Segmentation of Statistics Canada's Proximity Measures

Weekly Meeting

Week 2

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.)

Data Description

- Primary dataset is the PMD which includes the continuous proximity scores of every dissemination block (DB) in Canada for 10 amenities such as employment, grocery stores etc.
    - The proximity measures have been normalized across Canada
    - A lower proximity measure indicates that the amenity is located farther away from the dissemination block.
- Our secondary dataset is the Index of Remoteness (IoR), a continuous numeric remoteness score for each census subdivision (CSD) in Canada.
    - Index of Remoteness equal to zero for the least remote CSD and equal to one for the most remote CSD.
    - If possible, the IoR can be linked to the proximity measures dataset by a unique ID that is available in both dataset.

Data Description

- NA values are represented in Statistics Canada's publications in a systematic way:

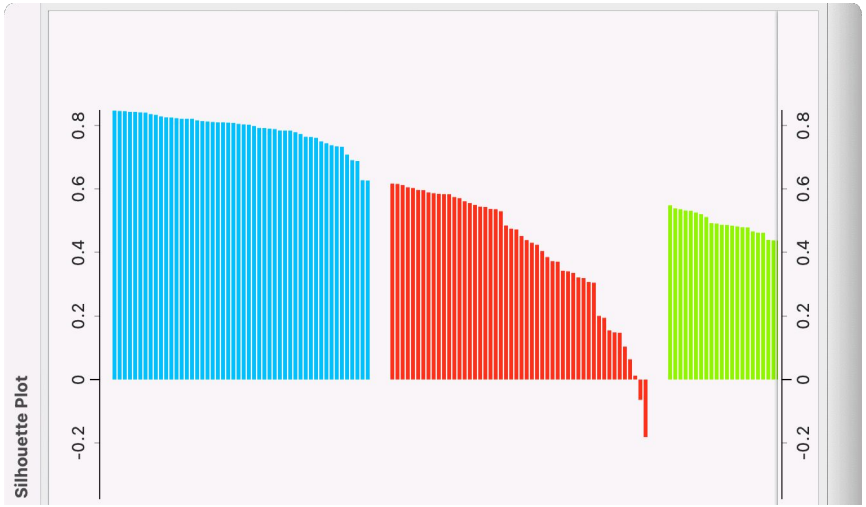
Symbol	Meaning
.	not available for any reference period
..	not available for a specific reference period
…	not applicable
F	too unreliable to be published

Methods

- Not much changed here from last meeting
- Jesse (TA) sent us several algorithms with R implementations that deal with missing values
- Comparison of algorithms to be done using Silhouette plots

Algorithm Types

- Connectivity
- Centroid
- Distribution
- Density
- Grid

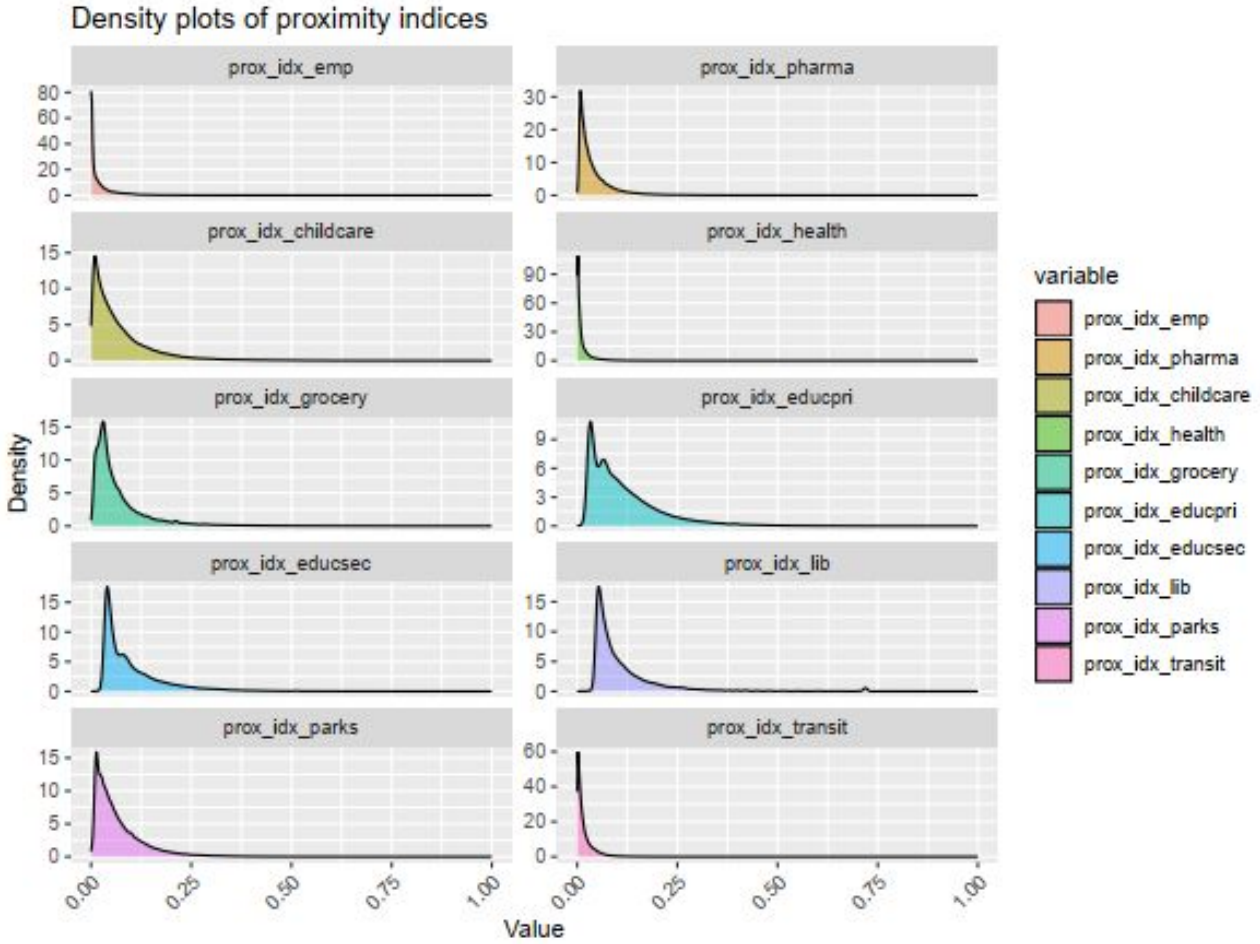


Progress

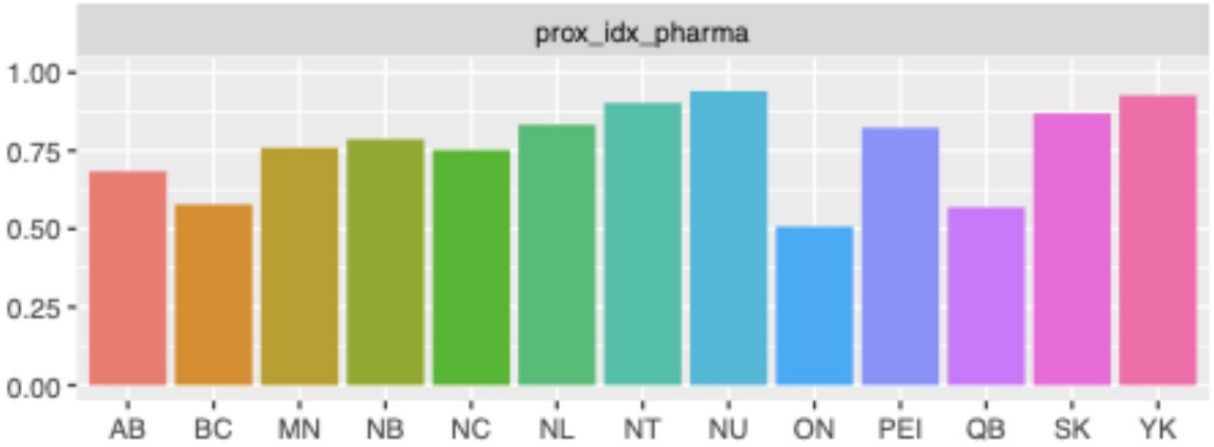
- EDA
  - Around Half Million observations but if remove NAs then left with 32,000 observations.
  - About 65,000 DBs where all the proximity measures are NA
  - All 10 Proximity measures distributions are right skewed
  - Some DBs which have no population (23%) but still have proximity measures for amenities.
  - Outliers in Proximity Measures (Chi-Square Test, Rosner's Test)
- EDA Visualization
  - Check Homogeneity
  - Missing Information from densely populated DBs
- Applied Clustering Techniques
  - K-means with Imputation
  - VarselLCM
  - MixAll

Progress

PMS Indices Distributions



Percentage of missing values by province: Pharmacies

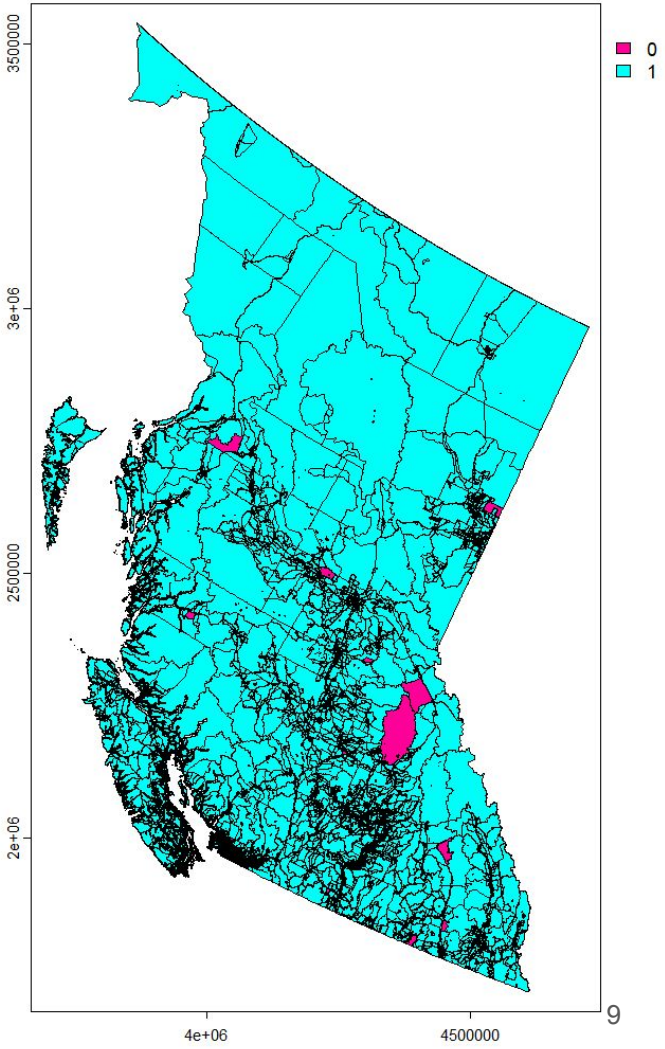


Progress

Null grocery proximity measures of BC

- 1 is NA grocery proximity measures

	NA	Not NA	Total
prox_idx_grocery	34,526	18,324	52,850
percentage	65.32%	34.68%	100%



Upcoming Goals

- Finishing up connectivity and centroid-based clustering approaches
  - Imputing values with Amelia
  - Hierarchical methods
  - Jonah, Noman

- Starting on density-based clustering approaches
  - MCLUST, MixAll, VarSelLCM algorithms
  - Ricky, Avishek

- Adding findings to Final Report document (Methods and Results sections)
  - Everyone

Roadblocks/Pivots

- Major Git LFS issues → cannot store all relevant data on GitHub
  - Resolved: everyone stores their data locally

- Computer memory issues → many algorithms cannot be run, as our computers run out of RAM and then crash
  - Work-around: 2-10% sub-sampling of original data makes algorithms viable

- Long running time → many operations can take up to 120 minutes to run
  - Work-around: 2-10% sub-sampling of original data takes less time
  - Work-around: when subsampling is not possible (ex. Plotting in space), time can be used to work on the final report

Feedback / Questions

APPENDIX


Missing values for each province by amenity
prox_idx_childcare
prox_idx_educpri
prox_idx_educsec
prox_idx_emp
prox_idx_grocery
prox_idx_health
prox_idx_lib
prox_idx_parks
prox_idx_pharma
prox_idx_transit
value
1.00
0.75
0.50
0.25
0.00
AB
BC
MN
NB
NC
NL
NT
NU
ON
PEI
QB
SK
YK
prov_name
prov_name