Segmentation of Statistics Canada's Proximity Measures

Weekly Meeting

Week 5

Research Questions

1. What are the optimal cut-off values and cluster boundaries determined by the chosen clustering algorithm in the PMD continuous metric?

2. What distinctive characteristics define each cluster of dissemination blocks, and how do these features contribute to both heterogeneity between clusters and homogeneity within each cluster?

   (Characteristics include: proximity measures, CSD type, DB population, IoR, and province breakdown.)

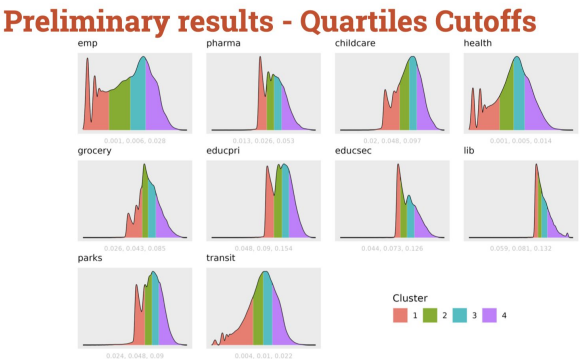
Methods

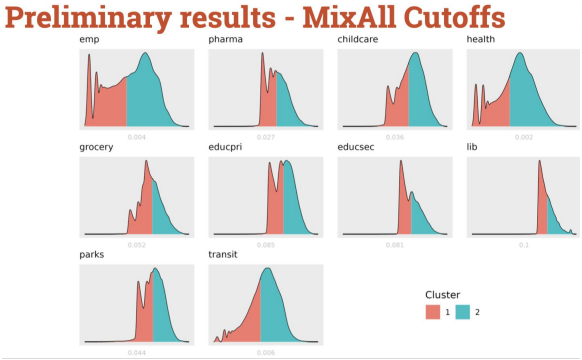
- Similar as before: apply algorithms, analyse clusters, draft report
  - Quartiles, HBDSCAN, MixAll, Varsell, K-means, fuzzy c-means, ‘manual’
  - (New) Focus on transformation of data

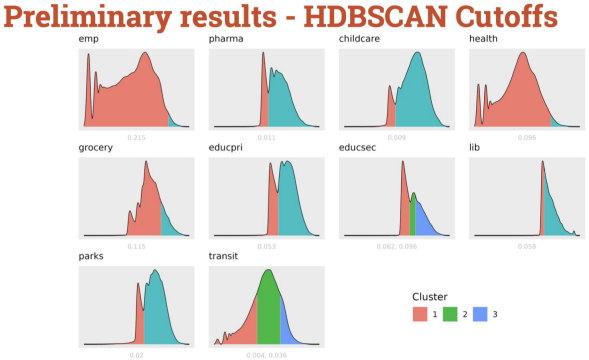
- Sending draft report for feedback early, practice presentation earlier

Progress

- Methods applied
  - Box-cox transformation
  - Quartile (base model), Quintile, HDBSCAN, MixAll, VarselLCM
- Unsuccessful Method
  - VarselLCM on log-transformed data







Client

- **Client + DEIL Lab:** Tuesday, May 16th 2023
  - Q and A session with data exploration and integration lab
- **Client Meeting:** Friday, May 26th 2023
  - Redid midterm presentation
  - Feedback on things to emphasise/improve for final presentation
    - Define end goal / motivation
    - NA clarity
    - Log transformation (Does not change OG data)

Team Effort - Weeks 3 & 4

Noman

- MixAll Methods
- Cutoff calculation
- Midterm Presentation
- 48.5 HRS

Ricky

- PMD understanding
- Managing
- Midterm pres
- 20 + 26 HRS

Jonah

- HDBSCAN Doc
- Presentation Plots
- Midterm Presentation
- 56.25 HRS

Avishek

- VarcelLCM
- Clustering Tendency & Validation
- Merge EDA
- Quartile
- Midterm Pres
- Quintile
- 63.57 HRS

As a Team:

- DEIL Team Meeting → better understanding of PMD
- Preparing midterm presentation

Upcoming Goals this week

- ‘Manual’ identification of cutoffs via minima of density plots
  - Ricky

- Algorithms to apply: k-means, fuzzy c-means, optics
  - Noman + Avishek

- Draft report
  - Jonah: Methods & Results so far

Upcoming Goals next weeks

- Cluster analysis
- Draft report
- Presentation

Roadblocks/Pivots

- Silhouette coefficient not confirming expected results → need different metric, investigate reasons

- MixAll results not great at all → investigate reasons

- Need to focus on report writing

Midterm Presentation Reflection

Went Well

- Slides were clear
- Talking time was evenly distributed
- Everyone talked at a good pace

Needs Improvement

- Some slides had too much text
- Increase cohesiveness between speakers
- Divide Q&A questions between members

Feedback / Questions