Statistics Canada Statistique Canada

THE UNIVERSITY OF BRITISH COLUMBIA

An Executive summary for the

Segmentation of the Statistics Canada's Set of Proximity Measures – A Clustering Algorithm Approach

Authors:
Ricky Heinrich, Noman Mohammad, Avishek Saha, Jonah Edmundson - The University of British Columbia

Report delivered to:
Jerome Blanchet, Ms.S and Bjenk Ellefsen, Ph.D –
Statistics Canada, Center for Special Business Projects, Data Exploration and Integration Lab, and
Firas Moosvi, Ph.D – University of British-Columbia

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 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). Our project aims to apply clustering algorithms to segment proximity measures for various amenities as provided by Statistics Canada. This clustering will allow for the continuous PMD metrics to be summarized as categorical variables, improving their usefulness in interpretation and application. The insights gained from this segmentation may help policymakers and urban planners to make better decisions and plans for community development.

The analysis began with exploratory data analysis, examining missing values, the distribution of proximity measures, outliers, and the impact of log-transformation on proximity measures. Univariate clustering was then conducted, applying clustering techniques to individual amenity log-transformed proximity measures. Before clustering each amenity, a clustering tendency check was performed to evaluate whether the data was suitable for clustering, as clustering techniques can produce clusters even when data is not inherently clusterable. Various clustering techniques were applied, including density-based (HDBSCAN, OPTICS), distribution-based (MixAll, MCLUST, Jenks Natural Breaks), and centroid-based (PAM) methods. Several cluster validation metrics were utilized to determine the appropriate number of clusters for each algorithm and assess the quality of clustering results. Finally, cluster profiling investigated additional variables such as the Index of Remoteness (IoR), number of DBs, and DB population to gain insights about the clusters.

The results of the current investigation were mixed. Even after log-transformation, assessment of clustering tendency demonstrated that the PMD is not particularly clusterable. This lack of natural divisions in the data led to inconsistent cluster cutoffs that were sensitive to the algorithm used. Not only did different clustering algorithms find an inconsistent number of clusters for the same amenity, but the location of the cutoffs between clusters also varied. However, there were instances where some cutoffs were relatively close to one another. Cluster profiling revealed that, for most amenities, different clusters have distinct characteristics. In most cases, as the proximity measure increases, the median DB population also tends to increase while median IoR decreases. This pattern suggests that areas with higher population tend to be less remote and have higher proximity to amenities.

The most significant takeaway from the current investigation is the lack of clear-cut segments in the PMD. While it is true that log-transforming the proximity measures did reveal certain density-sparse regions, the clustering algorithms utilized did not consistently identify these regions. As a result, we observed a lack of stability in the clustering results. This is also reflected by the lack of consensus suggested by the cluster validation metrics. Certainly, this does not invalidate the ability of the PMD to accurately judge proximity to amenities; rather, it suggests that proximity to amenities across Canada is a relatively smooth gradient without any obvious natural clusters.

June 19, 2023