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

Jonah Edmundson, Ricky Heinrich,
Noman Mohammad, Avishek Saha

Introduction

text

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. It will need to be linked to the proximity measures dataset by determining which DBs reside in each census subdivision.

Research Questions

1. Which clustering approach is best at identifying meaningful cutoff values/segments in the proximity measures?
2. Is our chosen clustering approach robust?
3. What are the characteristics of each cluster of DBs?
4. Can a generalized clustering approach be applied to all amenity types, or should specific clustering methods be used for different amenities?
5. Are there correlations between the identified clusters and socio-economic factors, such as population density, building density, or the proportion of rural/urban areas?
6. Can additional datasets provide further insights into amenity accessibility or more clear clusters?

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 perform various techniques on different subsets of the datasets, such as individual proximity measures, all proximity measures combined, and 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

- A document describing our chosen reproducible clustering methodology.
- Final report showing the different approaches attempted, their validity, and a sensitivity analysis.
- Mapbox interactive choropleth map visualization.
- 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.
- Week 3 ..... (May 15 - 19)
  - Start writing methods and results (using what we have so far).
  - Trying distribution-based clustering approaches, recording progress.
- Week 4 ..... (May 22 - 26)
  - Preparing for midway presentation.
  - Trying density and grid-based clustering approaches, recording progress.

  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)
  - Start working on interactive choropleth visualization.
  - 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 smaller groups (groups of 2), each focusing on different aspects of the project. Over the course of the project, we will rotate the members among these groups, 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 across the entire team.

Limitations

1. The proximity measures dataset has lots of missing data. Once data is available for many smaller/uninhabited/remote DBs, it may skew our clustering results and our algorithm may need to be replaced.

2. 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 *Workshop on Modernising Statistical Systems for better data ...* - oecd.org. (n.d.). Retrieved May 4, 2023, from https://www.oecd.org/cfe/regionaldevelopment/modernising-statistical-systems.htm

2 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.* Retrieved May 4, 2023, from https://www150.statcan.gc.ca/n1/pub/18-001-x/18-001-x2020001-eng.htm

3

4

5