Identification of homogenous regions of Texas based on demographic characteristics

Introduction

Understanding the demographics of a region is important for a country for decision making terms of social, political, medical and economic aspects. This study of unsupervised clustering analysis will group the regions of Texas into homogenous and spatially adjacent counties based on its their cultural, political, socio-economic, demographic and residential characteristics.

Feature Selection and Preparations

The features selected based on 4 groups as given below:

a) Cultural: WHINOHISP
b) Political: republic
c) Socio-economic: COLLEGEDEG, UNEMPL, INCOME, CRIMERATE
d) Demographic: AGE18TO64, POPDENSE, AGE65UP, UNINSURED
e) Residential: MEDVALHOME, BPOST2000

The features were selected from the Esri shapefile. I created BPOST2000 feature for the ease of interpretation, which consists of the percent of housing units built from 2000 to 2018. I selected these features in order to explore how the demographics are distributed in terms of their standard of living, ethnicity, culture and political belief. This exploration is based on the assumption higher unemployed people lead to higher crime rate, lower standard of living, most prevalent among the working aged (Age 18 to 64), culturally diverse group of people.

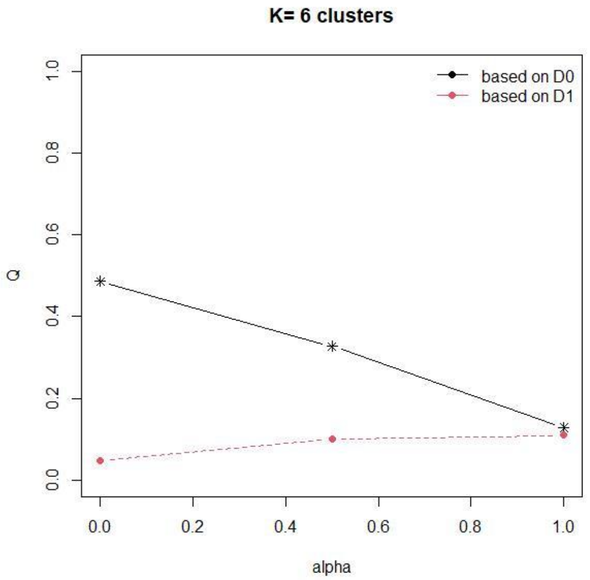
I believe the division of population based on gender is an important parameter which is missing in the attribute table of the **TXCnty2021** shape file.

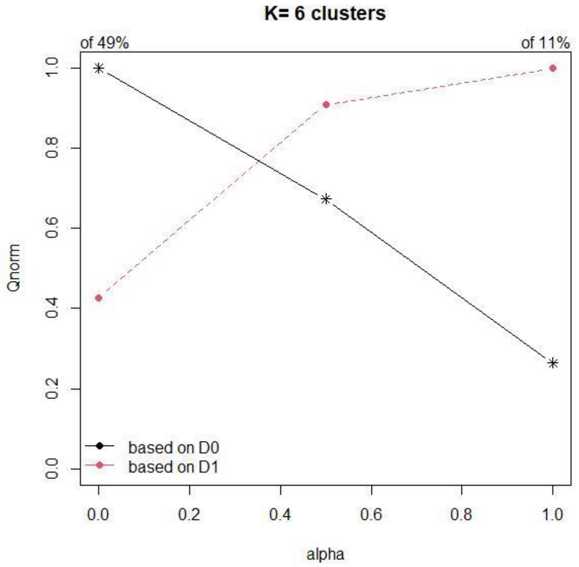
All the variables and different measurement units, so they were scaled for the ease of comparison.

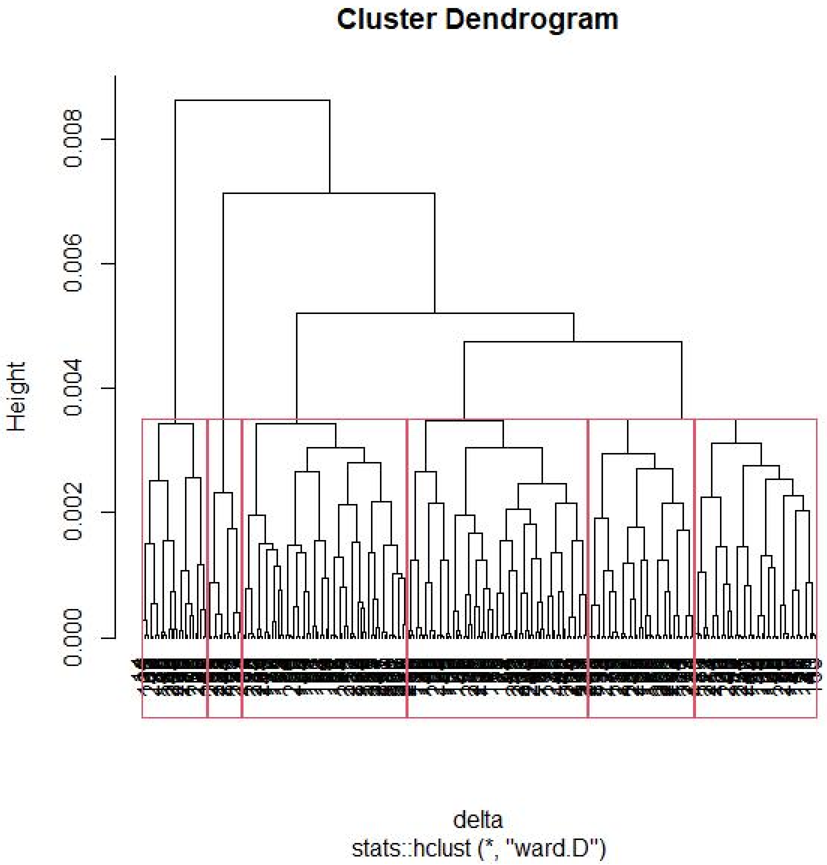
Finally, I selected the topoDist among the three spatial relationship distance matrices to avoid fragmented heterogenous cluster outcomes.

Iterative Cluster Identification

I have selected k = 6 and alpha = 0.3 since it gives a good cluster interpretation.

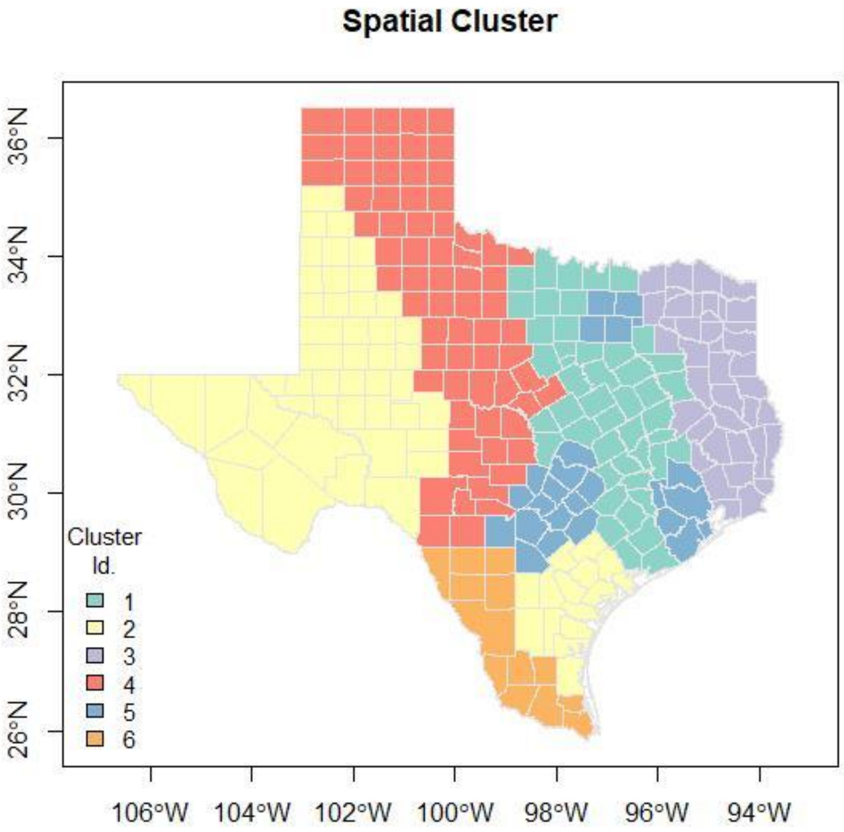


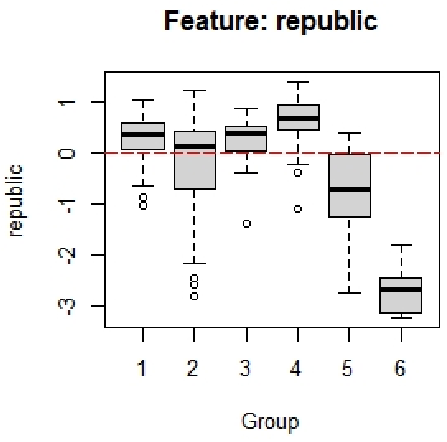


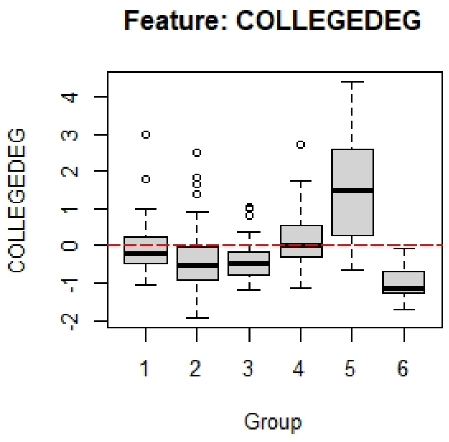


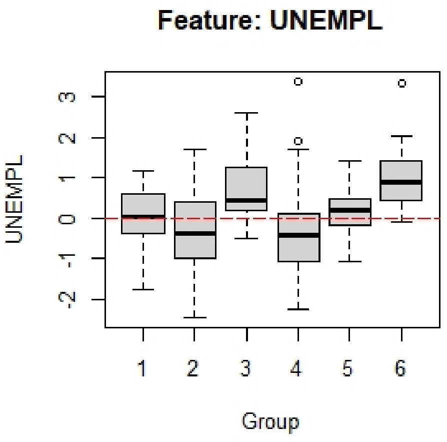
```
table(neighClus)
1  2  3  4  5  6
46 62 40 68 25 13
```

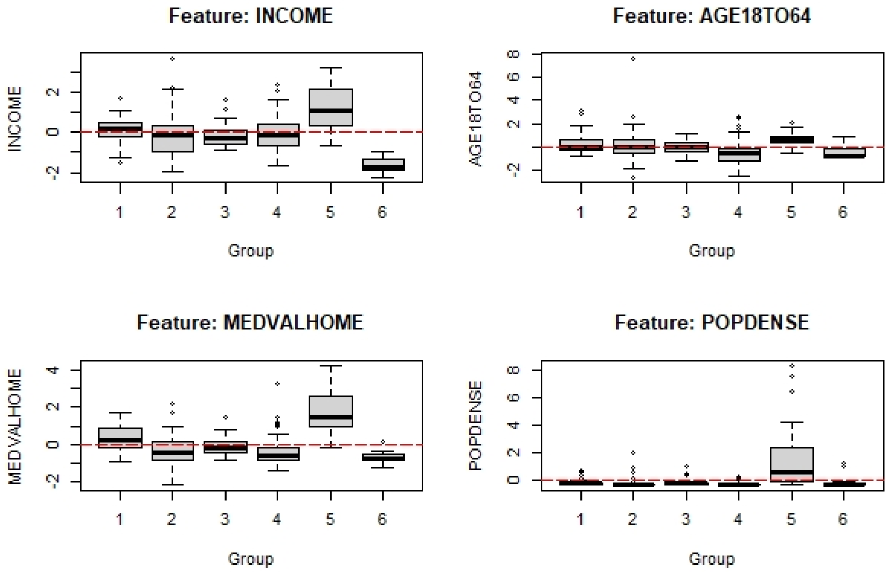
Interpretation of Results

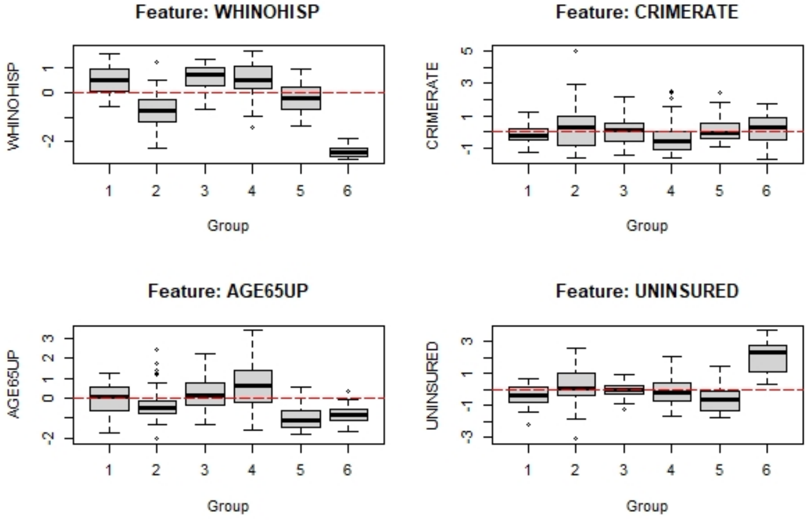



Feature: republic
republic
1
0
-1
-2
-3
1
2
3
4
5
6
Group


Feature: COLLEGEDEG
COLLEGEDEG
4
3
2
1
0
-1
-2
1
2
3
4
5
6
Group


Feature: UNEMPL
UNEMPL
3
2
1
0
-1
-2
1
2
3
4
5
6
Group


Feature: INCOME
INCOME
-2
0
2
Group
1
2
3
4
5
6
Feature: AGE18TO64
AGE18TO64
-2
2
4
6
8
Group
1
2
3
4
5
6
Feature: MEDVALHOME
MEDVALHOME
-2
0
2
4
Group
1
2
3
4
5
6
Feature: POPDENSE
POPDENSE
0
2
4
6
8
Group
1
2
3
4
5
6



Group 1: Major white population – This group is based on white male, mostly on the western part of Texas. They are mostly republican, less educated and less people with insurance.

Group 2: Less educated majority – Mostly located at the west coast, and some clustered around the south, this group has less education, but more employed and non-white majority people. The region is inexpensive in household characteristics.

Group 3: West coast white majority – Located at the very west coast of Texas, this group consists basically white majority, unemployed, republicans. Their income and population density is below average.

Group 4: Mid Texas older group – This group lies in the middle of Texas, mostly older majority. This locality shows lower crime rate, white non-Hispanic people. Job opportunity could be better in this location due to higher employment of the people.

Group 5: Scattered educated people – This group has higher education status and higher level of income, dense population. Home price is the most expensive in this area. Mainly the people who are active aged 18 to 64, occupy this region.

Group 6: Democratic minded people – clustered at the very south of Texas, are the democratic minded people, characterized by less education, more unemployment. They are mostly uninsured non-white group in the region where crime rate is higher than average.

Appendix

```
setwd("D:\\GISC 6323 Machine Learning\\Lab 05")   # Directory with the map
data
getinfo.shape("TXCnty2021.shp")
neig.shp <- rgdal::readOGR(dsn=getwd(), layer = "TXNeighbors",
                           integer64 = "allow.loss", stringsAsFactors=T)
interState.shp <- rgdal::readOGR(dsn=getwd(), layer = "InterStateHwy",
                                 integer64 = "allow.loss",
stringsAsFactors=T)
county.shp <- rgdal::readOGR(dsn=getwd(), layer = "TXCnty2021",
                             integer64 = "allow.loss", stringsAsFactors=T)



county.bbox <- bbox(county.shp)                             # county bounding
box for map region
county.centroid <- coordinates(county.shp)                  # Get county
centroids

county.shp$republic <- county.shp$TRUMPVOT16/county.shp$TOTALVOT16
county.shp$BPOST2000 <- county.shp$B2000PCT+county.shp$B2010PCT

varKeep <- c("republic","COLLEGEDEG", "UNEMPL", "INCOME","AGE18TO64",
             "MEDVALHOME","POPDENSE", "WHINOHISP", "CRIMERATE","AGE65UP",
             "UNINSURED","BPOST2000")
xVars <- county.shp@data
xVars <- as.data.frame(xVars[varKeep])
row.names(xVars) <- 1:nrow(xVars)
featDist <- dist(scale(xVars))
K <- 6                           # Number of distinct clusters
range.alpha <- seq(0, 1, by=0.1)   # Evaluation range

cr <- choicealpha(featDist, topoDist, range.alpha, K, graph=TRUE)
tree <- hclustgeo(featDist, topoDist, alpha=0.3)
plot(tree, hang=-1)
rect.hclust(tree, k=K)
neighClus <- as.factor(cutree(tree, K))        # Determine cluster membership
table(neighClus)                               # number of tracts in each
cluster
mapColorQual(neighClus, tractShp,
             map.title ="Spatial Cluster",
             legend.title="Cluster\nId.", legend.cex=0.9)
plotBoxesByFactor(rVars[,1:3], neighClus, ncol=2, zTrans=T, varwidth=F)
plotBoxesByFactor(rVars[,4:7], neighClus, ncol=2, zTrans=T, varwidth=F)
plotBoxesByFactor(rVars[,8:11], neighClus, ncol=2, zTrans=T, varwidth=F)
```