of a particular disease, might want to find out if there was a change of the regions over time, and if so, what associated variables could have served as contributing factors. An interactive visualisation or a .gif of the regions over a period of time would prove to be beneficial for such cases. Greater customisation: customisation options such as choosing the palette of the visualisations or being able to title said visualisations might be important for users who want to use the visualisations in their work.

References

[1] Shneiderman, B. The eyes have it: A task by data type taxonomy for information visualization. *IEEE Conference on Visual Languages.* 336–343.