Clustering Lab

Use this lab time to gain experience with clustering in R / Python. ULTIMATE GOAL: try to find a robust clustering scheme for various Canadian health regions.

Dataset

HR_2016_Census_simple.xlsx

Problem Description

The population of Canada is divided physically into provincial and territorial areas, most of which are further subdivided into **health regions**.

Census information (from 2016) is available for those health regions

- https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1710012201
- https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1710012301

The equivalent 2018 dataset has been clustered to produce **peer groups**: the result is shown at

- https://www150.statcan.gc.ca/n1/pub/82-402-x/2018001/maps-cartes/rm-cr14-eng.htm

In this lab, you will cluster the 2016 dataset in various ways, and compare your results with those obtained in the map above (the goal is not to re-create the map).

1. Load the data and **summarize/visualize** it (extract the rows with a 4-digit geocode).

2. Clean the data and **impute** missing values (if necessary). **Scale** the data and assign to a new set.

3. Run the ***k*-means** algorithm on the scaled data, using ALL the features, for $k = 3, \ldots, 10$. Use the *Davies-Bouldin* index and the *Within-SS* index to determine the optimal number of clusters. Is the optimal clustering scheme plausible?

4. Reduce the dimension of the dataset by running a **principal component analysis** (PCA) and keep the principal components that explain up to 80% of the variability in the data. Repeat step 3. Are the results significantly different?

5. Write a routine that selects a number of features, a set of features, and a clustering algorithm **randomly** (k-means, DBSCAN, hierarchical clustering, spectral clustering, etc., with random parameter choices), and that produces and records a cluster assignment for each health region, together with some internal cluster metrics (of your choosing).

6. Run your routine a number of times (50? 100? 200? 500?) on the scaled (but not PCA-reduced) dataset. Produce a similarity matrix that measures the percentage of times that each pair of health regions are in the same cluster. Are there health regions that seem to find themselves in the same cluster more than 95% of the time? 90%? 80%? 50%? 0?

7. Based on your results, provide a list of health regions which the data seem to indicate are **true peers**. Compare with the map. Are the results surprising?