Segmentation of Statistics Canada's Proximity Measures

Weekly Meeting

Week 7

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

- Final Draft
  - Finished final report methods and results drafting
  - Summarize results for clustering approaches
    - Tables
    - Visualizations

- Sending draft report for feedback early, practice presentation earlier

Progress

- Draft for final report
  - Profiling Clusters algorithm
  - Summarizing results with tables for clusters
  - Plots and visuals for report
- Sending draft report to client for feedback

Progress - Tables

	# of DBs	DB Population	Median IoR	CMA Type	Province	Amenity Dense	Employment	Range
Entire Population	423,602 (100.0%)	38	0.16	CMA (48.3%)	Ontario (18.2%)	Low (90.1%)	0.006	0 - 1
Quintiles C1	78,014 (18.4%)	10	0.29	None (80.9%)	NovaScotia (6.5%)	Low (100.0%)	0.000	0 - 4e-04
Min/Max C1	29,831 (7.0%)	5	0.32	None (83.6%)	NovaScotia (8.8%)	Low (100.0%)	0.000	0 - 0.0000
HDBSCAN C1	419,062 (98.9%)	38	0.16	CMA (47.7%)	Ontario (17.9%)	Low (90.9%)	0.006	0 - 0.2298
MixAll C1	179,334 (42.3%)	16	0.27	None (73.6%)	Ontario (7.2%)	Low (99.9%)	0.000	0 - 0.0036
MCLUST C1	29,831 (7.0%)	5	0.32	None (83.6%)	NovaScotia (8.8%)	Low (100.0%)	0.000	0 - 0.0000
PAM k-means C1	177,804 (42.0%)	16	0.27	None (73.8%)	Ontario (7.1%)	Low (99.9%)	0.000	0 - 0.0035
Quintiles C2	89,705 (21.2%)	23	0.24	None (69.6%)	Ontario (9.2%)	Low (99.9%)	0.001	4e-04 - 0.0030
Min/Max C2	22,179 (5.2%)	10	0.30	None (81.2%)	NovaScotia (5.6%)	Low (100.0%)	0.000	0.0000 - 2e-04
HDBSCAN C2	4,540 (1.1%)	122	0.03	CMA (100.0%)	Ontario (44.9%)	Med (45.9%)	0.292	0.2298 - 1
MixAll C2	244,268 (57.7%)	63	0.11	CMA (73.3%)	Ontario (26.3%)	Low (82.8%)	0.023	0.0036 - 1
MCLUST C2	56,902 (13.4%)	10	0.27	None (78.6%)	Ontario (6.3%)	Low (100.0%)	0.000	0.0000 - 4e-04
PAM k-means C2	245,798 (58.0%)	62	0.11	CMA (73.0%)	Ontario (26.2%)	Low (82.9%)	0.023	0.0035 - 1
Quintiles C3	85,928 (20.3%)	41	0.20	CMA (34.2%)	Ontario (12.7%)	Low (97.8%)	0.006	0.0030 - 0.0127
Min/Max C3	14,893 (3.5%)	10	0.27	None (78.3%)	Ontario (6.7%)	Low (100.0%)	0.000	2e-04 - 3e-04
MCLUST C3	39,730 (9.4%)	20	0.24	None (72.7%)	Ontario (9.4%)	Low (99.9%)	0.001	4e-04 - 0.0012
Quintiles C4	85,096 (20.1%)	65	0.11	CMA (79.7%)	Ontario (26.9%)	Low (89.9%)	0.022	0.0127 - 0.0368
Min/Max C4	26,887 (6.3%)	17	0.23	None (75.4%)	Ontario (8.6%)	Low (100.0%)	0.000	3e-04 - 5e-04
MCLUST C4	48,188 (11.4%)	27	0.25	None (64.2%)	Ontario (9.3%)	Low (99.7%)	0.002	0.0012 - 0.0033
Quintiles C5	84,859 (20.0%)	83	0.06	CMA (99.5%)	Ontario (37.2%)	Low (62.8%)	0.072	0.0368 - 1
Min/Max C5	329,812 (77.9%)	50	0.14	CMA (59.1%)	Ontario (21.9%)	Low (87.2%)	0.014	5e-04 - 1
MCLUST C5	53,628 (12.7%)	38	0.21	None (37.2%)	Ontario (11.3%)	Low (98.5%)	0.005	0.0033 - 0.0085
MCLUST C6	64,056 (15.1%)	57	0.14	CMA (61.2%)	Ontario (19.1%)	Low (93.8%)	0.014	0.0085 - 0.0206
MCLUST C7	69,082 (16.3%)	71	0.10	CMA (91.8%)	Ontario (34.7%)	Low (85.2%)	0.032	0.0206 - 0.0518
MCLUST C8	51,824 (12.2%)	82	0.06	CMA (99.8%)	Ontario (36.4%)	Low (63.8%)	0.081	0.0518 - 0.1629
MCLUST C9	10,361 (2.4%)	118	0.03	CMA (100.0%)	Quebec (37.9%)	Med (46.0%)	0.219	0.1629 - 1

Table 11: Summary statistics for each cluster found by all approaches for the employment amenity. DB Population, IoR and proximity value show the median, while CMA Type, Province and Amenity Dense show the mode.

	Silhouette	Dunn	Calinski Harabasz	Davies Bouldin
Quintiles	0.35	0.00000	3545	0.95
MixAll	0.62	0.00492	35404	0.60
HDBSCAN	0.69	0.00338	3656	0.40
PAM k-means	0.63	0.00498	36372	0.59
MCLUST	0.59	0.00126	98539	0.56
Min/Max	0.60	0.00014	256	1.01

Table 20: The validation metric values for each clustering approach for the employment amenity.

Progress - Visuals

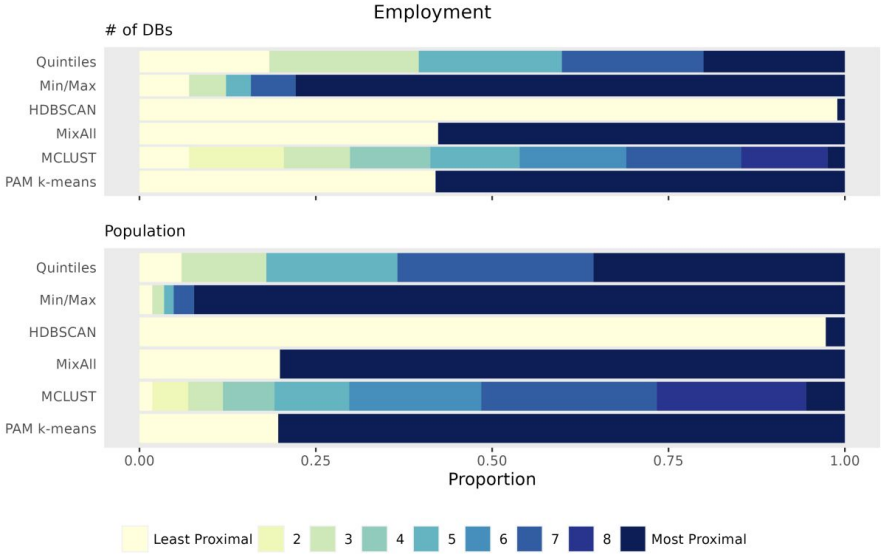
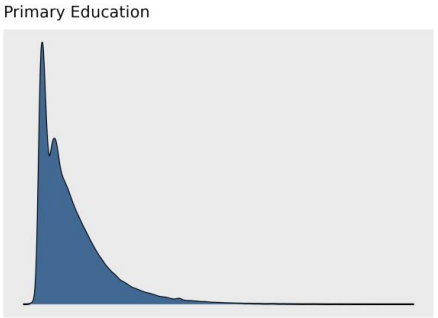
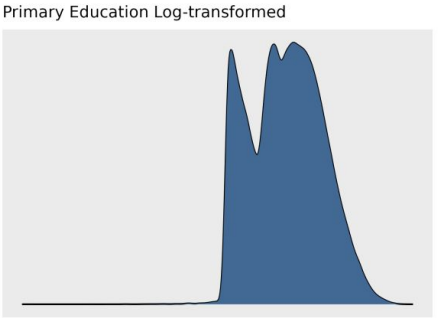


Figure 32: Proportion of DBs and population in each cluster for all approaches for the employment amenity.





Client

- **Client Meeting:** Friday, June 9th 2023
  - Update current progress on report (send draft report by sunday night)
  - Q & A on expectations for report
  - Planning / goals for next week

Team Effort - Week 6

Noman

- Profiling
- Final report draft
- 32.5 HRS

Ricky

- Profiling
- Final report draft
- 33.5 HRS

Jonah

- Profiling
- Final Report draft
- Tables & Visualizations
- 37.5 HRS

Avishek

- Profiling
- Final report draft
- QGIS
- 35 HRS

As a Team:

- Draft of final report sent to client for review
- Meeting with Jeff

Upcoming Goals this week

- Working on Final Report
  - Wrap up discussion
  - Brainstorm and implement more effective visualizations
  - Ensure up to standard

- Get feedback from client and TA
  - Update report based off feedback

- Final presentation prep
  - Reviewing all feedback from midterm presentation
  - Making slides
  - Practicing + get feedback from TA/Instructor

Upcoming Goals next weeks

- Submit final report

- Final Presentation

- Reflect on the last 10 months of MDS and mentally prepare for it all to come to an end next week

Roadblocks/Pivots

- No results is still a result?
- Discussion / result investigating is hard (so many methods/amenities
- How to effectively show a lot of information?
- Coordination in writing report: hard to effectively ‘split off’ work / work together

Feedback / Questions