A Proposal for the Segmentation of Statistics Canada's Proximity Measures

Jonah Edmundson, Ricky Heinrich,
Noman Mohammad, Avishek Saha

Introduction

We all live somewhere, and inhabit physical space. Unless one lives completely removed from others, amenities are usually present in the built environment: schools, places of employment, healthcare facilities, etc. These amenities serve to make residents' lives better, and are the result of policy and planning by multiple groups. Like people, they inhabit physical spaces, and not everybody is equidistant from them. As Alasia et al. (2021) outline: "having physical access to basic services and amenities is a key determinant of social inclusion, their capacity to meet basic needs, and their ability to fully participate in social and economic development."

The Proximity Measure Database (PMD) developed by the Data Exploration and Integration Lab (DEIL) at Statistics Canada serves to provide a granular measure of proximity to services and amenities to inform planning and policy questions (Alasia et al., 2021). The PMD contains continuous measures for 10 amenities at a 'dissemination block' (DB) level.

Our goal is to segment these continuous proximity measures to group similar dissemination blocks together based on their access to amenities. These clusters may provide valuable insights to policymakers and urban planners regarding how to prioritize efforts to improve accessibility and promote social and economic sustainability.

There are various clustering algorithms in the literature, like the k-means algorithm (MacQueen, 1967) and the fuzzy c-means algorithm (Bezdek et al., 1984). There are also various metrics defined for assessing clustering quality. These metrics can be used to evaluate the performance of different clustering algorithms and to determine which algorithm is the best fit for a particular dataset. Examples of such metrics include Rand Index (RI), Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), Adjusted Mutual Information (AMI), F-measure, Homogeneity, V-measure, Heterogeneity, Completeness, and Silhouette coefficient (Mehta et al., 2020).

Our motivation for this project is rooted in the need for improved urban planning and policy-making that can benefit individuals, businesses, and the overall community. Through our analysis, we hope to contribute to a better understanding of local access to amenities.

Data Sources

Our primary dataset of interest is the Proximity Measures Dataset (PMD) from the DEIL at Statistics Canada, which includes the continuous numerical proximity scores of every dissemination block (DB) in Canada for 10 services/amenities: employment, grocery stores, pharmacies, health care, child care, primary education, secondary education, public transit, neighborhood parks, and libraries. These proximity measures were calculated using a gravity model that takes into account the distance between a reference DB and all other DBs where the service is located within a specified range, as well as the size of those services. Additionally, the presence of services within the reference DB is factored into the measure. These measures are considered a reliable way to assess local access to various amenities. The data dictionary for this dataset can be found here.

Our secondary dataset is the Index of Remoteness, also from Statistics Canada. This dataset includes a continuous numeric remoteness score for each census subdivision in Canada. If possible, it can be linked to the proximity measures dataset by determining which DBs reside in each census subdivision.

Research Questions

1. Is the missing value structure of the PMD dataset randomly distributed in Canada? Or is it biased to some targeted geographic regions?
2. Which clustering approach is best at identifying meaningful cutoff values/segments in the proximity measures?
3. What are the cut-off values suggested by the clustering algorithm the PMD continuous metric? (This requires the analysis, not only of the content of each cluster, but also of the boundaries distinguishing the clusters.)
4. What are the characteristics of each cluster of DBs? (Heterogeneity between clusters and homogeneity within a cluster.)
5. Can a generalized clustering approach be applied to all amenity types, or should specific clustering methods be used for different amenities?
6. If it is possible to link relevant open source data: are there correlations between the identified clusters and socio-economic factors, such as population density, building density, or the proportion of rural/urban areas?
7. Can additional datasets provide further insights into amenity accessibility or more clear clusters?
8. Is our chosen clustering approach robust?

Methodology

Our methodology consists of three sequential parts: exploratory data analysis (EDA), statistical analysis, and visualizations.

The EDA includes:

- Data can be downloaded, already cleaned and prepped from the Statistics Canada Website.
- Investigating missing values and ways to deal with them.
- Base Model – Intuition/Violin Plots (individual measures only).

We will test various techniques on different subsets of the datasets, such as individual proximity measures, all proximity measures combined, and if possible, proximity measures in conjunction with population density, Index of Remoteness, and neighborhood income. Possible clustering algorithms we may explore include:

- Connectivity based (Hierarchical)
  - Complete linkage (Base R)
  - Average linkage (Base R)
  - Single linkage (Base R)

    - BIRCH (`stream` package)

- Centroid based
    - k-means (Base R)
    - fuzzy c-means (`ppclust` package)
    - Mean-shift (`meanShiftR` package)
    - Affinity propagation (`apcluster` package)
- Distribution based (mixture models)
    - Gaussian Mixture Modelling (`mclust` package)
    - Model-Based Clustering with the Multivariate t-Distribution (`teigen` package)
- Density based
    - Density-Based Spatial Clustering of Applications with Noise (`dbscan` package)
    - HDBSCAN (`dbscan` package)
    - OPTICS = Ordering points to identify the clustering structure (`dbscan` package)
- Grid based
    - CLIQUE (`subspace` package)

All clustering approaches will be compared and validated to assess clustering quality. We will explore different metrics defined in the literature, such as the Dunn Index and Silhouette Coefficient.

We will explore ways to visualize the results of this work, such as Silhouette plots for clustering validation and interactive maps similar to Statistics Canada's Proximity Measures Data Viewer for the final results.

Deliverables

- Final report showing documentation for all steps of our work: an exploration of the data, the different approaches attempted, their validity, a sensitivity analysis, a chosen reproducible clustering methodology, the characteristics of the clusters, the identification of the PMD cut-off values, and the interpretation of the final results.
- Final presentation slides.

Schedule

Key dates are in italics.

- Week 1 .... (May 1 - 5)
  - Proposal
  - Initial setup, getting oriented

  May 7 - Written Proposal

- Week 2 .... (May 8 - 12)
  - EDA - method to deal with missing values, exploring additional datasets, characteristics of data.
  - Trying connectivity and centroid-based clustering approaches, recording progress
    ⟶ research and apply method, interpret clustering results, report section draft.

- Week 3 .... (May 15 - 19)
  - Start writing methods and results (using what we have so far).
  - Trying distribution-based clustering approaches, recording progress
    ⟶ research and apply method, interpret clustering results, report section draft.

- Week 4 .... (May 22 - 26)
  - Preparing for midway presentation.
  - Trying density and grid-based clustering approaches, recording progress
    ⟶ research and apply method, interpret clustering results, report section draft.

  May 25 - Midterm Presentation

- Week 5 .... (May 29 - June 2)
  - Finishing up modelling approaches.
  - Piecing report together, start final draft (methods and results section should be mostly done).

- Week 6 .... (June 5 - 9)
  - Consolidate modelling results (cluster profiles, robustness check)
  - Finish draft report, submit to Jerome for major edits.

- Week 7 .... (June 12 - 16)
  - Finalizing report (minor edits).
  - Flex week (catch up on stuff or start working ahead).

- Week 8 .... (June 19 - 22)
  - Preparing for final presentation.

  June 20 - Final Report

  June 22 - Final Presentation

In order to ensure that every team member gains a well-rounded experience and contributes effectively to the capstone project, we will divide our team into subgroups of two, each focusing on different aspects of the project (on modelling weeks, algorithms will be divided equally between the two subgroups). We will rotate the members among these subgroups, allowing everyone the opportunity to work on various components and gain exposure to different challenges and skill sets. To optimize efficiency, we will hold daily team meetings where we will assess progress and distribute/prioritize tasks as needed. This collaborative approach will foster a deeper understanding of the project as a whole, while promoting teamwork.

Limitations

1. Since the proximity measures dataset was only recently released as "experimental statistics", it is possible that better, more comprehensive ways of calculating the proximity index using more/different data sources may be developed in the future, which may render our methodology obsolete.

Conclusion

Our project aims to apply clustering algorithms to segment proximity measures for various amenities as provided by Statistics Canada. The insights gained from this segmentation can help policymakers and urban planners make informed decisions on how to prioritize efforts to improve access and promote social and economic sustainability. By selecting a robust clustering methodology and exploring the relationships between clusters and socio-economic factors, we hope to contribute to a better understanding of local access to amenities and its implications on communities.

Bibliography

1 Alasia, A., Newstead, N., Kuchar, J., & Radulescu, M. (2021, February 15). *Measuring Proximity to Services and Amenities: An Experimental Set of Indicators for Neighbourhoods and Localities.* Reports on Special Business Projects, Statistics Canada. Retrieved May 4, 2023, from https://www150.statcan.gc.ca/n1/pub/18-001-x/18-001-x2020001-eng.htm

2 MacQueen, J. B. (1967). *Some Methods for classification and Analysis of Multivariate Observations.* Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. Vol. 1. University of California Press. pp. 281–297.

3 Bezdek, J.C., Ehrlich, R., & Full, W. *FCM: The fuzzy c-means clustering algorithm*, Computers & Geosciences,Volume 10, Issues 2–3,1984, Pages 191-203,ISSN 0098-3004, https://doi.org/10.1016/0098-3004(84)90020-7.

4 Mehta, V., Bawa, S. & Singh, J. *Analytical review of clustering techniques and proximity measures.* Artif Intell Rev 53, 5995–6023 (2020). https://doi.org/10.1007/s10462-020-09840-7

5 Alasia, A., Bédard, F., Bélanger, J., Guimond, E., & Penney, C. (2017). *Measuring remoteness and accessibility: A set of indices for Canadian communities.* Reports on Special Business Projects, Statistics Canada. https://www150.statcan.gc.ca/n1/pub/18-001-x/18-001-x2017002-eng.htm.