# 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)](https://www.investopedia.com/terms/d/demographics.asp)). 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]](#footnote-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)](http://www.spatialfiltering.com/ThinkR/Downloads/VacModelTX.html).

## 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)](https://arxiv.org/pdf/1707.03897.pdf))

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 -parameter and at a given number of clusters .

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

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.

## 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.

## Iterative Cluster Identification

Decide on the number of homogenous clusters (less than the distinct but homogeneous regions) and the mixing -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 -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.

## Interpretation of Results

Use your local knowledge of Texas to identify homogeneous regions within the Texas[[2]](#footnote-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?

## 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**.

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**. [↑](#footnote-ref-1)
2. See, for instance, the Texas State Historical Association, 2020. *Texas Almanac 2020-2021*. 70th edition. See also [www.TexasAlmanac.com](http://www.TexasAlmanac.com) [↑](#footnote-ref-2)