Standard and Spatial PCA Analysis on Third World Countries

Jamp Vongkusolkit

Introduction

The purpose of this paper is to analyze the dataset for four different third world countries with the same variables and visualize the differences/relationships between each region for each country. In order to do this, principal components analysis (PCA) and spatial principal components analysis (sPCA) are used.

Motivation & Goal

The goal of this study is to understand the relationships of physical and cultural variables associated with each region in each country to determine if the importance of each variables vary across the regions. Understanding these relationships may enable us to gain insight on some factors that may affect poverty in third world countries spatially.

Required Data

The dataset was obtained from Kaggle, where the owner of the dataset provided it as a contribution to Kiva (a non-profit organization that allows people to lend money to low-income entrepreneurs, farmers, or students online in order to help alleviate poverty). This dataset was be converted into spatial points in R. The polygons for the countries and their regions/provinces were obtained from Natural Earth Data. The polygon for each country was filtered using ArcGIS.

Methods

When doing principal component analysis on spatial data, there are typically two ways to approach this problem. The first method would be performing standard PCA on the features excluding the location (i.e. latitude and longitude). Standard PCA was used in this study to analyze four different third world countries and find important features within them and see if some features have similar importance throughout different locations.

The second method would be to perform PCA while taking into account the spatial attributes. SpatPCA, an R package, was implemented in this study to compare the results with the standard PCA approach on four different countries from a spatial dataset consisting of different cultural and physical attributes, such as poverty percentage and the elevation of the region.

Conclusion

In this study, it was found that each country was relatively unique in terms of the features of our dataset. However, some features did seem to be important across different countries, such as evaporation levels in Malawi and Nigeria. Considering these countries are geographically close to each other and exhibit similar patterns, the geospatial location of a certain region plays a role what features were important. This property was shown when regions within a country were observed to exhibited similar patterns to regions that were geographically near.

References

Wen-Ting Wang & Hsin-Cheng Huang (2017) Regularized Principal Component Analysis for Spatial Data, Journal of Computational and Graphical Statistics, 26:1, 14-25, DOI: https://doi.org/10.1080/10618600.2016.1157483

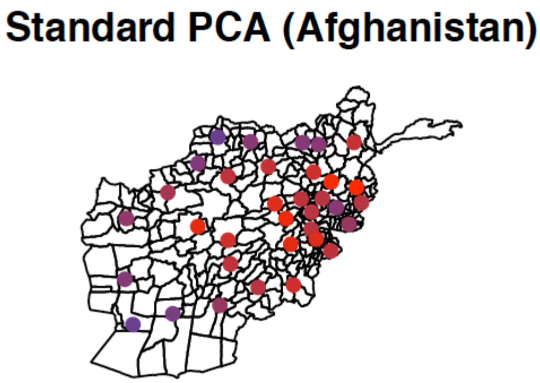

Figure 1.4

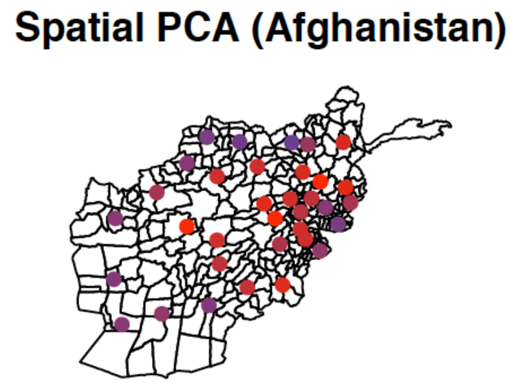

Figure 1.5

Brazil Analysis

The scree plot for standard PCA on Brazil (Figure 4.1) shows that the first principal component captures almost 80% of the variance, while PC2 captures around ~10% of the variance. Like Nigeria and Malawi, evaporation was a strong influencer in PC1. The strongest influence in PC2 was time to city.

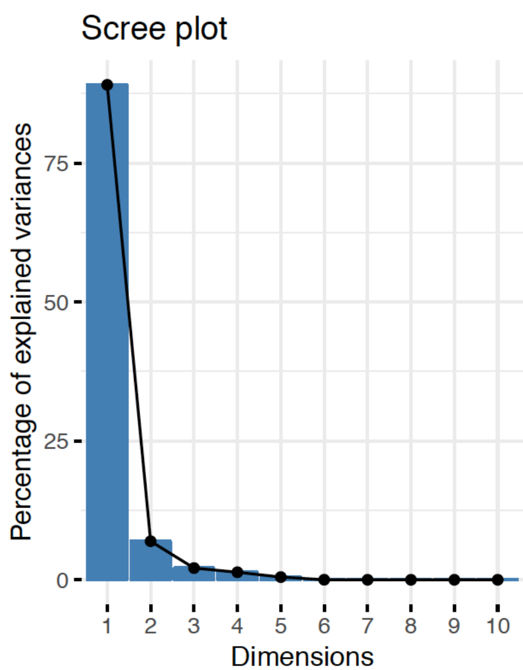

Figure 4.1

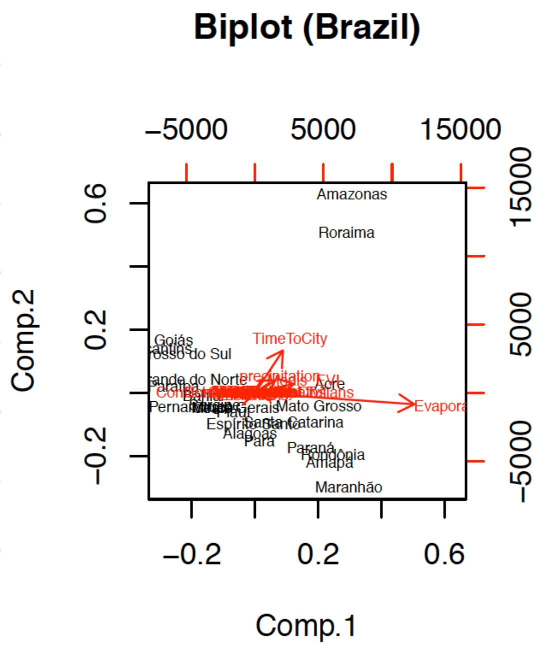

Figure 4.2