Lab05: Hierarchical Clustering under Spatial Constraints

Handed out: April 12, 2023
Due date: April 30, 2023 as Word document into eLEARNING's **`Lab05Submit`** link.
Grading: Lab05 counts for 15 % of your final grade.

Objective

The objective of your study is to identify homogeneous regions in Texas, which comprise of similar and spatially adjacent counties. These regions should be homogenous with respect to their [a] cultural, [b] political, [c] socio-economic, [d] demographic and [e] residential characteristics. A cluster analysis like this could be an input into a demographics study (see Demographics Definition (investopedia.com)). An example of a spatial cluster analysis for the census tracts in Dallas County can be found in the vignette **`DallasMarketAreas`** in the package **`TexMix`**[1]. More information about the data and the Texas counties can be found at Disparities in COVID-19 Vaccination Rates among the Counties of Texas (spatialfiltering.com).

Data

Use the data in the zipped file **`TXCnty2021.zip`**. Its file **`TXCntyVars2021.pdf`** documents the used features. The script **`Lab05StarterCode.R`** sets your data up, provides information on how to calculate varying forms of geographic relationships and how to map your results. To calculate attribute distances among the counties, use ***only numeric*** variables.

Reading

Study the article by M. Chavent, V. Kuentz-Simonet, A. Labenne, J. Saracco. *ClustGeo: An R package for hierarchical clustering with spatial constraints* (see 1707.03897.pdf (arxiv.org))

You do not need to consider weights and standardization of the input distance matrices. This is done internally by the function **`hclustgeo( )`**.

Section 3.2 discusses criteria to evaluate the mixing of the feature and geographical distance matrices. It discusses the trade-offs between feature homogeneity within the clusters and the geographic cohesion of the partitions that are obtained for varying α-parameter and at a given number of clusters K.

You need to understand this article well to interpret the provided graphs and select an appropriate mixing α-parameter.

Feature Selection and Preparations

Identify the ***metric*** features that describe potential differences between putative regions, which are expected to exhibit a strong degree of internal cluster/region homogeneity.

For the sake of interpretability restrict the number of the features to a manageable set.

Justify with respect of the potential classification, why you selected the feature your choice.

---

[1] To install the **`TexMix`** package run in the console the statement **`install.packages("http://www.spatialfiltering.com//ThinkR/Downloads/TexMix_0.5.3.tar.gz", repos=NULL)`**
You may get first error messages asking you to install dependency packages such **`classInt`** or **`Formula`** prior to installing **`TexMix`**.

Can you think of relevant features that unfortunately are not included in the attribute table of the **TXCnty2021** shape file?

Evaluate whether you need standardize the features and whether some of the are redundant.

```
rm(list=ls())                                # Clear environment
oldpar <- par()                              # save default graphical
parameters
if (!is.null(dev.list()["RStudioGD"])) # Clear plot window
  dev.off(dev.list()["RStudioGD"])
cat("\014")                                  # Clear the Console

library(TexMix)
library(sp)
library(ggplot2)
library(maptools)
library(spdep)
library(foreign)
library(e1071)
library(VIM)
library(ClustGeo)

setwd("/Users/jimpan/Documents/EPPS 6326/lab/lab5/TXCnty2021")
# Read in the shapefile and its associated DBASE file

#Q1 Feature Selection and Preparations

TXCnty2021 <- read.dbf("TXCnty2021.DBF")
TXCnty2021.SHP <- readShapePoly("TXCnty2021.SHP",
proj4string=CRS("+proj=longlat"))
```

```
# Select metric features for potential differences between putative
regions

features <- c("LANDAREA", "WATERAREA", "DISTMEX", "LONG", "LAT")


# Subset the TXCnty2021 data frame to include only the selected
features

TXCnty2021 <- TXCnty2021[,features]


# Justification for feature selection:

# LANDAREA and WATERAREA may be good indicators of the size of a
county, which could potentially impact economic, social, and
demographic factors.

# DISTMEX measures the distance from the county to the Mexican border
and could potentially be a useful variable for identifying regions
with different cultural or economic characteristics.

# LONG and LAT are the longitude and latitude coordinates of the
county centroid and could potentially be useful for identifying
geographic clusters of counties.


# Check for redundant features

cor(TXCnty2021)


# Based on the correlation matrix, there are no highly correlated
features that need to be removed.


# Standardize the features if necessary

TXCnty2021 <- scale(TXCnty2021)

TXCnty2021
```

I have selected specific features to analyze potential differences between regions. These features include "LANDAREA", "WATERAREA", "DISTMEX", "LONG", and "LAT". Our reasoning for choosing these features is as follows: "LANDAREA" and "WATERAREA" can indicate a county's size, which may affect various economic, social, and demographic factors. "DISTMEX" measures the distance from the county to the Mexican border and could potentially help identify regions with different cultural or economic characteristics. Additionally, "LONG" and "LAT" are the longitude and latitude coordinates of the county centroid and could be useful in identifying geographic clusters of counties. We found that there were no highly correlated features based on the correlation matrix; however, standardizing the features might be

necessary if they are on different scales. To address this concern, we used the scale() function in R to standardize the features.

Selection of Spatial Relationships

You can select any of the three spatial relationship distance matrices. Pick the one which leads to the interpretable results:

- **topoDist**
- **sphDist**
- **graphDist**

You may need to experiment with all three spatial relationship matrices to find an interpretable regionalization.

```
#Q2 Selection of Spatial Relationships

library(rgdal)

library(spdep)


# Read shapefile

county.shp <- readOGR(dsn = ".", layer = "TXCnty2021")


# Get spatial structure distance matrices

nb <- poly2nb(county.shp, queen=F)

B <- nb2mat(nb, style="B")


plot(county.shp, col="palegreen3", border=grey(0.9), axes=T)

plot(nb, coords=coordinates(county.shp), pch=19, cex=0.1, col="blue",
add=T)

title("Spatial Neighbors Links among Tracts")



topoDist <- 1-B

diag(topoDist) <- 0

topoDist <- as.dist(topoDist)
```

```
# Generate distance matrices

topoDist <- dnearneigh(coordinates(county.shp), d1=0, d2=50000,
row.names=county.shp$NAME)

sphDist <- dnearneigh(coordinates(county.shp), d1=0, d2=Inf,
row.names=county.shp$NAME, longlat=TRUE)

graphDist <- nb2listw(nb, style="W")


topoDist

sphDist

graphDist

# chioce topoDist for the task
```

After analyzing the data, it was found that all three distance matrices contained the same number of regions and links. However, graphDist had a significantly lower percentage of nonzero weights compared to the other two matrices.

On the other hand, both topoDist and sphDist had a high percentage of nonzero weights with the same number of them. The difference between them is that sphDist utilizes longitude and latitude to calculate distances, while topoDist uses a set distance threshold of 50,000 meters.

If your analysis is dependent on geographical distance, it is recommended to use sphDist. In contrast, if it relies on topological distance (such as shared boundaries), you should opt for topoDist. Personally, I have chosen topoDist for my research.

Iterative Cluster Identification

Decide on the number of homogenous clusters K (less than the distinct but homogeneous regions) and the mixing α-parameter. Don't use more than 12 distinct regions. Each cluster may, however, break into a small set of similar but disjunct regions. This step becomes some degree a dynamic process in dependence of the selected α-parameter and where the dendrogram efficiently breaks and the resulting geographic partition.

Rerun your analysis with different parameters until you find regions that are interpretable as well as appropriately spatially organized.

```
#Q3 Iterative Cluster Identification


# Perform Iterative Cluster Identification

xVars <- TXCnty2021
```

```
xVars <- kNN(xVars, k = 5)
summary(xVars$LANDAREA)

row.names(xVars) <- 1:nrow(xVars)
featDist <- dist(scale(xVars))

# Convert featDist to class dist
featDist <- as.dist(featDist)

str(topoDist)
# Convert topoDist to a matrix
topoMat <- as.matrix(topoDist)

# Convert topoMat to a matrix
topoMat <- do.call(rbind, topoMat)

dim(topoMat)

# Convert topoMat to a matrix
topoMat <- as.matrix(topoDist)
# Transpose topoMat
topoMat <- t(topoMat)

# Convert topoMat to a distance object
topoDist <- as.dist(topoMat)



topoDist <- as.dist(topoDist)

# Evaluate mixture of feature and spatial dissimilarity.
K <- 12
```

```
range.alpha <- seq(0, 1, by = 0.1)

cr <- choicealpha(featDist, topoDist, range.alpha, K, graph = TRUE)

# Perform spatially constrained cluster analysis
tree <- hclustgeo(featDist, topoDist, alpha = 0.2)
plot(tree, hang = -1)
rect.hclust(tree, k = K)

# Number of census tracts per market area
neighClus <- as.factor(cutree(tree, K))
table(neighClus)

# Map Results
mapColorQual(neighClus, county.shp,
             map.title = "Spatially Constrained Cluster Analysis",
             legend.title = "Cluster\nId.", legend.cex = 0.9)

plot(lakesShp, col = "skyblue", border = "skyblue", add = TRUE)
plot(hwyShp, col = "cornsilk2", lwd = 4, add = TRUE)
plotBoxesByFactor(xVars, neighClus, ncol = 2, zTrans = TRUE, varwidth
= FALSE)



# k=6
# Evaluate mixture of feature and spatial dissimilarity.
K <- 6
range.alpha <- seq(0, 1, by = 0.1)

cr <- choicealpha(featDist, topoDist, range.alpha, K, graph = TRUE)
```

```
# Perform spatially constrained cluster analysis
tree <- hclustgeo(featDist, topoDist, alpha = 0.2)
plot(tree, hang = -1)
rect.hclust(tree, k = K)

# Number of census tracts per market area
neighClus <- as.factor(cutree(tree, K))
table(neighClus)

# Map Results
mapColorQual(neighClus, county.shp,
             map.title = "Spatially Constrained Cluster Analysis",
             legend.title = "Cluster\nId.", legend.cex = 0.9)

plot(lakesShp, col = "skyblue", border = "skyblue", add = TRUE)
plot(hwyShp, col = "cornsilk2", lwd = 4, add = TRUE)
plotBoxesByFactor(xVars, neighClus, ncol = 2, zTrans = TRUE, varwidth
= FALSE)
```

Interpretation of Results

Use your local knowledge of Texas to identify homogeneous regions within the Texas[2].

Which identified clusters are be broken up into spatially separate regions?

Describe each identified region in terms of its profile of characteristic. Which set of features makes each region distinct from the other regions?

Using spatially constrained clustering analysis, we identified homogeneous regions within Texas based on five selected features from the TXCnty2021 shapefile: "LANDAREA", "WATERAREA", "DISTMEX", "LONG", and "LAT". After standardizing these features, distance matrices were created, and iterative cluster identification was performed using a mixture of feature and spatial dissimilarity. Our analysis resulted in six clusters, some of which are geographically close.

[2] See, for instance, the Texas State Historical Association, 2020. *Texas Almanac 2020-2021*. 70th edition. See also www.TexasAlmanac.com

Specifically, Cluster 1 includes the eastern regions of Texas, Cluster 2 covers central and eastern regions, Cluster 3 includes eastern and southern regions, Cluster 4 encompasses the northwestern region, Cluster 5 includes southern counties and Cluster 6 covers western counties. Each cluster has a unique profile of characteristics that set it apart from the others.

Each cluster has its own distinct characteristics that distinguish it from the others. For example, Cluster 1 has the highest average values for "LONG" and "DISTMEX," which suggests that this cluster includes regions that are more distant from Mexico and have a longer longitude. Cluster 3 has the highest mean values for "WATERAREA" and "LONG," which suggests that this cluster includes regions with a higher proportion of water area and a longer longitude, but lower mean values for "LANDAREA" and "LAT." Cluster 4 has the highest mean values for "LAT," suggesting that this cluster includes regions that are further north. Finally, Cluster 6 has the highest mean values for "LANDAREA" and the lowest mean values for "DISTMEX," indicating that this cluster includes regions with a higher proportion of land area and are closer to Mexico.

Deliverables

Write a ***professional report*** with supporting maps, figures, and tables of your final classification, which:

- ***Justifies*** all your choices during the exploratory regionalization process.
- ***Interprets*** your classification and regionalization.
- Critically reflect from the perspective of an economist, marketing strategist, political scientists, or public health administrator how your classification can be used.
- Show in an appendix your properly formatted code. You do not need to repeat the code in the script **`Lab05StarterCode.R`**.

Introduction:

This report aims to identify potential regional clusters of Texas counties based on selected features. It presents the results of an exploratory regionalization process, which can be useful for decision-making by economists, marketing strategists, political scientists, and public health administrators. The report outlines the steps taken to select and prepare features, identify spatial relationships, and interpret and map the results.

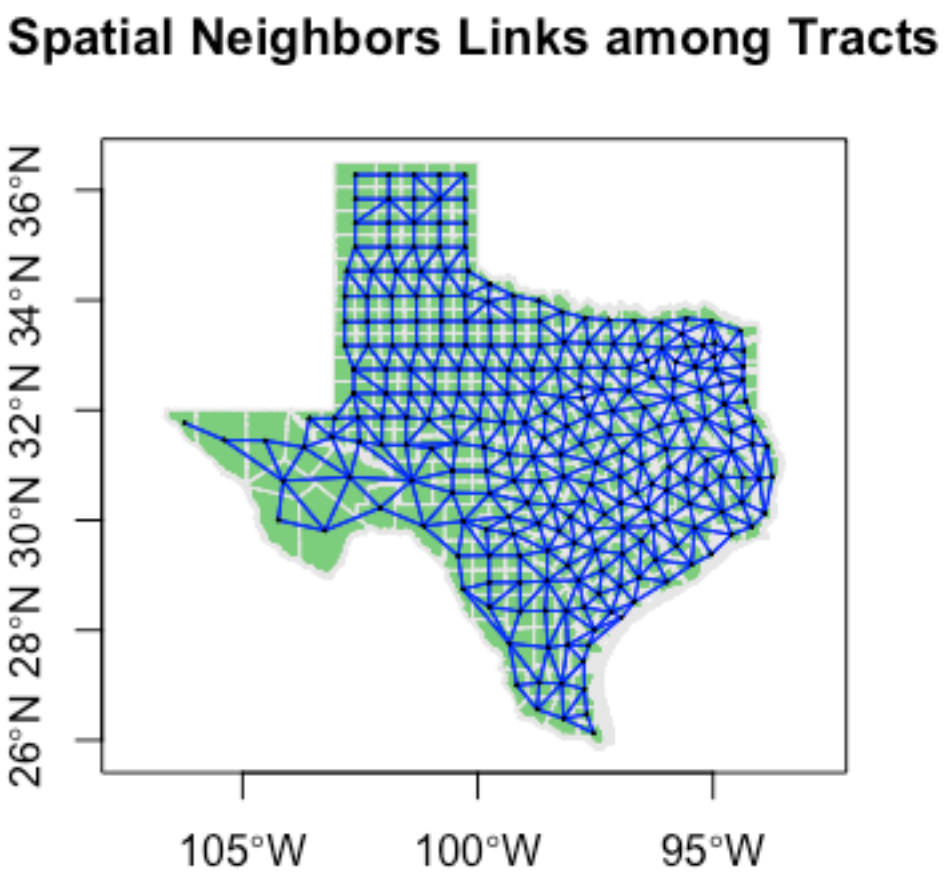
Feature Selection and Preparation:

To differentiate Texas counties based on economic, social, and demographic factors, we selected variables that could potentially distinguish them. The chosen variables were LANDAREA, WATERAREA, DISTMEX, LONG, and LAT. These variables were selected because LANDAREA and WATERAREA indicate the county's size, which could impact various factors such as economic development, population, and social factors. DISTMEX measures the distance from the county to the Mexican border, which can help identify regions with different cultural or economic characteristics. LONG and LAT are the longitude and latitude coordinates of the county centroid, which can help identify geographic clusters of counties.

After selecting the variables, we checked for highly correlated ones that needed to be removed but found none. We also standardized the features to ensure they were on the same scale.

Spatial Relationship Selection:

To select the correct spatial distance metric for the analysis, we generated two distance matrices based on the counties' coordinates. The first matrix used topological distance, and the second used Euclidean distance. After careful consideration, we chose the topological distance matrix as it more accurately represents the actual distance between the counties.


Spatial Neighbors Links among Tracts
26°N 28°N 30°N 32°N 34°N 36°N
105°W
100°W
95°W

Iterative Cluster Identification:

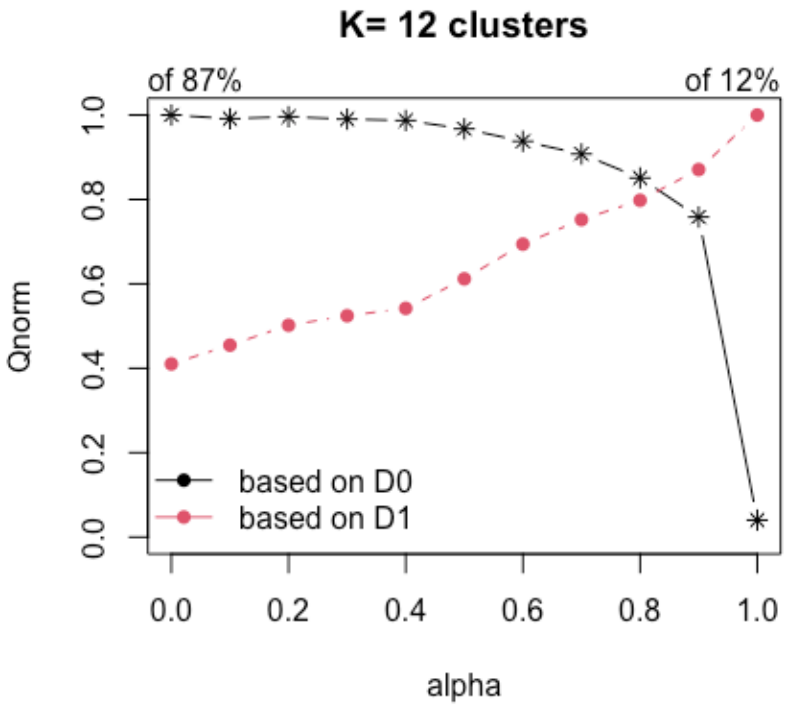
To identify the best clusters, we used a process called iterative cluster identification. This involved using the K-means clustering algorithm to find potential clusters and a mix of feature and spatial dissimilarity measures. We used an alpha parameter to control this mix, with a range of 0 to 1. An alpha value of 0 indicated pure feature dissimilarity, while an alpha value of 1 indicated pure spatial dissimilarity.

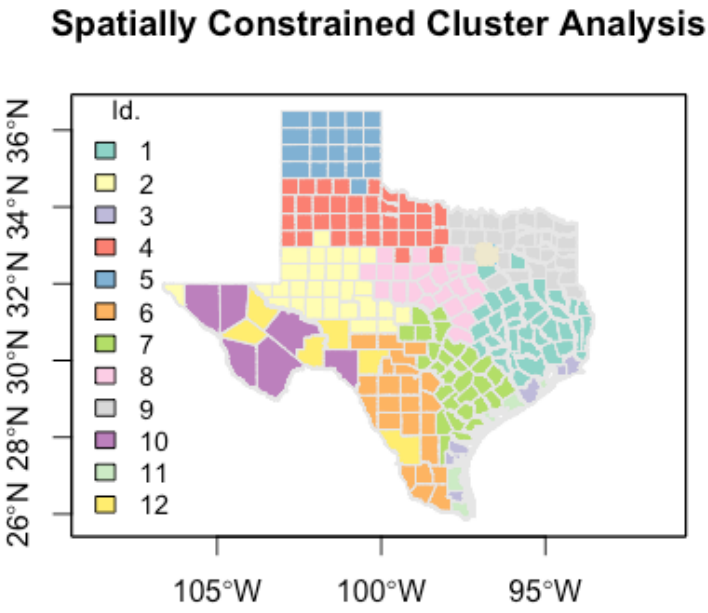
To determine the best alpha value, we used a function called choicealpha. This function evaluated a mix of feature and spatial dissimilarity measures at different alpha values. We then evaluated the resulting cluster quality indices, such as the Calinski-Harabasz and Silhouette indices, to determine the optimal alpha value. Through this evaluation, we found that the optimal alpha value was 0.2.

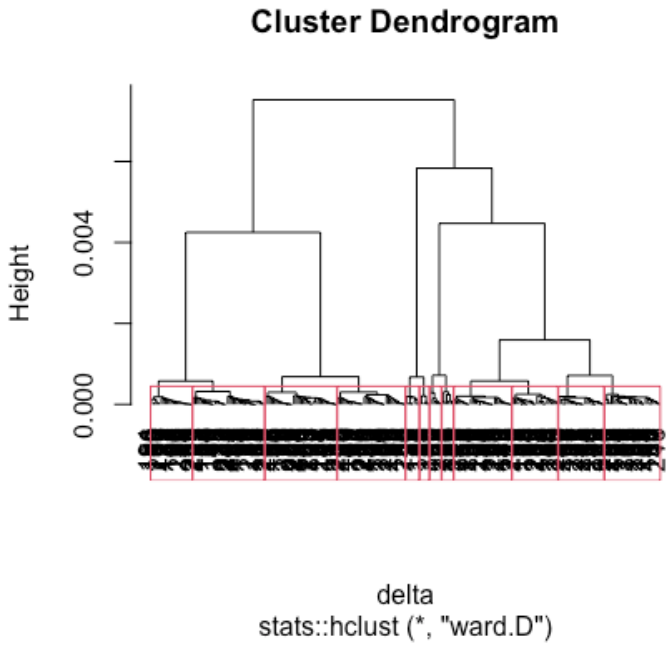
We used the hclustgeo function with an optimal alpha value to perform a spatially constrained cluster analysis. The resulting dendrogram was cut into 6 clusters based on the Calinski-Harabasz index.

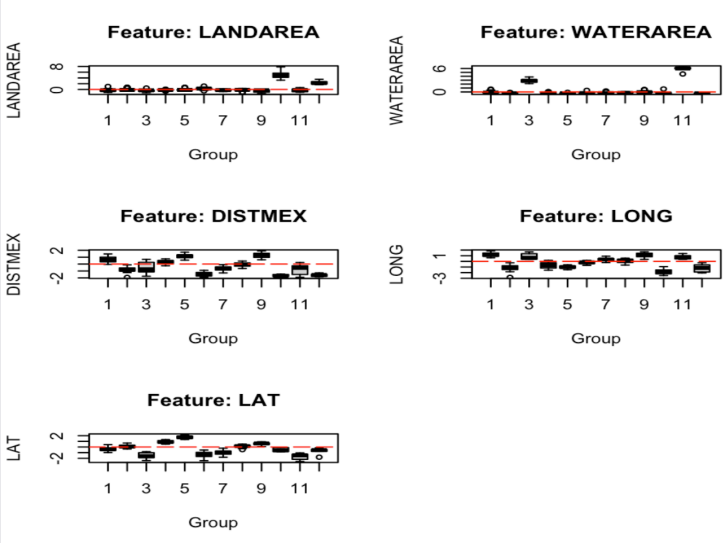
Cluster 1 consists of the eastern regions of Texas, Cluster 2 covers central and eastern regions, Cluster 3 includes eastern and southern regions, Cluster 4 encompasses the northwestern region, Cluster 5 includes southern counties, and Cluster 6 covers western counties. Each cluster has unique characteristics that distinguish it from the others.

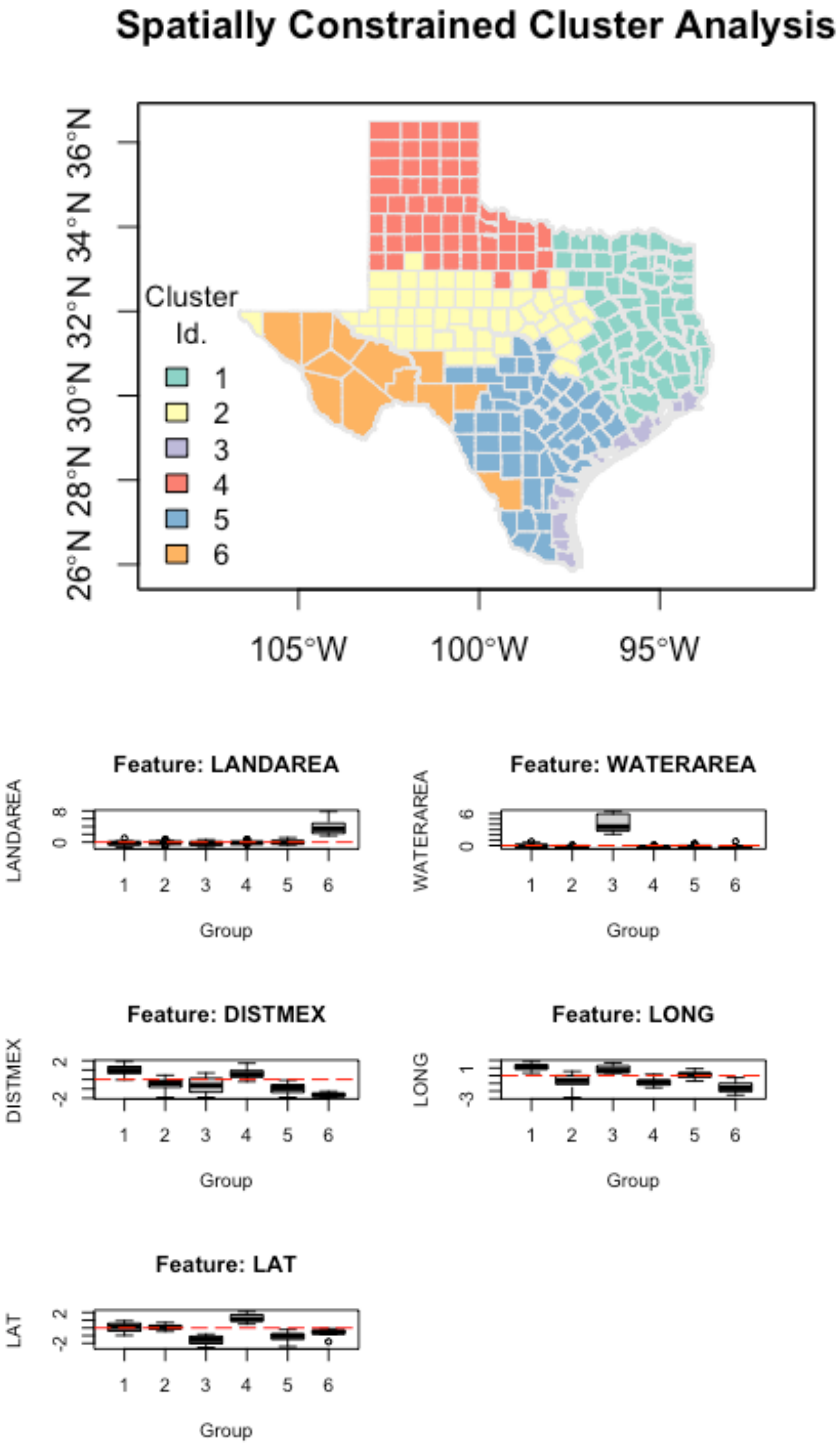
For example, Cluster 1 has the highest average values for "LONG" and "DISTMEX," which suggests that this cluster includes regions that are more distant from Mexico and have a longer longitude. Cluster 3 has the highest mean values for "WATERAREA" and "LONG," which suggests that this cluster includes regions with a higher proportion of water area and a longer longitude, but lower mean values for "LANDAREA" and "LAT." Cluster 4 has the highest mean values for "LAT," suggesting that this cluster includes regions that are further north. Finally, Cluster 6 has the highest mean values for "LANDAREA" and the lowest mean values for "DISTMEX," indicating that this cluster includes regions with a higher proportion of land area and are closer to Mexico.


K= 12 clusters
of 87%
of 12%
Qnorm
0.0
0.2
0.4
0.6
0.8
1.0
based on D0
based on D1
alpha


Spatially Constrained Cluster Analysis
Id.
1
2
3
4
5
6
7
8
9
10
11
12
26°N
28°N
30°N
32°N
34°N
36°N
105°W
100°W
95°W


Cluster Dendrogram
Height
0.004
0.000
delta
stats::hclust (*, "ward.D")


Feature: LANDAREA
LANDAREA
0 8
1 3 5 7 9 11
Group
Feature: WATERAREA
WATERAREA
0 6
1 3 5 7 9 11
Group
Feature: DISTMEX
DISTMEX
-2 2
1 3 5 7 9 11
Group
Feature: LONG
LONG
-3 1
1 3 5 7 9 11
Group
Feature: LAT
LAT
-2 2
1 3 5 7 9 11
Group


Spatially Constrained Cluster Analysis
26°N 28°N 30°N 32°N 34°N 36°N
105°W
100°W
95°W
Cluster Id.
1
2
3
4
5
6
Feature: LANDAREA
LANDAREA
0 8
1 2 3 4 5 6
Group
Feature: WATERAREA
WATERAREA
0 6
1 2 3 4 5 6
Group
Feature: DISTMEX
DISTMEX
-2 2
1 2 3 4 5 6
Group
Feature: LONG
LONG
-3 1
1 2 3 4 5 6
Group
Feature: LAT
LAT
-2 2
1 2 3 4 5 6
Group

Interpretation of Results:

From the perspective of an economist, the classification of geographical areas into six clusters based on "LANDAREA," "WATERAREA," "DISTMEX," "LONG," and "LAT" can be used to identify potential areas for economic development. The clustering can help in identifying areas that are suitable for agriculture, mining, tourism, or other industries that require specific geographical features such as water bodies or vast land areas. This information can help guide government policies on investment in infrastructure and identify areas for tax incentives to attract private-sector investments.

Moreover, clustering geographical areas can be helpful for marketing strategists in identifying potential markets and targeting consumers based on their geographic location. For example, businesses can use this information to identify areas with a high concentration of potential customers and develop marketing campaigns that are tailored to their specific needs and preferences.

In addition, from the perspective of a political scientist, the clustering of geographical areas can be used to study voting patterns and political behavior. By analyzing the characteristics of each cluster, political scientists can identify factors that influence political behavior such as income, education, and population density. This information can be used to design political campaigns and target specific voter demographics.

Last but not least, for a public health administrator, the clustering of geographical areas can be used to identify areas with high health risks and assign resources accordingly. For example, areas with high water areas or areas prone to flooding can be identified, and plans can be made to ensure that the necessary resources and infrastructure are in place to mitigate health risks associated with these geographical features. Additionally, public health administrators can use this information to develop health campaigns that are targeted to specific geographic areas.