Cluster Analysis and COVID-19 Immigration Policies

Thu Pham

thupham@college.harvard.edu
Harvard Statistics Department

20 April 2022

Introduction

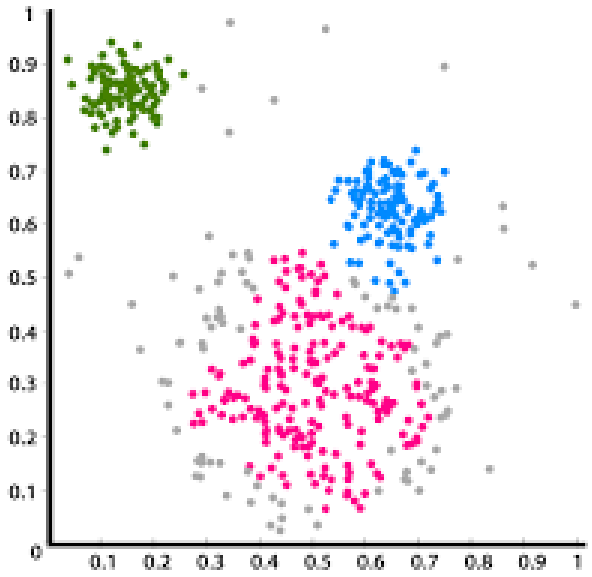
- Cluster analysis: unsupervised machine learning method to group observations, with little to no prior knowledge of what the groups should look like

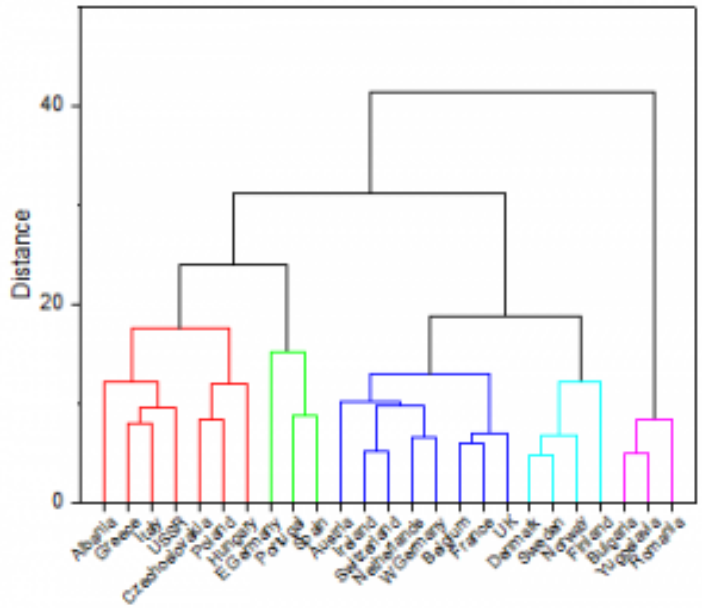
Introduction

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

Introduction

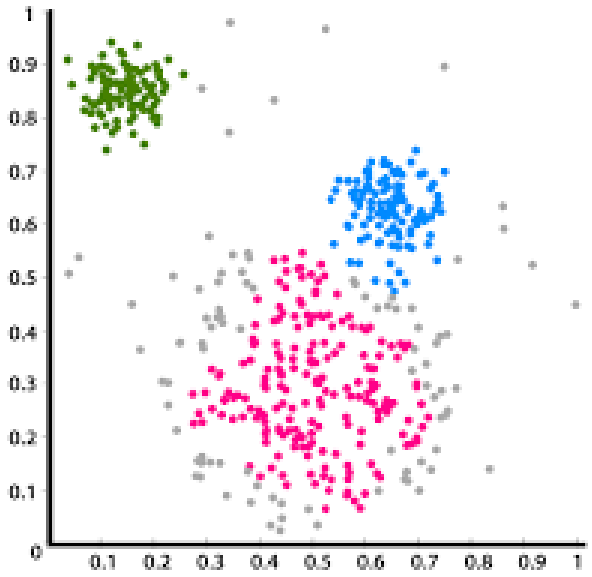
- Cluster analysis: unsupervised machine learning method to group observations, with little to no prior knowledge of what the groups should look like
- Common applications: image research and market applications

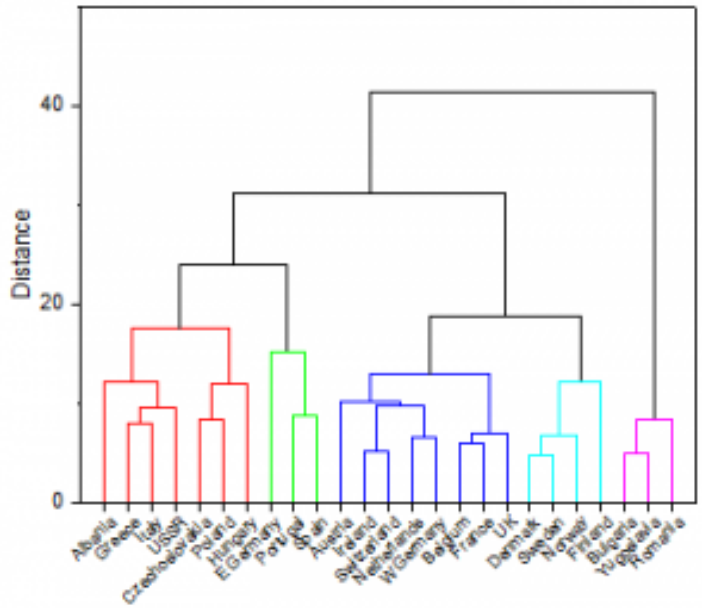




Introduction

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





Research Question: Can we find patterns in the cluster analyses of countries' COVID-19 immigration policies and their demographic characteristics?

Literature Review

- In general, clustering is not as common in social sciences – typically used in sociology (social groups)

Literature Review

- In general, clustering is not as common in social sciences – typically used in sociology (social groups)
- Latent class models, Kamila, K-prototypes typically performed best in heterogenous data settings

Literature Review

- In general, clustering is not as common in social sciences – typically used in sociology (social groups)
- Latent class models, Kamila, K-prototypes typically performed best in heterogenous data settings
- Variable selection significantly affect clustering – methods to select variables based on their "clusterability"

Literature Review

- In general, clustering is not as common in social sciences – typically used in sociology (social groups)
- Latent class models, Kamila, K-prototypes typically performed best in heterogenous data settings
- Variable selection significantly affect clustering – methods to select variables based on their "clusterability"
- "Worse still, because there are several clustering methods and proximity measures, for each combination (method, proximity measure) cluster analysis output (dendogram) is quite different ..."

Data and Data Cleaning

- COVID Border Accountability Project and World Bank

Data and Data Cleaning

- COVID Border Accountability Project and World Bank
- Narrowing down variables of interest: 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 and World Bank
- Narrowing down variables of interest: Visa bans, history of travel bans, citizen bans, policy length, policy type, travel blockage (air, land, sea), refugee bans, country exceptions, work exceptions
- Dealing with NA values: visa bans, history of travel bans, citizen bans

Data and Data Cleaning

- COVID Border Accountability Project and World Bank
- Narrowing down variables of interest: Visa bans, history of travel bans, citizen bans, policy length, policy type, travel blockage (air, land, sea), refugee bans, country exceptions, work exceptions
- Dealing with NA values: visa bans, history of travel bans, citizen bans
- Other data cleaning: one-hot encoding, assumptions for strange data entries

Data and Data Cleaning

- COVID Border Accountability Project and World Bank
- Narrowing down variables of interest: Visa bans, history of travel bans, citizen bans, policy length, policy type, travel blockage (air, land, sea), refugee bans, country exceptions, work exceptions
- Dealing with NA values: visa bans, history of travel bans, citizen bans
- Other data cleaning: one-hot encoding, assumptions for strange data entries
- Aggregated by country; merged on demographic data after clusters were determined by policy

Preliminary Results: HAC

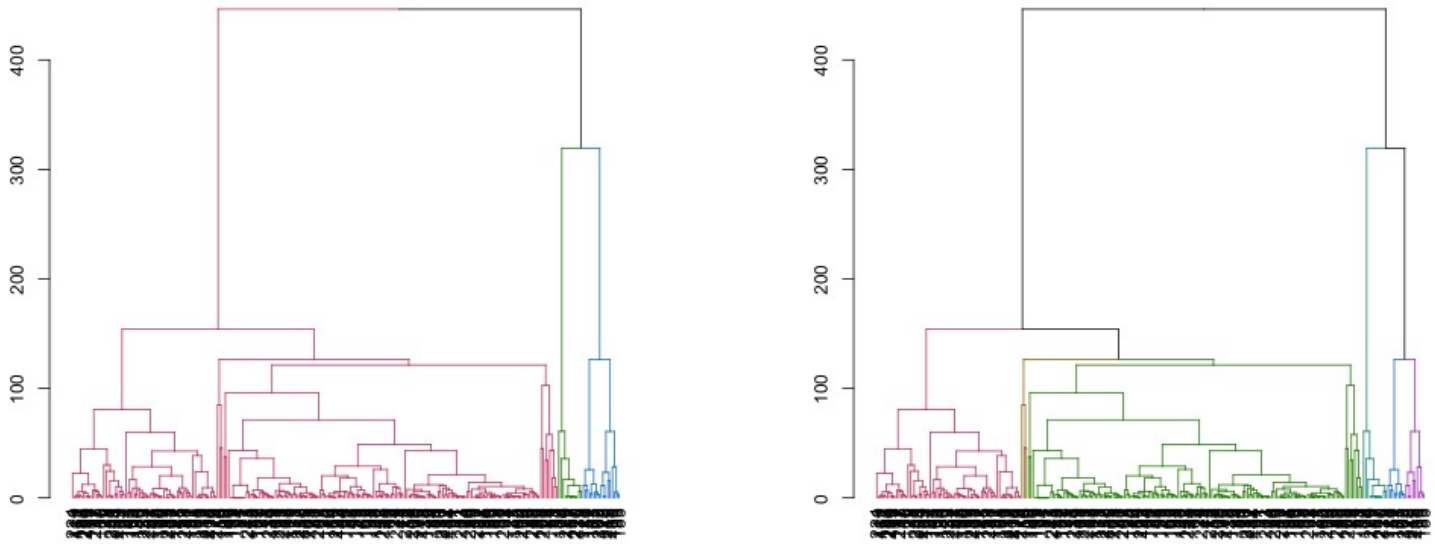
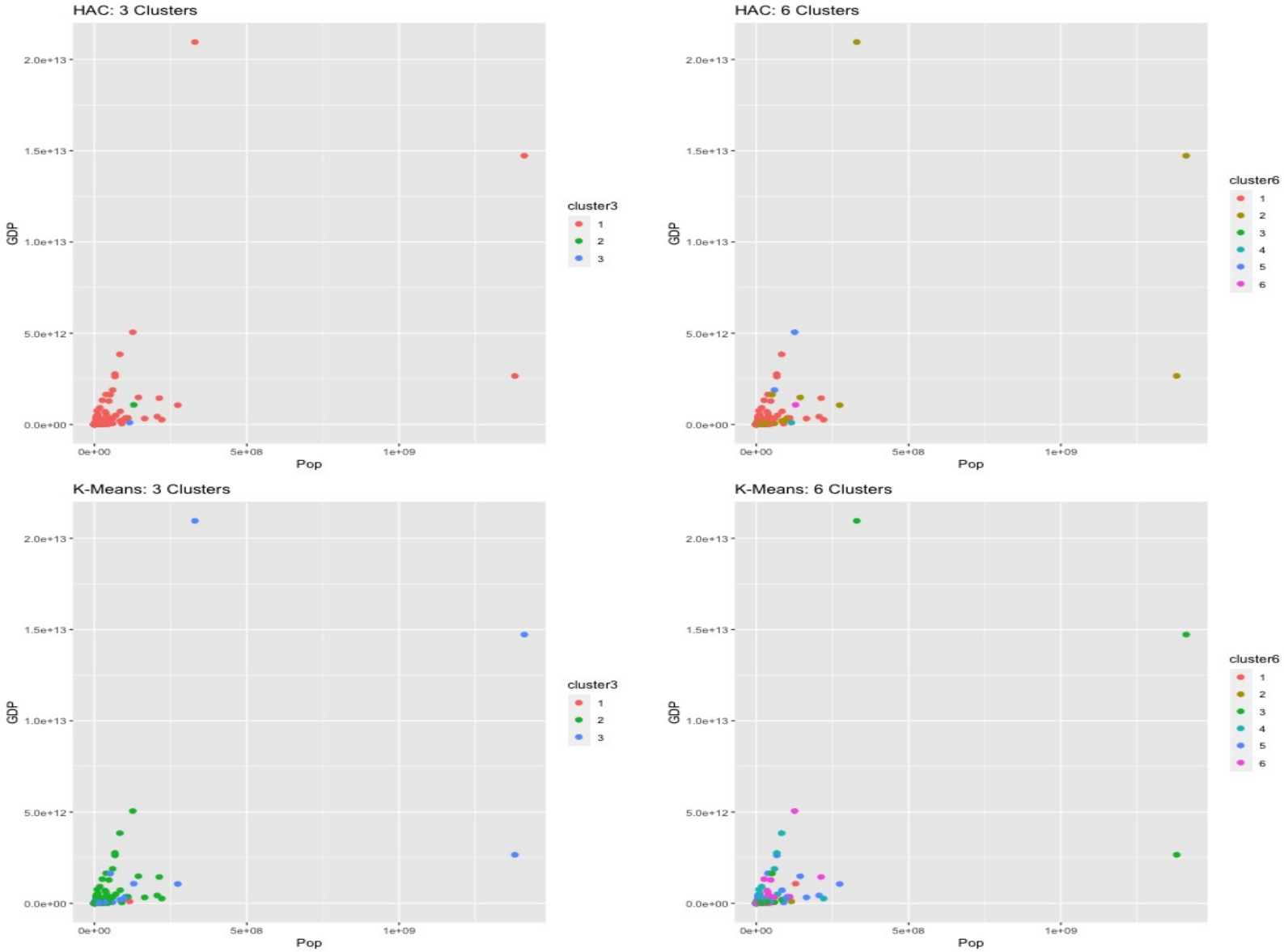


Figure: Hierarchical Clustering using 3 (left) and 6 (right) clusters and Euclidian distance between cluster centroids

Preliminary Results: Finding Clustering Patterns



GDP and Population with HAC and K-Means Clusters of 3 and 6.

Future Plans and Questions

- The effect of including/excluding certain variables on the structure of the clusters

Future Plans and Questions

- The effect of including/excluding certain variables on the structure of the clusters
- Running cluster analysis on more demographic features – are countries in the same clusters?

Future Plans and Questions

- The effect of including/excluding certain variables on the structure of the clusters
- Running cluster analysis on more demographic features – are countries in the same clusters?
- Changing the number of clusters

Future Plans and Questions

- The effect of including/excluding certain variables on the structure of the clusters
- Running cluster analysis on more demographic features – are countries in the same clusters?
- Changing the number of clusters
- Investigating cluster analysis for more "mixed" data

Future Plans and Questions

- The effect of including/excluding certain variables on the structure of the clusters
- Running cluster analysis on more demographic features – are countries in the same clusters?
- Changing the number of clusters
- Investigating cluster analysis for more "mixed" data
- Uncertainty: visualizing data

Future Plans and Questions

- The effect of including/excluding certain variables on the structure of the clusters
- Running cluster analysis on more demographic features – are countries in the same clusters?
- Changing the number of clusters
- Investigating cluster analysis for more "mixed" data
- Uncertainty: visualizing data
- Uncertainty: scope of project and time constraints

Future Plans and Questions

- The effect of including/excluding certain variables on the structure of the clusters
- Running cluster analysis on more demographic features – are countries in the same clusters?
- Changing the number of clusters
- Investigating cluster analysis for more "mixed" data
- Uncertainty: visualizing data
- Uncertainty: scope of project and time constraints
- Uncertainty: finding meaningful results in general :(

Sources

- Fionn Murtagh and Pedro Contreras. Algorithms for hierarchical clustering: an overview. *WIREs Data Mining Knowl Discov* 2012, 2: 86–97 doi: 10.1002/widm.53
- Hana Řezanková. Cluster Analysis and Categorical Data. *Statistika* 2009, 89: 216-32
- Jaime R.S. Fonseca. Clustering in the field of social sciences: that is your choice. *International Journal of Social Research Methodology* 2013, 16:5, 403-428, DOI: 10.1080/13645579.2012.716973
- Preud'homme, G., Duarte, K., Dalleau, K. et al. Head-to-head comparison of clustering methods for heterogeneous data: a simulation-driven benchmark. Sci Rep 2021, 11 4202. https://doi.org/10.1038/s41598-021-83340-8
- Steinley, D., Brusco, M.J. Selection of Variables in Cluster Analysis: An Empirical Comparison of Eight Procedures. *Psychometrika* 73, 125 (2008).