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, such as schools, places of employment,and healthcare facilities. These amenities serve to make residents' lives better, and their distribution is often the result of careful policy and planning by governing bodies. Like people, they inhabit physical space, 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." Therefore, it is imperative for these governing bodies to make deliberate, well-informed decisions as to the location of new amenities and services.

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, the most granular area defined by StatCan (Statistics Canada, 2021). In an urban area, a DB corresponds to a city block, whereas in rural areas they are areas "bounded by roads or other natural features" (Alasia et al., 2021). Thus, DBs different broadly in their proximity to these amenities. In order to facilitate intuitive understanding in how these DBs differ in this regard, they can be clustered based on the proximity values calculated.

There are various clustering algorithms in the literature, such as the k-means algorithm (MacQueen, 1967) and the fuzzy c-means algorithm (Bezdek et al., 1984). There are also various validation 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), Normalized Mutual Information (NMI), and the F-measure (Mehta et al., 2020).

The goal of this project is to segment the continuous proximity measures in the StatCan PMD in order to identify stable cutoffs that distinguish dissemination blocks based on their proximities to amenities. Supplementary factors beyond the proximity indices may impact the resulting clusters. Intuitively, we would expect that additional information about the DBs such as remoteness, rural character, population average income or zoning information help increase the separation between clusters. Rigolon and Németh investigated causes of uneven access to urban amenities in their specific case study of Denver parks. They found that factors such as funding systems, public policies like zoning, and social mechanisms could significantly alter access (Rigolon & Németh, 2021). While we would expect the aforementioned variables such as zoning to factor into the PMD values, we need to investigate whether there is data available to add to the PMD in order to facilitate more distinct clustering results. The results of this clustering analysis will provide valuable insights to policymakers and urban planners regarding how to prioritize efforts to improve accessibility and promote social and economic sustainability. We hope that this research contributes to a better understanding of local access to amenities for Canadian citizens.

Data Sources

Our primary dataset of interest is the PMD from the DEIL at Statistics Canada, which includes the continuous numerical proximity scores of every dissemination block (DB) in Canada for 10 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. In the PMD, the proximity measures have been normalized across Canada. In this case, a lower proximity measure indicates that the amenity is located farther away from the dissemination block. So, if the proximity measure is low, it means that the amenity is more distant from the dissemination block than if the proximity measure were high. The data dictionary for this dataset can be found here.

Our secondary dataset is the Index of Remoteness (IoR), also from Statistics Canada. This dataset includes 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 or the most remote CSD. A census subdivision is a general term used by Statistics Canada to refer to a municipal or equivalent area. It can be a city, town, village, township, regional municipality, or other type of local government unit. DBs are smaller geographic units used by Statistics Canada which fall within CSDs. If possible, the IoR can be linked to the proximity measures dataset by a unique ID that is available in both dataset.

"NA" is a symbol that represents missing values, which means that the data for a particular observation or variable is not available or is incomplete. In the context of Statistics Canada, the term "NA" is used to indicate missing data in a standardized and consistent way. The following standard symbols are used in Statistics Canada publications: [1]

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

NA values are most likely due to missing data in the Business Register, or other authoritative open data sources, such as the Linkable Open Data Environment, the Open Database of Educational Facilities and the General Transit Feed Specification.

Research Questions

Throughout the project, we will be trying various clustering algorithms. The clusters returned by these algorithms will be of varying quality; some will be well-defined, and others may be more muddled. In order to choose one best algorithm, the clusters returned from each algorithm will have to be compared using a validation metric such as the Dunn Index or Silhouette Coefficient. These metrics are generalizable between algorithms because they compare intra- and inter-group variance. Once a "best" clustering algorithm has been chosen, the following two questions can be asked:

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.)

Methodology

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

The EDA includes:

- Investigating additional data sources and how to link them.
- Investigating missing values and ways to deal with them, including imputation and removal.
- Investigating characteristics and distributions of variables in the PMD

We will test various techniques on the individual proximity measures, all proximity measures combined, and if possible, proximity measures in conjunction with population density, IoR, and CSD type. 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)
  - k-means with built-in missing data imputation (`ClustImpute` package)
- Distribution based (mixture models)
  - Gaussian Mixture Modelling (`mclust` package)
  - Model-Based Clustering with the Multivariate t-Distribution (`teigen` package)
  - Model-Based Clustering with variable selection and estimation of the number of clusters (`VarSelLCM` package)
  - Clustering Mixed data with Missing Values (`MixAll` 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)

Note: Among these algorithms, some are capable of handling missing values (NAs) directly without imputation, some have built-in imputation methods, and others require that the missing values be addressed (either imputed or removed) prior to running the algorithm.

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. We will select and apply different methods based on how we handle missing values (NAs) in our data, such as imputing or removing them.

We will use different approaches to determine the cutoff values or thresholds proximity indices. So, by comparing the cutoff values suggested by each approach, we can assess the sensitivity of the results to the choice of method and determine whether the findings are sensitive (robust) to changes in the method used.

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
    $\longrightarrow$ 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
    $\longrightarrow$ 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
    $\longrightarrow$ 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

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 Rigolon, A., & Németh, J. (2021). *What shapes uneven access to urban amenities? Thick injustice and the legacy of racial discrimination in Denver's parks.* Journal of Planning Education and Research, 41(3), 312-325.

3 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.

4 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.

5 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

6 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.