# Lab02: Hierarchical Clustering under Spatial Constraints.

Handed out: October 2, 2019

Due date: October 16, 2019, as hard copy

Grading: Lab02 counts for 12 % of your final grade.

## Objective

Partition Dallas County based on its census tracts into homogeneous neighborhoods and districts. Use the features of the census and their spatial relationships to identify these districts. Properly interpret the final regionalization.

## Data

Use the data in the package **DallasTracts**. The script **Lab02StarterCode.R** provides information on how to calculate varying forms of geographic relationships.

## Reading

Study the M. Chavent, V. Kuentz-Simonet, A. Labenne, J. Saracco. *ClustGeo:* *An R package for hierarchical clustering with spatial constraints* arXiv:1707.03897

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 and the geographic cohesion of the partitions that are obtained for varying -parameter and a given number of clusters .

## Feature Selection and Preparations

Identify the metric features that describe potential differences between putative district clusters, which are expected to exhibit a strong degree of internal homogeneity.

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

Verbally 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 **tractShp** spatial data-frame?

Evaluate whether you need standardize the features and whether you need combine them into latent features of component scores.

## Selection of Spatial Relationships

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

1. **geoDist <- as.dist(stepDist)**
2. **geoDist <- as.dist(sphDist)**
3. **geoDist <- 1-B; diag(geoDist) <- 0; geoDist <- as.dist(geoDist)**

It is preferred to experiment with all three spatial relationship matrices.

## Iterative Cluster Identification

Decide on the number of geographic clusters and the mixing -parameter. Don’t use more than 12 distinct cluster. This to some degree is a dynamic process in dependence of the selected -parameter and where the dendrogram efficiently breaks and the resulting geographic partition.

## Interpretation of Results

Use your local knowledge of Dallas County to identify homogeneous neighborhoods and districts within the county. A tolerable, preferably interpretable, spatial fragmentation of the identify homogenous district clusters is acceptable.

Which identified clusters can be broken up into spatially separate districts?

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

Perform a discriminant analysis using the set of selected features to evaluate the quality of your resulting cluster analysis and interpret the result.

## Deliverables

Write a ***professional report*** with supporting maps, figures and tables which:

* ***Justifies*** all your choices during the exploratory regionalization process.
* Perhaps, contrasts your “best” results against your “second best” results.
* Critically reflects from the perspective of a residential real estate broker on how the methodology of identifying homogeneous residential districts in Dallas County can be improved.
* Show in an appendix your properly formatted code. You do not need to repeat the code in the script **Lab02StarterCode.R** .