Cluster Analysis and Policy

Dr Jinjing Li

IGPA, University of Canberra

New Forms of Discovery for Policy Making Workshop, 19-20 February 2015
Cluster Analysis and Policy

• What is Cluster Analysis?
• How to Cluster Data?
• Cluster Analysis Examples @ IGPA
What is Cluster Analysis

Cluster Analysis and Policy
Introduction

• Cluster analysis is a statistical technique that can classify units into (relative) homogenous groups
Introduction
Cluster Analysis

• This is a well established techniques for many fields
  – Group similar documents
  – Find genes and proteins with similar functions
  – Spatial groups for earthquake prone regions

• Less common in the field of social and political science
What can it be used for

• Search for topologies automatically
  – How many social classes are there in Australia
  – How many types of career trajectories are there for a high school graduate?
  – What are the common viewpoints towards the Australian’s political system
  – Identify the welfare systems around the world
What can it be used for

• Search for the difference between known associations and statistically generated cluster
  – Any polling station has been tempered with?
  – Does tax offset claim matches others with the same occupation?
  – Does different generations of Australians have different views?
Why use Cluster Analysis instead of

• Typical parametric analysis such as regressions (e.g. OLS)
  – Cluster analysis can be non/semi-parametric, less stringent assumptions

• Cross-tabulations
  – You need to know the groups first
How to Cluster Data

Cluster Analysis and Policy
How to Cluster Data

• Define similarities
• Choose a clustering algorithm
• Number of clusters (*)
• Evaluate the cluster quality
Define Similarities / Distance

• Distance can be tricky
  – Think about distance between two cities (air, road, time etc.)
  – The choice of distance measure depends on the area of study.

• Commonly used measures
  – Squared distance \((a-b)^2\)
  – For high dimensional data like text: Cosine distance
  – Self defined distance measure
Find a clustering algorithm

• Most common two
  – Hierarchical Clustering
    • Need to also define the distance between clusters
    • Each cluster can be divided into smaller clusters
  – K-means
    • Each run may have a different result
Find a clustering algorithm

- Hierarchical Clustering

Dendrogram for _clus_1 cluster analysis
Find a clustering algorithm

- K-means
Evaluate the cluster quality

• External Evaluation
  – Compare with the known results (mostly happen in classification task, e.g. Fraud detection)

• Internal Evaluation
  – Looking to optimise the cluster quality by trying different number of clusters
Implementation

• Cluster Analysis is supported by almost all major statistical packages
  – SPSS
  – SAS
  – Stata
  – R
  – Many others
Cluster Analysis @ IGPA

Cluster Analysis and Policy
Ongoing work with Cluster analysis

• Political Science
  – Cluster the opinions about democracies in Australia (see the museum exhibition for results)
• Underemployment
  – Cluster the types of employment trajectory and explore the path dependency of underemployment
• Migration
  – Cluster the regions with similar patterns of migration flows and demographic profile in MDB region
• Education
  – Cluster the primary and secondary education system attributes and performances in Australia
Example: Political Views
Example: Political Views
Example: Underemployment
Concluding Remark

• Cluster Analysis is a less-explored statistical technique in the field of social and political science
• Particularly useful with large number of attributes, and when patterns are complex
• (sometimes) Imposes less model assumptions and offers an easy-to-interpret result
Thank you

Cluster Analysis and Policy