# Lab02: Hierarchical Clustering under Spatial Constraints.

Handed out: March 11, 2021

Due date: April 1, 2021, as Word document into eLearning’s **Lab02Submit** link.

Grading: Lab02 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 and [d] residential characteristics. Partition the counties of Texas into homogeneous regions. 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)).

## Data

Use the data in the zipped file **TXCnty2020.zip**. The script **Lab02StarterCode.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 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 **TXCnty2020** 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 and the mixing -parameter. Don’t use more than 12 distinct cluster. Each cluster may, however, break into a set of similar but disjunct regions. This step is to some degree is 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 clusters 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[[1]](#footnote-1).

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

Describe each identified cluster in terms of its characteristic profile of feature. Which set of features makes each cluster distinct?

## 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.
* Perhaps, contrasts your “best” classification against your “second best” classification.
* 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 TX**Lab02StarterCode.R** .

1. 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-1)