Cluster Analysis and COVID-19 Immigration Policies

Thu Pham

thupham@college.harvard.edu
Harvard Statistics Department

03 May 2022

Introduction

- COVID-19 and the onset of many travel restrictions
- What makes immigration policies "similar?"
- Interesting to find similarities between countries with "similar" policies

Introduction

- COVID-19 and the onset of many travel restrictions
- What makes immigration policies "similar?"
- Interesting to find similarities between countries with "similar" policies

Research Question: Using cluster analysis, can we find demographic patterns in countries' COVID-19 immigration policies?

Clustering Methods

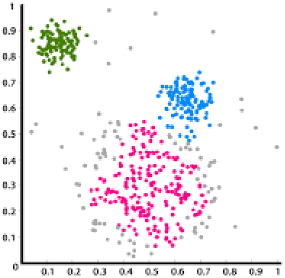
- Cluster analysis: unsupervised machine learning method to group observations, with little to no prior knowledge of what the groups should look like

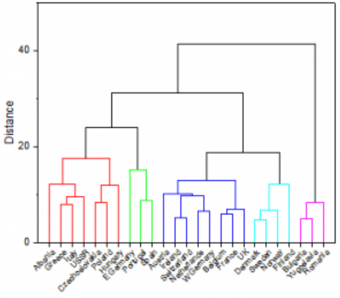
Clustering Methods

- Cluster analysis: unsupervised machine learning method to group observations, with little to no prior knowledge of what the groups should look like
- Common application: market analysis

Clustering Methods

- Cluster analysis: unsupervised machine learning method to group observations, with little to no prior knowledge of what the groups should look like
- Common application: market analysis





K-Means

- Hartigan-Wong method: less prone to converge to a local optima [1]

[1] Morissette, Laurence and Sylvain Chartier. The k-means clustering technique: General considerations and implementation in Mathematica

K-Means

- Hartigan-Wong method: less prone to converge to a local optima [1]
- Randomly initializes point to K clusters

[1] Morissette, Laurence and Sylvain Chartier. The k-means clustering technique: General considerations and implementation in Mathematica

K-Means

- Hartigan-Wong method: less prone to converge to a local optima [1]
- Randomly initializes point to K clusters
- Repeat until clusters converge:

[1] Morissette, Laurence and Sylvain Chartier. The k-means clustering technique: General considerations and implementation in Mathematica

K-Means

- Hartigan-Wong method: less prone to converge to a local optima [1]
- Randomly initializes point to K clusters
- Repeat until clusters converge:
  - Calculate the within-cluster sum of squares error

$$SSE = \sum_{k}^{K} \sum_{x_i \in c_k} (x_i - \mu_k)^2$$

[1] Morissette, Laurence and Sylvain Chartier. The k-means clustering technique: General considerations and implementation in Mathematica

K-Means

- Hartigan-Wong method: less prone to converge to a local optima [1]
- Randomly initializes point to K clusters
- Repeat until clusters converge:
  - Calculate the within-cluster sum of squares error

$$SSE = \sum_{k}^{K} \sum_{x_i \in c_k} (x_i - \mu_k)^2$$

  - Re-assign x_i to the cluster c_ℓ that has the lowest SSE with the inclusion of x_i

[1] Morissette, Laurence and Sylvain Chartier. The k-means clustering technique: General considerations and implementation in Mathematica

Hierarchical Agglomerative Clustering (HAC)

- Deterministic, and does not require a choice of K clusters ahead of time

[2] Saraçli, Sinan, et al.: Comparison of hierarchical cluster analysis methods by cophenetic correlation

Hierarchical Agglomerative Clustering (HAC)

- Deterministic, and does not require a choice of K clusters ahead of time
- Each observation belongs to its own cluster

[2] Saraçli, Sinan, et al.: Comparison of hierarchical cluster analysis methods by cophenetic correlation

Hierarchical Agglomerative Clustering (HAC)

- Deterministic, and does not require a choice of K clusters ahead of time
- Each observation belongs to its own cluster
- Repeat until we have a single cluster:
  - Closest clusters are merged (by some linkage and metric criteria)
  - Clustering at each step is recorded

[2] Saraçli, Sinan, et al.: Comparison of hierarchical cluster analysis methods by cophenetic correlation

Hierarchical Agglomerative Clustering (HAC)

- Deterministic, and does not require a choice of K clusters ahead of time
- Each observation belongs to its own cluster
- Repeat until we have a single cluster:
  - Closest clusters are merged (by some linkage and metric criteria)
  - Clustering at each step is recorded
- Construct and cut dendrogram

[2] Saraçli, Sinan, et al.: Comparison of hierarchical cluster analysis methods by cophenetic correlation

Hierarchical Agglomerative Clustering (HAC)

- Deterministic, and does not require a choice of K clusters ahead of time
- Each observation belongs to its own cluster
- Repeat until we have a single cluster:
  - Closest clusters are merged (by some linkage and metric criteria)
  - Clustering at each step is recorded
- Construct and cut dendrogram
- Choosing a linkage criteria with the cophenetic correlation [2]:

$$c = \frac{\sum_{i<j}[d(x_i, x_j) - \bar{d}][t(x_i, x_j) - \bar{t}]}{\sqrt{\sum_{i<j}[x(i,j) - \bar{x}]^2 \sum_{i<j}[t(i,j) - \bar{t}]^2}}$$

[2] Saraçli, Sinan, et al.: Comparison of hierarchical cluster analysis methods by cophenetic correlation

Choosing the Number of Clusters

Gap statistic, the difference of total intra-cluster variation between observed data and reference data [3]

[3] Tibshirani, Robert et al. Estimating the number of clusters in a data set via the gap statistic.

Choosing the Number of Clusters

Gap statistic, the difference of total intra-cluster variation between observed data and reference data [3]

- For K clusters, calculate the intra-cluster variation:

$$W_K = \sum_{k=1}^{K} \frac{1}{2n_k} \sum_{i,i' \in C_k} d_{i,i'},$$

[3] Tibshirani, Robert et al. Estimating the number of clusters in a data set via the gap statistic.

Choosing the Number of Clusters

Gap statistic, the difference of total intra-cluster variation between observed data and reference data [3]

- For K clusters, calculate the intra-cluster variation:

$$W_K = \sum_{k=1}^{K} \frac{1}{2n_k} \sum_{i,i' \in C_k} d_{i,i'},$$

- Generate N reference distributions and cluster

[3] Tibshirani, Robert et al. Estimating the number of clusters in a data set via the gap statistic.

Choosing the Number of Clusters

Gap statistic, the difference of total intra-cluster variation between observed data and reference data [3]

- For K clusters, calculate the intra-cluster variation:

$$W_K = \sum_{k=1}^{K} \frac{1}{2n_k} \sum_{i,i' \in C_k} d_{i,i'},$$

- Generate N reference distributions and cluster
- Compute the gap statistic:

$$\text{Gap}(K) = \frac{1}{N} \sum_{n=1}^{N} [\log(W_{K,n}) - \log(W_K)]$$

[3] Tibshirani, Robert et al. Estimating the number of clusters in a data set via the gap statistic.

Choosing the Number of Clusters

Gap statistic, the difference of total intra-cluster variation between observed data and reference data [3]

- For K clusters, calculate the intra-cluster variation:

$$W_K = \sum_{k=1}^{K} \frac{1}{2n_k} \sum_{i,i' \in C_k} d_{i,i'},$$

- Generate N reference distributions and cluster
- Compute the gap statistic:

$$\text{Gap}(K) = \frac{1}{N} \sum_{n=1}^{N} [\log(W_{K,n}) - \log(W_K)]$$

- Choose the number of clusters as the smallest value of k such that $Gap(k) \geq Gap(k+1) - \sigma(k+1)$.

[3] Tibshirani, Robert et al. Estimating the number of clusters in a data set via the gap statistic.

Methods of Analysis

- Multiple sample T-test (ANOVA) across each chosen demographic factor

Methods of Analysis

- Multiple sample T-test (ANOVA) across each chosen demographic factor
- Rand index to compare clustering

$$R = \frac{a+b}{\binom{n}{2}},$$

where for a partition X and Y of some set S of n elements, a is the number of pairs in S that are in the same subset in both X and Y, and b is the number of pairs that are in different subsets in X and Y

Data and Data Cleaning

- COVID Border Accountability Project
- Variables Visa bans, history of travel bans, citizen bans, policy length, policy type, travel blockage (air, land, sea), refugee bans, country exceptions, work exceptions

Data and Data Cleaning

- COVID Border Accountability Project
- Variables Visa bans, history of travel bans, citizen bans, policy length, policy type, travel blockage (air, land, sea), refugee bans, country exceptions, work exceptions
- Data cleaning: NA values, one-hot encoding, assumptions, aggregating by country

Data and Data Cleaning

- COVID Border Accountability Project
- Variables Visa bans, history of travel bans, citizen bans, policy length, policy type, travel blockage (air, land, sea), refugee bans, country exceptions, work exceptions
- Data cleaning: NA values, one-hot encoding, assumptions, aggregating by country
- World Bank data: GDP, population, life expectancy, fertility rate, and adult literacy rate (2020)

Chosen Hyperparameters

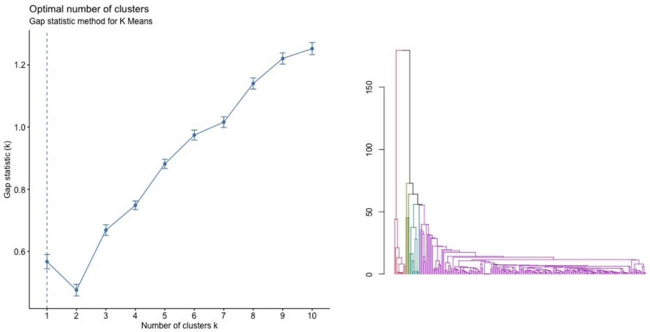


Figure: Gap statistic for K-Means (left), final dendrogram (right).

The cophonetic correlation was highest for minimum linkage (0.862).

ANOVA and Rand Index

Demographic	K-Means	HAC
GDP	0.485	0.334
Population	0.155	0.984
Life Expectancy	**0.00542**	**0.039**
Fertility Rate	**0.0067**	0.089
Literacy Rate	**0.0273**	0.149

Table: Significant p-values are bolded.

Policy Clustering	Continent	Development Level
K-Means	0.640	0.612
HAC	0.300	0.342

Table: Another interesting result: the rand index for K-Means vs HAC was 0.405.

ANOVA

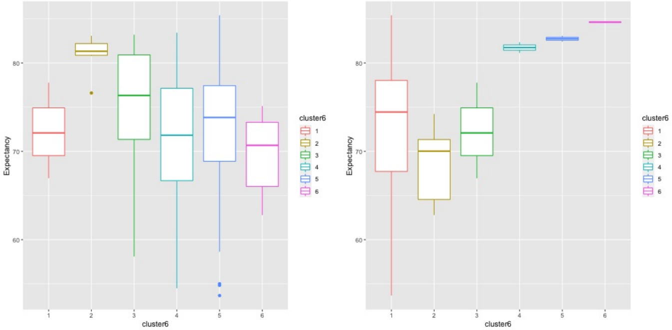


Figure: K-Means (left) and HAC (right)

ANOVA

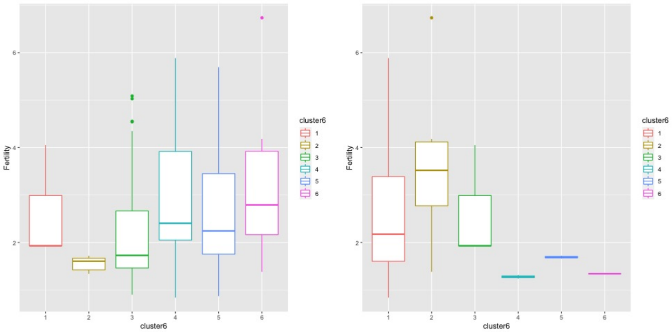


Figure: K-Means (left) and HAC (right)

Rand Index

- Minimum distance linkage $\implies$ stringier clusters; may merge clusters whose centroids are far apart

Rand Index

- Minimum distance linkage $\implies$ stringier clusters; may merge clusters whose centroids are far apart
- K-Means are more compact; takes into account the size of the clusters and the internal variance

Rand Index

- Minimum distance linkage $\implies$ stringier clusters; may merge clusters whose centroids are far apart
- K-Means are more compact; takes into account the size of the clusters and the internal variance
- K-Means had more similar clustering to our "natural" metrics – why?
  - Continents: 43, 41, 50, 15, 40
  - Development level: 33, 29, 48, 61

Conclusion

- More effective visualizations

Conclusion

- More effective visualizations
- Some clusters seem "intuitive:" Belgium, Denmark, Greece, Iceland, Poland, and Sweden
- Others, not so much: Iraq, United States, Egypt, Mexico

Conclusion

- More effective visualizations
- Some clusters seem "intuitive:" Belgium, Denmark, Greece, Iceland, Poland, and Sweden
- Others, not so much: Iraq, United States, Egypt, Mexico
- Impact of choosing cluster method