Statistics Canada Statistique Canada

THE UNIVERSITY OF BRITISH COLUMBIA

A Report
For the

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

June 2023

Authors:
Ricky Heinrich, The University of British Columbia
Noman Mohammad, The University of British Columbia
Avishek Saha, The University of British Columbia
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 – The University of British Columbia

In compliance with the Capstone Project requirements of the MDS program at the University of British Columbia - Okanagan.

Contents

List of Figures

List of Tables

Acknowledgements

We want to express our gratitude to Jerome Blanchet, who served as our industry advisor from Statistics Canada. We are also thankful to Professor Dr. Firas Moosvi and Irene Vrbik, who guided us as our capstone project advisors. Additionally, we appreciate the support and feedback provided by our Teaching Assistant, Jesse Ghashti. Their consistent guidance helped us to successfully complete our Master's of Data Science Capstone project within a tight two-month timeframe.

1 Executive Summary

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 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), 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.

2 Introduction

Each individual lives somewhere and inhabits physical space. Unless one lives completely removed from others, amenities such as schools, places of employment, and healthcare facilities are usually present in the built environment. 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." These amenities play a vital role in improving residents' quality of life, with their distribution being a product of meticulous policy and planning by governing bodies. In the context of urban development, accurately predicting population movement relies on the accessibility of land for the population. By utilizing accessibility measures at the 'dissemination block' (DB) level, it becomes possible to enhance the accuracy of population movement predictions and make urban planning more precise. Like people, amenities inhabit physical space, and not everybody is equidistant from them. 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 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 regions they are areas "bounded by roads or other natural features" (Alasia et al., 2021). Thus, DBs differ broadly in their size as well as in their proximity to these amenities.

The aim of this project is to group the continuous proximity measures in the Statistics Canada's PMD into distinct categories using different clustering techniques. By doing so, we can create a more straightforward and easier-to-understand measure. Categorizing the data means putting similar values together in one group and dissimilar values in other groups. This segmentation research helps to preprocess and clean the dataset for use in future research, as a highly detailed continuous variable may sometimes offer too much information. Transforming it into a categorical variable makes it easier to analyze with descriptive statistics or regression models. Using categorical variables in regression allows for better interpretation of coefficients and other statistical results. This research aims to simplify the work of researchers who may not have the time to preprocess their datasets. Similar efforts have been made by Statistics Canada in the past to transform continuous metrics into categorical ones (Subedi et al., 2020). Improving the understanding and use cases of the PMD will enable policymakers and urban planners to prioritize efforts effectively to enhance accessibility and promote social and economic sustainability within communities. In this report, we outline the methodologies used to explore segmentation of the continuous proximity measures, the robustness of the group boundaries, and the characteristics of the groups. The two specific research questions we aimed to addressed in the project are:

1. What are the optimal cut-off values and cluster boundaries for each amenity proximity measure in the PMD determined by appropriate clustering algorithms?

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, Census Metropolitan Area type, DB population, Index of Remoteness (IoR), and provincial breakdown.)

3 Background

The methodology used to generate the PMD is presented in Measuring proximity to services and amenities, by Alasia et al. with Statistics Canada (2021). Accompanying this report is the Proximity Measures Data Viewer, an online mapping application that allows users to view proximity measures by DB for a selected amenity (Statistics Canada, 2020a). The continuous measure is segmented by quintiles and assigned a colour, as shown in Figure 1, giving a user a rough idea of proximity differences between DBs. For this reason, we used the 'quintile method' as our base model in this project.

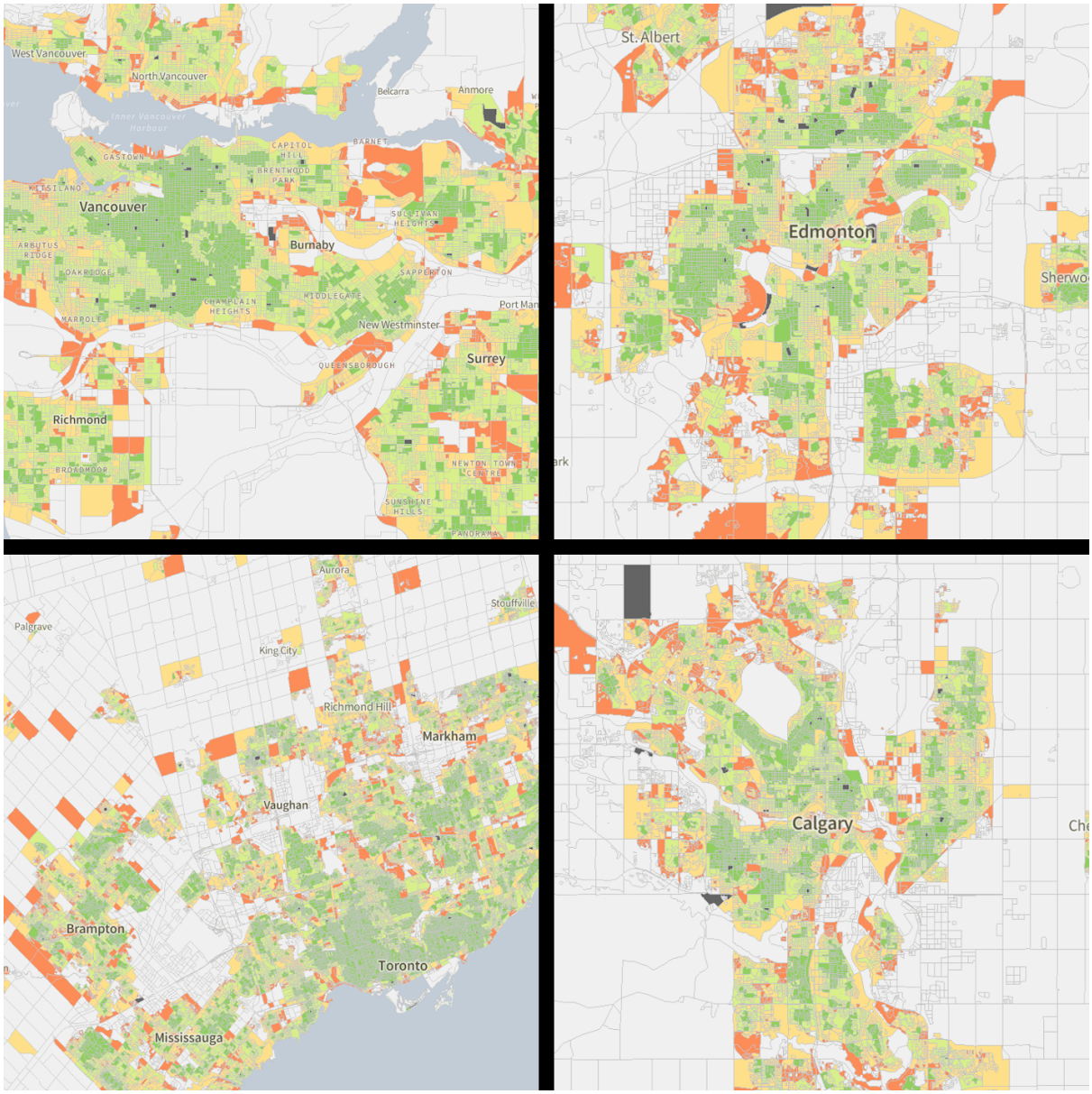

Figure 1: Statistics Canada's Proximity Measures Data Viewer showing the proximity to the primary education amenity in Vancouver (top left), Edmonton (top right), Toronto (bottom left) and Calgary (bottom right).

A 'sister' measure to the PMD is the Index of Remoteness, presented in *Measuring remoteness and accessibility* by Alasia et al. (2017). They outline how a key factor affecting socioeconomic and health outcomes is geographic proximity to population and service centers. As a result, remoteness is relevant when analyzing and implementing policies and programmes.

The distance to all the population centers as well as their population size are taken into account when calculating the index for each census subdivision (CSD). Figure 2 from the original publication summarizes their results. This index was generated as a continuous measure, which was then partitioned into categories as outlined in *Developing Meaningful Categories for Distinguishing Levels of Remoteness in Canada* by Subedi et al. (2020). The authors present five approaches to categorize the continuous measure, which included methods like Jenks natural breaks, k-means, and quintile classification. They aimed to examine various ways to group the continuous remoteness index values of CSDs into meaningful categories (Subedi et. al 2020). This is similar to our goal of categorizing the continuous proximity measures of amenities in the PMD.

The Jenks natural breaks classification and k-means classification techniques are examples of distribution and centroid based techniques, respectively. Another type of clustering technique that may be appropriate for our project are density based methods (De Smith et al., 2021). As Kassambara, a Bioinformatics R&D Scientist at Veracytes, outlines in his self-published book, the selection of an appropriate clustering technique depends on the nature of the data under investigation. He explains how before applying clustering techniques, it is important to assess the clustering tendency to ensure meaningful results, as clustering algorithms may identify clusters even in cases where no clear clusters exist in the data. Additionally, determining the optimal number of clusters is a necessary step in the process. Kassambara outlines how once clustering techniques have been applied, it is important to evaluate their performance using cluster validation statistics such as the Silhouette Coefficient (Rousseeuw, 1987), Dunn index (Dunn, 1974), and Davies Bouldin index (Davies & Bouldin, 1979). These metrics assist in identifying the most suitable clustering technique for the given data. By considering these factors, researchers can make informed decisions and select the appropriate clustering methodology for specific analysis (Kassambara, 2017b).

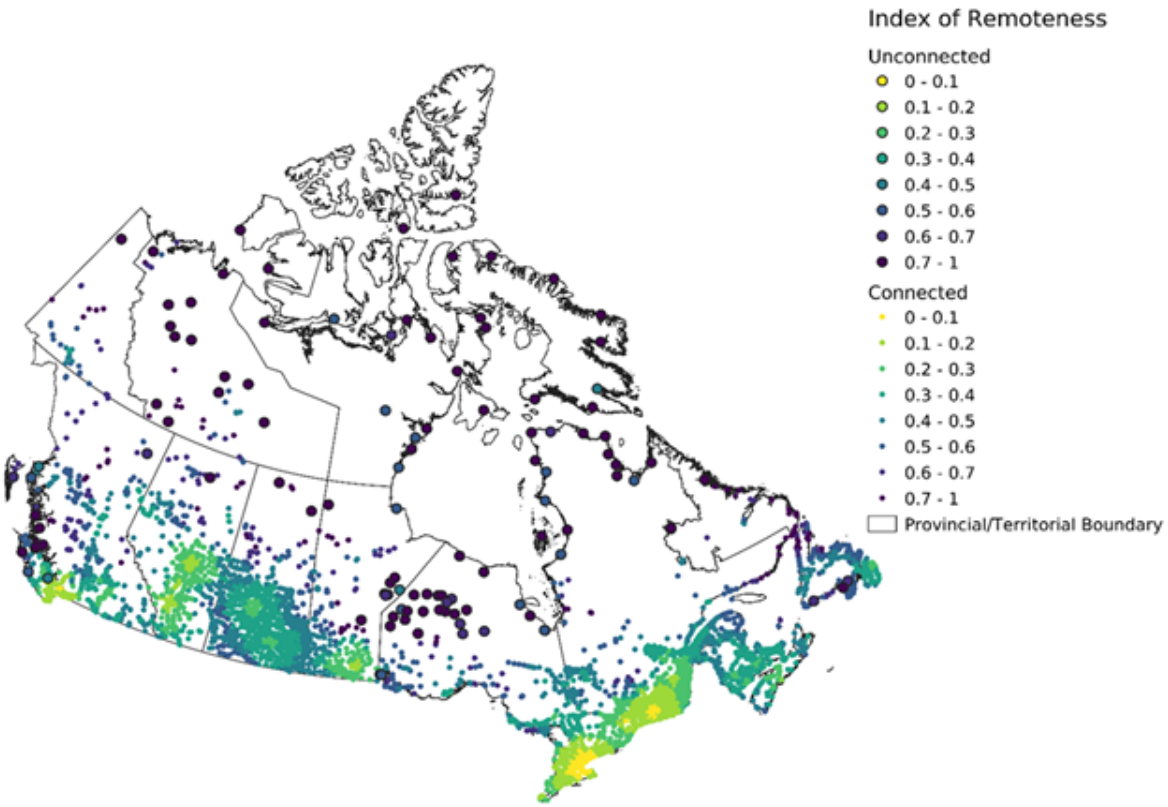

Figure 2: The spatial distribution of the 2021 Index of Remoteness over census subdivisions in Canada.

4 Data

4.1 Primary Dataset

The primary dataset for this study is the early release of the PMD, available online and provided by the DEIL at Statistics Canada (2020b). The updated PMD will be published online on the Statistics Canada webpage on June 27th, 2023. The PMD contains continuous numerical proximity measures for every DB in Canada within a select radius for 10 amenities: employment, grocery stores, pharmacies, health care, child care, primary education, secondary education, public transit, neighborhood parks, and libraries.

The proximity measures are based on a gravity model that accounts for the distance between a reference DB and all the DBs within a given travel distance in which the service is available. The proximity measures also take into account the 'mass' of services within the given distance, representing the number of services and their size. The proximity measures are published as normalized index values, meaning that the values resulting from computations were converted to a scale from 0 to 1, where 0 indicates the lowest proximity value across Canada, and 1 indicates the highest proximity value. The proximity level can be seen as the quantity of service relative to the distance traveled (Alasia et al., 2021). These measures are considered a reliable way to assess local access to various amenities (OECD, 2018). The data dictionary for this dataset can be found in Figure 27 of the appendix A.3.

4.2 Data Limitations

Statistics Canada uses a specific convention for representing different types of missing values. Table 1 shows the standard symbols that are used by Statistics Canada. The symbols present in the PMD are '..' and 'F'.

Table 1: Missing value symbol convention from Statistics Canada.

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

Values "too unreliable to be published" (F) are due to unavailability of data in the many data sources used to construct the PMD. Data that is 'not available' (..) for a DB is the result of that DB being out of scope: while producing the PMD, the authors considered a maximum travel radius for each amenity, as a mean to reduce computational complexity as well as to "reflect the fact that there is an upper limit to how far a person will likely travel for most services" (Alasia et al., 2021). The authors assigned ".." when no amenity was available within a given travel radius for a select DB. As a result, not every DB has proximity values for amenities.

In summary, data points may be unavailable either because the supporting databases are incomplete, or because there is no access to the amenity within the specified travel radius. The fact that a sizable portion of DBs don't have an associated proximity measure is not a concern for this project, as we want to segment the measures that are within the scope set by the authors of the PMD.

4.3 Other Data

We linked the IoR dataset to the PMD to add to the cluster profile analysis. This dataset includes a continuous numeric remoteness score for each CSD in Canada. The IoR is equal to zero for the least remote CSD and equal to one for the most remote CSD.

5 Methods

The study begins with an exploratory data analysis to gain familiarity with the dataset and understand its characteristics. R and the `tidyverse` package (Wickham et al., 2019) are used for data handling and analysis. In parallel, clustering algorithms and validation metrics are researched. The clustering tendencies of each amenity are evaluated to assess the natural grouping potential of the data. Different clustering algorithms and intuitive categorization methods, such as quintiles and identifying minima, are applied to each amenity. Computational constraints require clustering analyses to be performed on a single 3% subsample of the dataset for feasibility. The `ClusterCrit` package (Bernard, 2018) is used for cluster validation. Profiling the resulting clusters for each technique provides insights into their robustness and characteristics. The findings lead to conclusions and recommendations for future work.

5.1 Data Preprocessing and Exploration

In order to better understand the structure of the data, we performed an exploratory data analysis (EDA). We analyzed numerical variables (all ten amenities plus DB population) in the PMD using summary statistics, and we counted unique values for categorical variables. We also visualized the distributions of each of the ten amenities using density plots. The distributions showed a strong right skew. To see any improvements, we also visualized the log-transformed proximity values of the amenities. To avoid issues with infinite values during the log transformation, we added a small value of 0.0001 to all the proximity values. This adjustment was necessary because some proximity values were zero. This small value was chosen as a practical compromise to allow the log transformation of zero proximity values without significantly altering the structure of the data or the relative order of the observations. Importantly, this value is significantly smaller than the lower bounds of our proximity measures, ensuring that they essentially remain zero in the transformed scale, hence preserving the original structure. Thus, it ensured the transformed data maintains its integrity for further analysis. Finally, we identified outliers via boxplots, validated them using Rosner's test, and counted them before and after log-transformation.

Logarithmic transforms are a common way to normalize skewed data, as well as to reduce the effect of outliers. Figure 3 shows the curves of a few representative logarithm functions. The natural (base e) logarithmic curve can be broken down into three main sections: a segment near zero where the function is quite vertical, with very large slope, a second segment to the left of 1, where the slope is still very large but not as large as the segment near zero, and a third segment that extends from 1 to infinity where the function becomes more horizontal and behaves as a data compressor. This compressive property is useful for transforming exponential trends into linear ones and exponential seasonality into a linear seasonality. It is also important to note that the log function is the inverse of the exponential function. Therefore, any exponential behaviour in the data will be offset by the log transform. Lastly, it is common to add a constant value ("epsilon") to the data before log transformation in order to prevent the creation of infinite values, which occurs when 0 is log transformed.

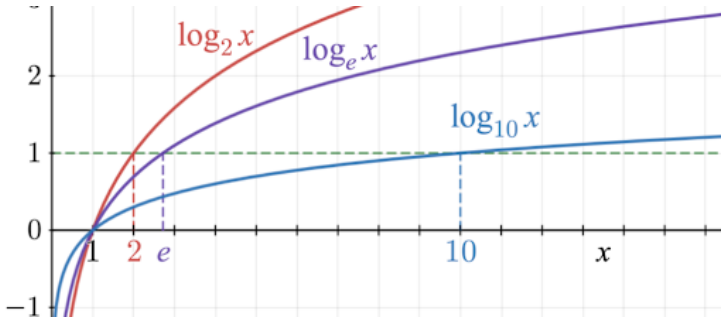

Figure 3: A line plot showing the log function with several different bases.

5.2 Preliminary Clustering Analysis

We evaluated the clustering tendency of each amenity via Visual Assessment of Tendency (VAT) as well as sort plots. VAT works by plotting the distance matrix between all observations in the dataset. Sort plots highlight natural breaks in a continuous vector by sorting the values and then plotting them by index. We used the `fviz_dist` function from the `factoextra` package (Kassambara et al., 2020) to produce the VAT plots.

The quintile method is chosen to reflect the current segmentation in the Proximity Data Viewer (Statistics Canada, 2020a). In this method, data are sorted, and then split into 5 groups, each with the same number of observations. This approach is therefore "blind" to the data, since the actual values are not used in the creation of clusters. We considered this approach as the "base model" to which comparisons will be made.

5.3 Advanced Clustering Analysis and Profiling

We tried different clustering techniques like HDBSCAN, MixAll, PAM, VarSelLCM, and OPTICS to find suitable cutoff values for the amenity proximity measures. Among them, only the results from HDBSCAN, MixAll, MCLUST, and PAM algorithms were useful. VarselLCM wasn't able to produce results with individual amenity proximity measures, and OPTICS produced clusters that overlapped and had subclusters within a cluster. Summaries of the unsuccessful approaches can be found in section A.2 of the Appendix.

In many clustering techniques, the user is required to specify the desired number of clusters (k) to be generated (Kassambara, 2017). To determine the appropriate number of clusters for each clustering technique, various metrics were employed, including the silhouette coefficient, Dunn index, Calinski-Harabasz, and Davies-Bouldin.

The silhouette coefficient is the average of silhouette values for all individual data points. The silhouette value is the difference between the average distance to points within the same cluster and the average distance to points in the nearest neighboring cluster, divided by the largest of the two. The resulting value is bounded between -1 and 1, where values nearer to 0 indicate poor distinction between clusters, and values nearer -1 or 1 are the result of better separation between clusters. Negative values indicate that some points within a cluster are closer to points in a neighbouring cluster than its own, suggesting a 'wrong' assignment. This metric should be maximized. (Rousseeuw, 1987).

The Dunn index reflects the separation between clusters and the distances between observations in the same cluster: it is the ratio of the largest intra-cluster distance and the smallest inter-cluster distances. It ranges from zero to infinity and should be maximized. (Dunn, 1974).

The Davies-Bouldin index calculates, for every pair of clusters, the sum of the within clusters scatter divided by the separation between the clusters. This metric should be minimized: if a value is smaller, that means that either the sum of the cluster scatters is small, and/or the separation between the clusters is large. (Davies & Bouldin, 1979).

The Calinski-Harabasz index calculates the ratio of the variance between clusters and the variance within clusters, where the variance is taken as the difference between cluster centroids and the global centroid for the former and the difference between points in a cluster and their cluster centroid for the latter. This metric should be maximized: higher values are indicative of denser, well separated clusters, since the metric may only be increased by either increasing the distances between a cluster centroid and that of the global centroid, or by decreasing the distances between points in a cluster and its centroid. (Calinski & Harabasz, 1974).

Figure 4 is an example of a set of internal evaluation schemes values per number of clusters for the MixAll clustering technique applied on the Employment proximity values. The number of clusters suggested by each metric is as follows:

- Silhouette coefficient: 2 clusters
- Dunn index: 3 clusters
- Calinski-Harabasz: 8 clusters
- Davies-Bouldin: 8 clusters

To determine the final number of clusters, the majority recommendation from these metrics was considered. As 8 clusters were suggested by the majority of the metrics, this number was chosen.

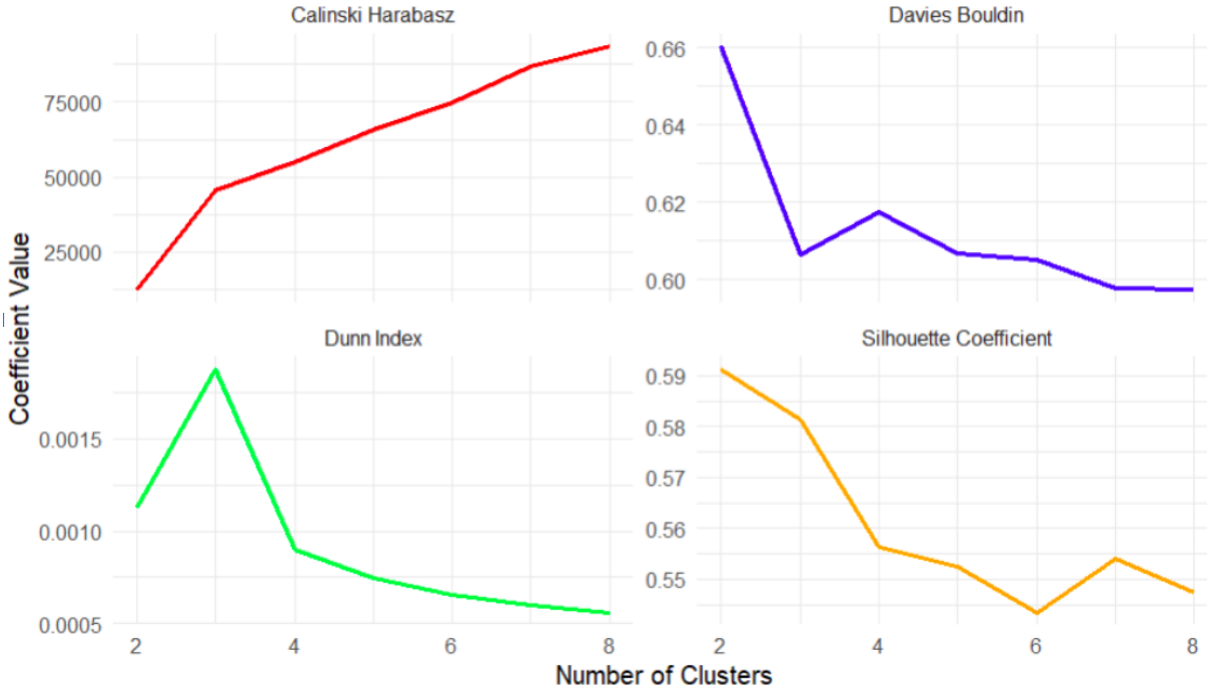

Figure 4: Number of clusters suggested by different metrics for employment amenity in MixAll clustering algorithm. We want to maximize all the metrics except for the Davies-Bouldin, which we want to minimize.

Originally, we expected the metrics to provide insight into which clustering method performed the 'best'. Due to the conflicting recommendations from clustering validation metrics, such as silhouette coefficients, Dunn index, Calinski-Harabasz, and Davies-Bouldin, it became challenging to select a single clustering technique for profiling purposes. In light of this challenge, a different approach was adopted.

Instead of relying on a single algorithm for cluster profiling, multiple techniques that had proven effective in the univariate case and produced satisfactory results were utilized. This approach allowed for a more comprehensive exploration of the data and ensured that the clustering results were robust, and not solely dependent on a single technique.

The success of several algorithms may also be compared intuitively by looking at how the cutoff values divide the log-transformed density plots for each amenity. Successful algorithms will find the cutoff values to be near the "troughs" or "density sparse regions" in these density plots, with the clusters themselves being the "peaks" or "dense regions". Conversely, poorly performing algorithms will miss these troughs by placing the cutoffs randomly or near the peaks.

Following the identification of clusters, an investigation was conducted to examine the profiles of these clusters. This involved comparing various factors, including the number of DBs, median DB population, median IoR, the mode of Census Metropolitan Area type, top province, mode of amenity density, median proximity measure of the clustered amenity, and the corresponding cutoff values.

6 Results

6.1 Data Exploration

6.1.1 Summary Statistics

The PMD contains 489,676 rows, each of which corresponds to a unique DB. Each row contains information about DB population, the encompassing CSD and province, an indicator of amenity density, whether or not the DB is within a census metropolitan area (CMA), plus all of the ten proximity measures. The amenity dense indicator is split into low, medium and high density, with around 5,000 getting an 'F' for "too unreliable to be published." CMA type is divided into four groups: a CMA, not a CMA, a tracted census agglomeration (CA), or an untracted CA ('tracted' in this case refers to whether or not the CA has been subdivided into smaller sections for census purposes). In addition, there is an indicator for each of the ten amenities that relates whether or not the amenity in question resides in the same DB for which the proximity is being calculated. Lastly, there is an indicator for whether or not a DB is considered "amenity dense." We've outlined earlier reasons for which DBs may not have proximity measures. Table 2 shows the amount of DBs that have proximity values for each amenity: Employment has the greatest coverage, at 86.5%, whereas Library has the least at 23%. It is assumed that this is a result of the set travel radius for Employment being much larger, as it covers 10 driving kilometers, whereas libraries are only searched within 1.5 walking kilometers.

Table 2: Counts and percentages of missing values of numerical variables in the PMD.

	DBs with Data Available	Percentage
Employment	423,602	86.5
Pharmacy	178,521	36.5
Childcare	243,964	49.8
Healthcare	300,465	61.4
Grocery	141,063	28.8
Primary Education	225,359	46.0
Secondary Education	141,213	28.8
Library	112,655	23.0
Parks	234,068	47.8
Transit	181,305	37.0
DB Population	487,526	99.6

Table 3 shows the summary statistics for the numerical variables in the PMD, while Table 4 shows the counts for each type of the categorical variables. We see that the 90th percentile of the proximity measures for all amenities are values that are much nearer the smaller end of the domain; the largest value is in Primary Education, at 0.233.

Table 3: Summary statistics of numerical variables in the PMD.

Moments	Employment	Pharmacy	Childcare	Healthcare	Grocery	Pri. Educ.	Sec. Educ.	Library	Parks	Transit	DB Pop.
10% Dec.	0.0001	0.0075	0.0079	0.0002	0.0144	0.0319	0.0374	0.0508	0.0127	0.0011	0
20% Dec.	0.0004	0.0098	0.0152	0.0007	0.0221	0.0416	0.0421	0.0558	0.0203	0.0026	0
30% Dec.	0.0013	0.0146	0.0241	0.0018	0.0289	0.0582	0.0485	0.0624	0.0278	0.0045	5
40% Dec.	0.0030	0.0193	0.0348	0.0032	0.0348	0.0720	0.0586	0.0707	0.0372	0.0067	16
50% Dec.	0.0065	0.0256	0.0476	0.0050	0.0434	0.0900	0.0745	0.0814	0.0481	0.0094	29
60% Dec.	0.0127	0.0341	0.0636	0.0074	0.0555	0.1105	0.0910	0.0960	0.0614	0.0131	45
70% Dec.	0.0217	0.0457	0.0846	0.0111	0.0719	0.1366	0.1141	0.1168	0.0793	0.0184	66
80% Dec.	0.0368	0.0641	0.1167	0.0184	0.0985	0.1720	0.1492	0.1488	0.1050	0.0272	100
90% Dec.	0.0726	0.0983	0.1751	0.0343	0.1540	0.2330	0.2128	0.2106	0.1494	0.0442	173
Min.	0.0000	0.0000	0.0000	0.0000	0.0001	0.0004	0.0005	0.0001	0.0000	0.0000	0
Median	0.0065	0.0256	0.0476	0.0050	0.0434	0.0900	0.0745	0.0814	0.0481	0.0094	29
Mean	0.0254	0.0444	0.0758	0.0137	0.0699	0.1162	0.1040	0.1146	0.0692	0.0181	72
Max.	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	7607
Std. Dev.	0.0491	0.0579	0.0874	0.0279	0.0783	0.0917	0.0869	0.0978	0.0685	0.0270	146
Skew	4.656	4.555	2.807	7.041	3.201	1.963	2.462	3.439	2.824	5.692	8
Kurtosis	38.08	37.81	14.82	95.45	17.83	8.72	11.84	18.48	17.20	72.96	152

Table 4: Summary statistics for categorical variables in the PMD.

Variable	**Counts**
DBs Per Province	
Alberta	66,749
British Columbia	52,850
Manitoba	30,669
New Brunswick	14,345
Newfoundland and Labrador	8,756
Northwest Territories	1,495
Nova Scotia	15,279
Nunavut	792
Ontario	133,214
Prince Edward Island	3,639
Quebec	106,251
Saskatchewan	54,118
Yukon	1,519
CMA Type	
CMA (B)	206,709
Untracted CA (D)	53,061
Tracted CA (K)	16,992
Not a CMA or CA	212,914
Amenity Dense	
Low Density (0)	442,179
Medium Density (1)	37,303
High Density (2)	4,827
Too unreliable to publish (F)	5,367
Suppressed	
Not suppressed (0)	484,309
Info. Suppressed (1)	5,367

6.1.2 Distributions

The distributions of the proximity scores for all amenities are heavily right-skewed, with the majority of the values being grouped near zero, as seen in Figure 3 (left). These distributions then appear to decay smoothly. The strong right-skew results from a relatively small number of high access outliers, which influences the distribution when the measures are normalized.

The presence of these outliers is further demonstrated through the Interquartile Range (IQR), with values lying beyond 1.5 times the IQR above the third quartile or below the first quartile classified as outliers. However, in this paper, the outliers are considered valid since they are not the result of measurement errors. Figure 5 shows the comparison of the density distributions before and after log-transformation. Figures 40 and 41 in section A.3 of the Appendix show these same distributions for all ten amenities. We can already see that the log-transformed distributions are better because the distribution of the proximity values are more normally distributed, and density sparse regions are now visible. Box-Cox and Arcsine transformations were also attempted, but did not yield distributions that were as consistently normally-distributed as those that were log-transformed. It is important to note that log-transforming the proximity measures does not change the structure of the data. In other words, a particular DB 'A' in the non-log-transformed data with less proximity than another DB 'B' will still have less proximity in the log-transformed data. Statistical summary tablee 3 for the log-transformed data can be found in section A.3 of the Appendix as table 30. In contrast, the skew and kurtosis values for the logged data are much smaller, which proves that the log-transformation was successful in reducing the extreme right skew of these proximity values. Kurtosis is a statistical measure that quantifies the shape of a distribution, specifically focusing on the presence and extent of outliers or extreme values. Distributions with a large kurtosis have more tail data than normally distributed data, which appears to bring the tails in toward the mean. Distributions with low kurtosis have fewer tail data, which appears to push the tails of the bell curve away from the mean (Kenton, 2023).

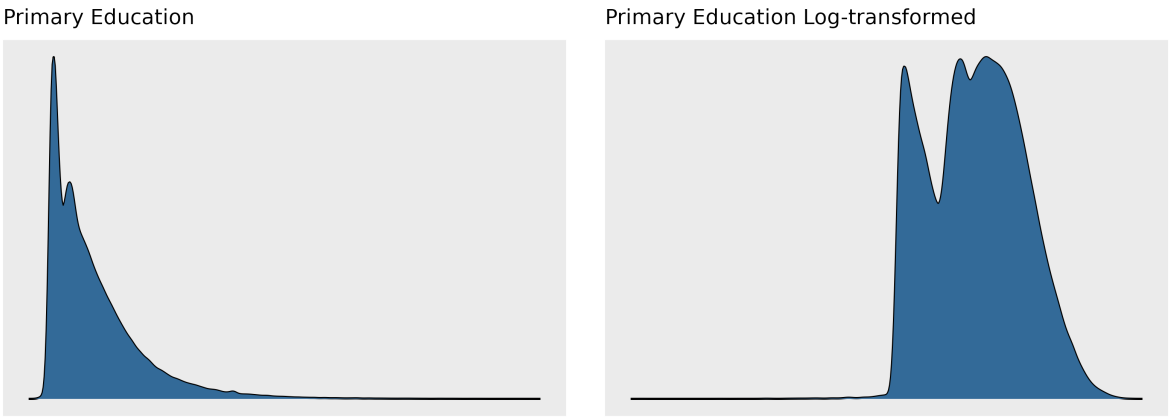

Figure 5: Distribution of the proximity measure to primary education services before and after log-transformation.

In this paper, log-transforming the data is critically important to clustering because of skewness of the data, and helps reveal the underlying structure. Data points near zero in the non-transformed data are "clumped" around particular values, as opposed to being smoothly distributed. This preference for particular values is what creates the miniature peaks that can be seen on the left hand side of the log-transformed density distribution in the employment amenity in Figure 41 of Appendix section A.3. These miniature peaks represent real clusters in the data, and are not simply an artifact of the transformation itself.

Before log-transforming the data, the outliers present had a significant effect on the skew of the data. Many clustering algorithms form clusters based on the distance measure they employ. For instance, algorithms like k-means utilize a squared Euclidean distance, leading to the formation of circular, spherical, or hyperspherical clusters. Outliers can significantly distort the centroids of these clusters, thereby exerting a substantial influence on the overall shape of the clusters. The number of outliers was significantly reduced by log-transforming the data, as

this reduced the relative distance between points. The reduction in the number of outliers after log-transformation can clearly be seen in table 5 (boxplots for visualizing outliers can be found as figures 38 and 39 in Appendix section A.3). In addition to log-transforming the proximity measures to reduce the number of outliers, statistical modeling techniques that are robust to outliers were chosen for clustering. In the future, we can include outliers (values that are much larger than the normal ones) in a single cluster to see what happens when they are treated separately. This will allow us to examine how clustering techniques perform when outliers are not present in the data.

Due to the reduction in the number of outliers as well as the improvement in distribution shape (high right skew to quasi-normal), the following clustering analyses were performed on the individual log-transformed measures as opposed to the original measures. For clarity, it is also pertinent to mention that the proximity measures were clustered individually as opposed to being clustered in concert with other variables. This approach was chosen for its simplicity, but other multivariate clustering approaches should be attempted in the future.

Table 5: The effect of outliers in each amenity in the PMD before and after log-transformation.

	Counts	Percentages	Log Counts	Log Percentages
Employment	45,390	9.27	0	0.00
Pharmacy	13,416	2.74	478	0.10
Childcare	15,397	3.14	140	0.03
Healthcare	31,007	6.33	50	0.01
Grocery	11,904	2.43	794	0.16
Primary Education	10,205	2.08	98	0.02
Secondary Education	8,683	1.77	215	0.04
Library	8,867	1.81	2,295	0.47
Parks	12,703	2.59	910	0.19
Transit	14,165	2.89	3,596	0.73

6.2 Clustering Tendency

The first of the two assessments of clustering tendency was the VAT. In this test, highly clusterable data is visualized having clearly defined rectangles that lie along the diagonal. In contrast, data with low clustering tendency does not have clear rectangles lying along the diagonal, but instead has a jumble of lines and inconsistent colouring. In the VAT plots for the non-transformed data, consistent low clustering tendency is observed. For the log-transformed data, it seems as though the data is semi-clusterable, as there are rectangles, but they are poorly defined and not as distinct as they could be. The VAT plots for all log-transformed amenities in the PMD can be seen in Figure 6.

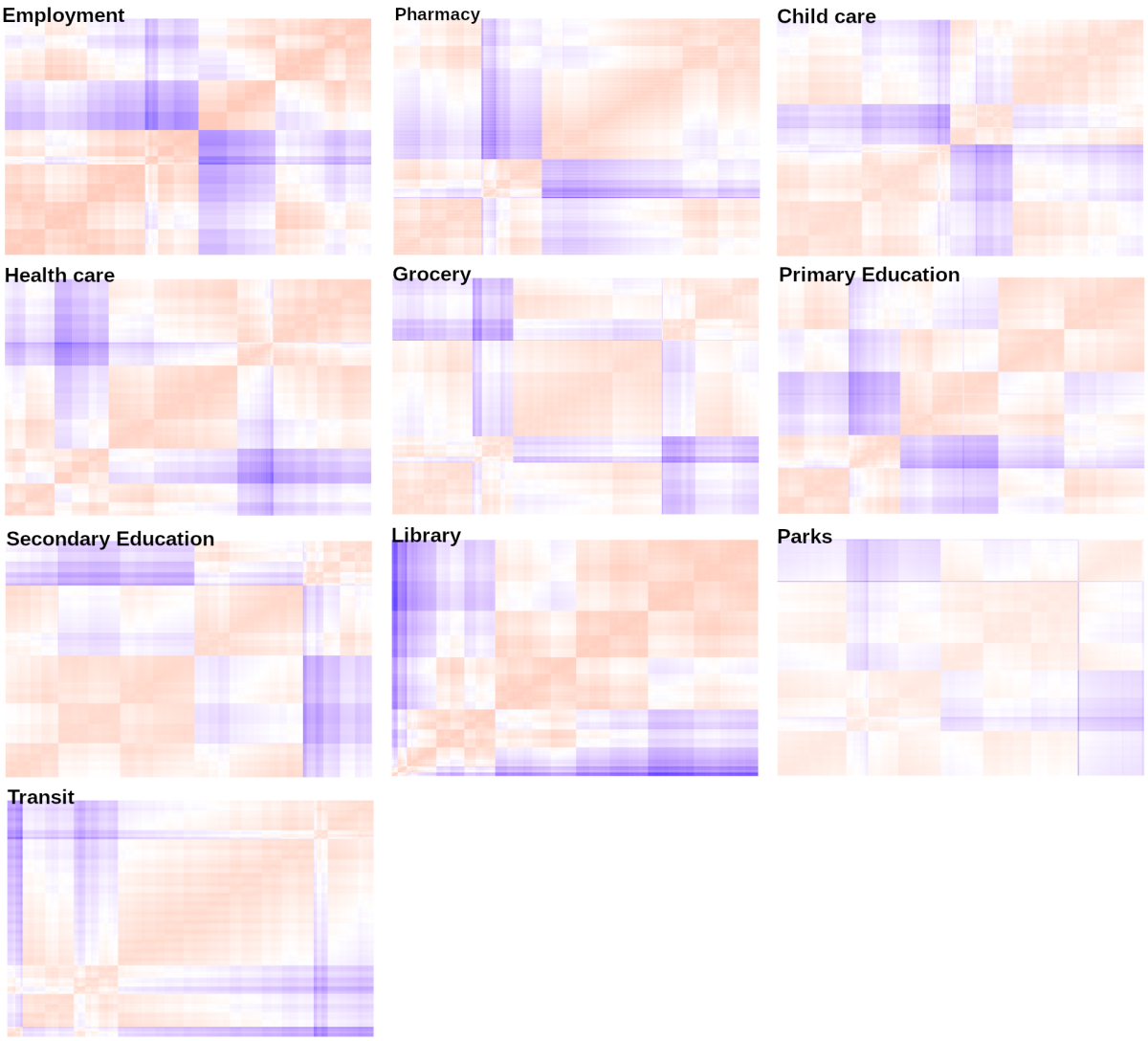

Figure 6: VAT plot results for all log-transformed proximity measures.

The second of the two assessments of clustering tendency was sort plots. If a unidimensional dataset is highly clusterable, then the sort plots will show obvious discontinuous points and changes in slope which separate the clusters. However, this is not observed, as is shown in Figure 7. The non-transformed and log-transformed sort plots for the primary education amenity are shown here for an example. Instead of showing obvious breaks, the lines are smooth. This indicates that there are not any obvious clusters in either the non-transformed or log-transformed data. This has implications for the interpretability of our results, as the cutoff points identified between clusters may be sensitive to changes in the data.

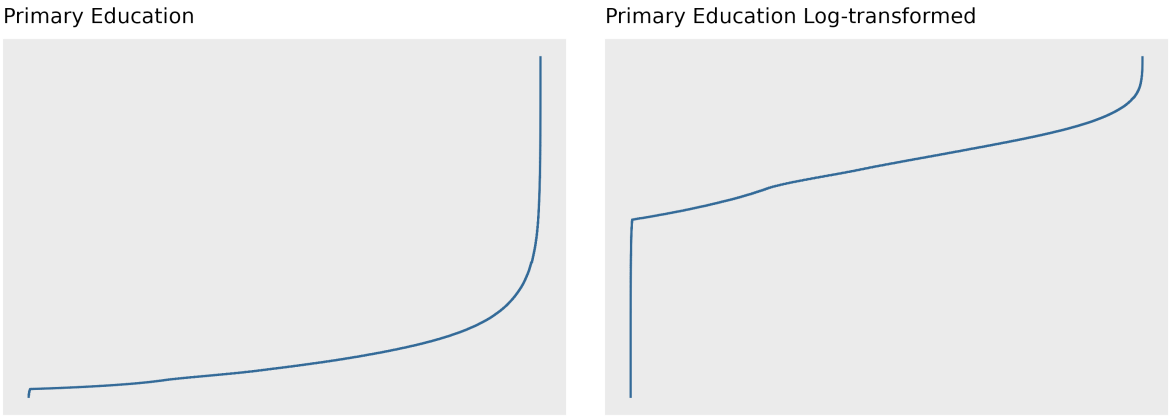

Figure 7: Sort plots of the proximity measure to primary education services before and after log-transformation.

6.3 Quintiles

While easy to understand, the quintile method is a "blind" algorithm, and therefore fails to find good cutoff values. As seen in figure 8, the cutoffs mostly miss the density sparse regions, and are able to find them only in a few cases by mistake.

6.4 Minima Identification

This method is the most intuitive: the minima of the kernel density curves represent density sparse regions, which may be appropriate areas to segment between naturally occurring groups. However, as seen in Figure 8, for many amenities there are large portions of the curve that do not have local minima, resulting in some groups being much larger than others. If choosing cutoff values fully manually, one may choose a point where the curve plateaus or has a flatter slope. Future work may include inflection points on the distribution curve as potential cutoff values.

We used the `density` function in the `stats` package with the default bandwidth and the default gaussian kernels to create the density curves. Changing the bandwidth of the kernel density has an effect on the results: smaller bandwidths result in more density sparse regions and more minima, whereas greater bandwidths result in 'flatter' curves and less minima. Future work should investigate how the size of the bandwidth affects the resulting clusters.

There were many unexpected mathematical minima in the density curves in areas where the density was very small and flat. To retain only the minima that represent density sparse regions amongst regions with higher density, a limiting threshold of 0.001 was set for the difference between neighbouring maxima and minima. Intuitively, if the difference between a local maximum and a local minimum was very small, then the minimum is not representative of a good segmentation point.

Given that the cutoff points selected in this method directly represent density sparse regions, which we intuitively think of as 'gaps' between groups, we expect the validation metrics to be better than those for the 'quintiles' method. We see in Table 7 that, for example, in the case of the Primary Education amenity, only some of the metrics are better, like the silhouette coefficient and the Dunn index, whilst the Davies-Bouldin and the Calinski-Harabasz actually perform worse. This incongruity is a result of what each of these metrics calculates and represents.

6.5 Clustering

6.5.1 Comparison of Algorithms

We applied multiple clustering algorithms, with the specifics outlined in Appendix A.1. Among these, MixAll, HDBSCAN, PAM, and MCLUST emerged as successful. In this context, 'successful' refers to the algorithms that were not only able to run with our univariate data but also provided intuitive results in the form of distinct and non-overlapping clusters. Figure 8 shows the plots of the logged-transformed density distributions with the resulting groups coloured for the representative amenity of Primary Education, while Table 6 outlines specifically the number of clusters each method suggested. We can already see how most of the time, the cutoff values for the different methods don't align with each other: each method finds different points where to segment the data. Even when some methods find more groups than others, aggregating some of the smaller groups doesn't necessarily result in the larger group of another method.

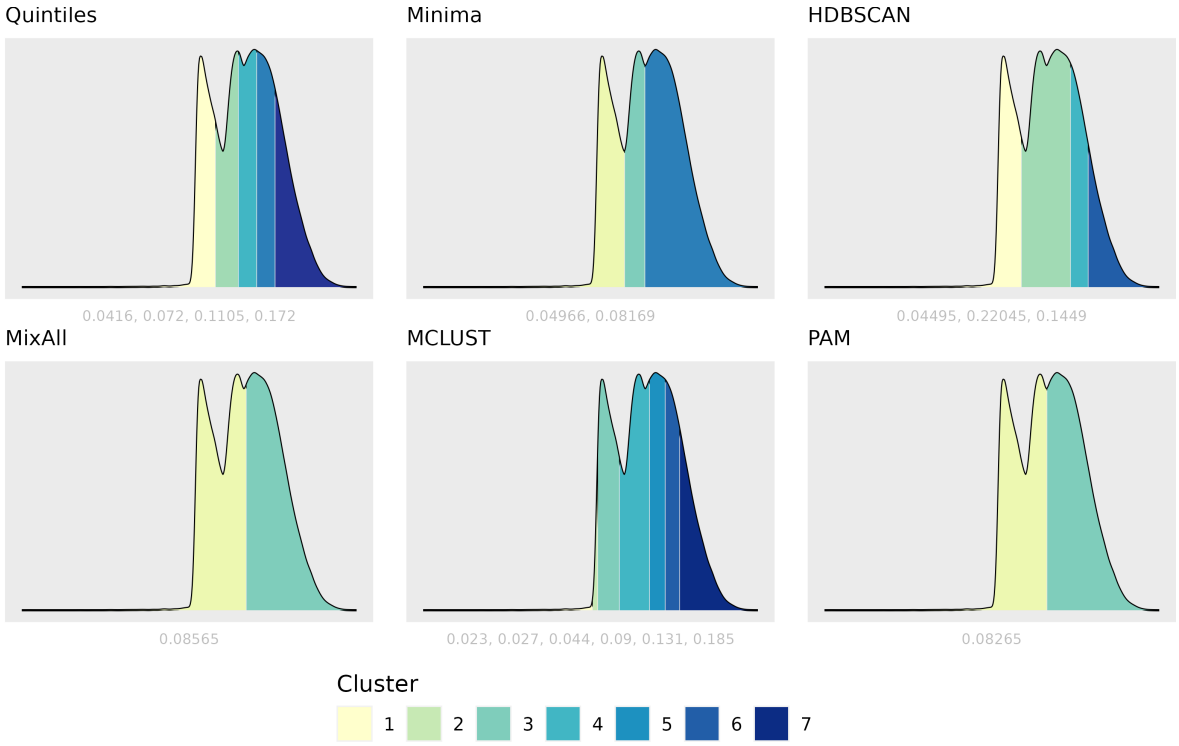

Figure 8: Cutoff values from each segmentation approach displayed on the log-transformed density distributions for the primary education amenity.

Table 6: The number of clusters suggested by all approaches for each amenity in the PMD.

	Emp.	Pharm.	Child.	Health.	Groc.	Pri. Educ.	Sec. Educ.	Lib.	Parks	Transit
Quintiles	5	5	5	5	5	5	5	5	5	5
Minima identification	5	3	3	4	3	3	2	2	3	4
HDBSCAN	2	3	2	2	3	4	3	4	2	2
MixAll	2	2	2	2	2	2	3	2	2	2
MCLUST	9	7	3	4	3	7	8	7	8	3
PAM	2	2	2	2	8	2	4	2	2	2

Table 7: The validation metric values for each clustering approach for the primary education amenity.

	Silhouette	Dunn	Calinski Harabasz	Davies Bouldin
Quintiles	0.47	0.00000	6013	0.71
MixAll	0.58	0.00033	15104	0.67
HDBSCAN	0.33	0.00009	2594	2.69
PAM	0.59	0.00038	15239	0.66
MCLUST	0.46	0.00043	18424	0.65
Minima identification	0.45	0.00015	2853	0.64

6.5.2 Cluster Profiles

Table 8 shows the profiles of each of the clusters defined by each of the successful univariate clustering algorithms for the primary education amenity. We assigned numerical labels to the clusters generated by each algorithm based on their proximity values. The clusters with the lowest proximity were labeled as cluster 1, and the clusters with higher proximity were assigned higher numbers. The other summary variables seem to be roughly correlated with proximity to primary education: as this proximity increases, DB population seems to increase, median IoR seems to decrease, percentage of CMA DBs increases, percentage of DBs in Ontario increases, and the percentage of low amenity dense DBs decreases. All of these trends seem to indicate that proximity to primary education is highest in densely populated cities. Figure 9 summarizes these same clusters by showing the number of DBs and number of people in each cluster for each algorithm.

Table 8: Summary statistics for each cluster found by all approaches for the primary education amenity. DB Population, IoR and proximity value show the median, while CMA Type, Province and Amenity Dense show the mode.

	# of DBs	DB Population	Median IoR	CMA Type	Province	Amenity Dense	Pri. Educ.	Range
Entire Population	225,359 (100.0%)	61	0.12	CMA (65.6%)	Ontario (24.3%)	Low (81.3%)	0.090	0 - 1
Quintiles C1	44,802 (19.9%)	47	0.15	CMA (53.4%)	Ontario (17.4%)	Low (93.1%)	0.032	0 - 0.0416
Quintiles C2	44,830 (19.9%)	51	0.14	CMA (56.1%)	Ontario (19.3%)	Low (89.8%)	0.058	0.0416 - 0.0720
Quintiles C3	45,503 (20.2%)	60	0.12	CMA (65.4%)	Ontario (24.9%)	Low (84.6%)	0.090	0.0720 - 0.1105
Quintiles C4	45,120 (20.0%)	67	0.11	CMA (73.7%)	Ontario (29.9%)	Low (77.3%)	0.137	0.1105 - 0.1720
Quintiles C5	45,104 (20.0%)	77	0.10	CMA (79.3%)	Ontario (29.9%)	Low (61.9%)	0.233	0.1720 - 1
Minima identification C1	57,009 (25.3%)	47	0.15	CMA (52.6%)	Ontario (17.1%)	Low (92.9%)	0.034	0 - 0.0497
Minima identification C2	45,865 (20.4%)	53	0.14	CMA (59.7%)	Ontario (21.2%)	Low (88.4%)	0.066	0.0497 - 0.0817
Minima identification C3	122,485 (54.4%)	69	0.11	CMA (73.9%)	Ontario (28.8%)	Low (73.3%)	0.146	0.0817 - 1
HDBSCAN C1	50,263 (22.3%)	47	0.15	CMA (53.0%)	Ontario (17.2%)	Low (93.0%)	0.033	0 - 0.0449
HDBSCAN C2	113,383 (50.3%)	59	0.12	CMA (64.3%)	Ontario (24.3%)	Low (84.9%)	0.085	0.0449 - 0.1449
HDBSCAN C3	35,780 (15.9%)	70	0.11	CMA (76.4%)	Ontario (29.9%)	Low (71.8%)	0.174	0.1449 - 0.2204
HDBSCAN C4	25,933 (11.5%)	82	0.09	CMA (80.7%)	Ontario (30.0%)	Low (55.9%)	0.285	0.2204 - 1
MixAll C1	107,488 (47.7%)	50	0.14	CMA (56.1%)	Ontario (19.1%)	Low (90.7%)	0.047	0 - 0.0857
MixAll C2	117,871 (52.3%)	69	0.11	CMA (74.3%)	Ontario (29.0%)	Low (72.8%)	0.149	0.0857 - 1
MCLUST C1	518 (0.2%)	127	0.30	None (72.6%)	NovaScotia (10.0%)	Low (100.0%)	0.018	0 - 0.0235
MCLUST C2	1,794 (0.8%)	48	0.15	CMA (53.2%)	Ontario (16.6%)	Low (93.9%)	0.026	0.0235 - 0.0265
MCLUST C3	47,196 (20.9%)	46	0.15	CMA (53.4%)	Ontario (17.4%)	Low (92.9%)	0.033	0.0265 - 0.0444
MCLUST C4	63,570 (28.2%)	54	0.14	CMA (59.1%)	Ontario (20.9%)	Low (88.3%)	0.067	0.0444 - 0.0901
MCLUST C5	40,185 (17.8%)	64	0.11	CMA (70.0%)	Ontario (28.0%)	Low (81.5%)	0.109	0.0901 - 0.1312
MCLUST C6	33,300 (14.8%)	69	0.11	CMA (75.0%)	Ontario (29.8%)	Low (75.0%)	0.154	0.1312 - 0.1850
MCLUST C7	38,796 (17.2%)	78	0.10	CMA (79.9%)	Ontario (30.0%)	Low (60.1%)	0.247	0.1850 - 1
PAM C1	104,320 (46.3%)	50	0.14	CMA (55.8%)	Ontario (19.0%)	Low (90.8%)	0.046	0 - 0.0827
PAM C2	121,039 (53.7%)	69	0.11	CMA (74.1%)	Ontario (28.9%)	Low (73.1%)	0.147	0.0827 - 1

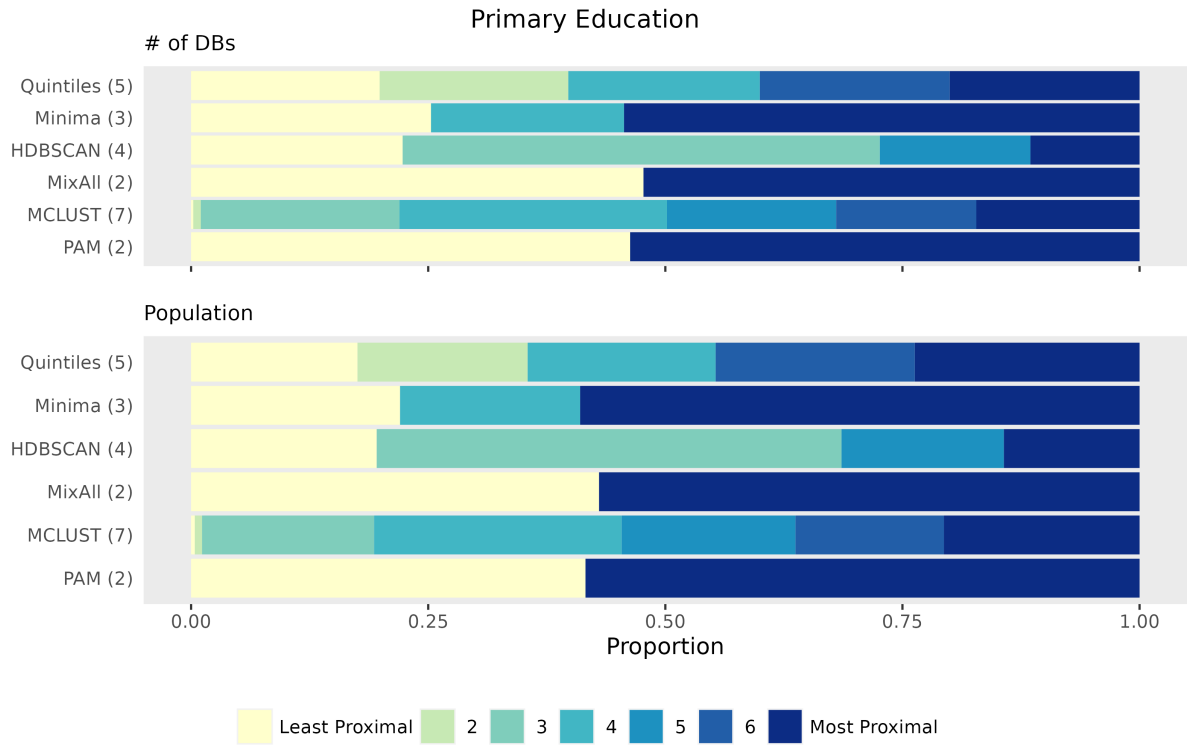

Figure 9: Proportion of DBs and population in each cluster for all approaches for the primary education amenity.

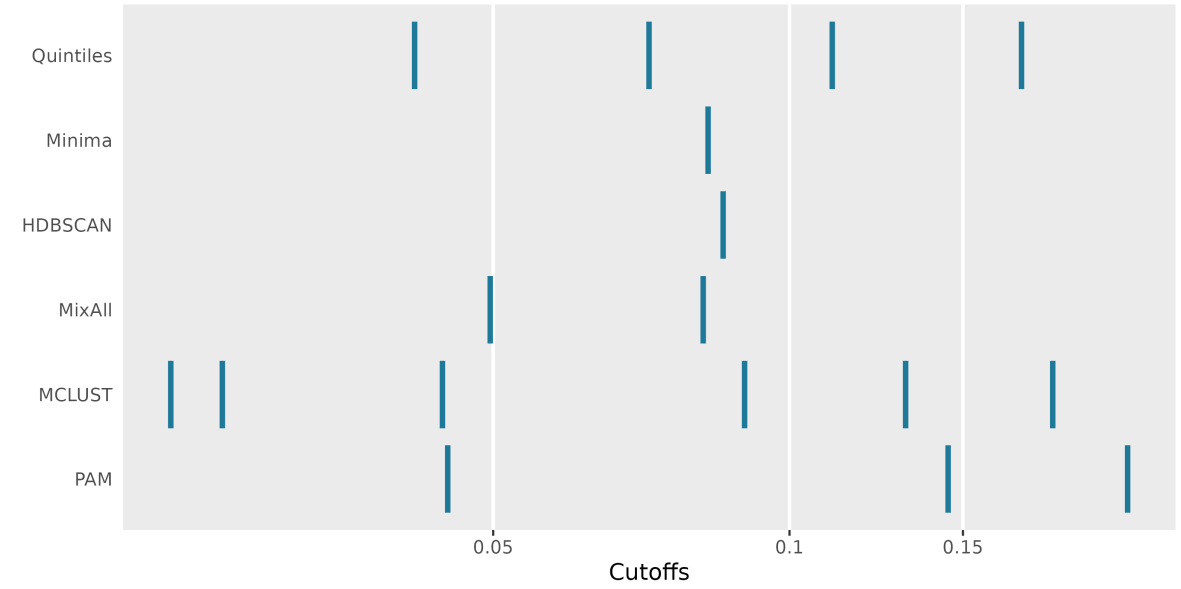

Figure 10: Cutoff values compared for the primary education amenity for all clustering approaches.

7 Analysis

In general, when analyzing various amenities and their clustering results, a pattern emerges where the median proximity measure or cluster range is directly related to the median DB population and median IoR. As the proximity measure or range increases, the median DB population also tends to increase and median IoR tends to decrease. This pattern suggests that areas with higher population tend to be less remote and have better access to amenities.

However, it's important to note that this pattern may not hold true for every clustering technique in all amenities. In the case of amenities like pharmacies, the clustering results obtained from MCLUST may not precisely follow this pattern for the median DB population. This discrepancy could be attributed to MCLUST identifying multiple clusters, some of which may be relatively small in terms of no of observations included in that cluster, resulting in a narrower range of data and affecting the representation of the median.

Overall, while the general trend of increasing median DB population with higher proximity measures holds across multiple amenities and clustering techniques, there may be variations and exceptions in specific cases, particularly when the clustering results include small clusters with limited population.

Moreover, upon examining table 10 to table 26, it becomes evident that all amenities demonstrate a low mode of amenity density for the entire population of Canada. This observation holds true not only for the overall population but also for the majority of individual clusters. This aligns with the findings from the EDA, which indicated that approximately 90% of the DBs in Canada exhibit a low level of amenity density.

7.1 Employment

In Figure 11, we can see that the methods provide mostly different numbers of groups at different cutoff points. MCLUST has a lot more cutoffs, providing nine almost balanced groups, in terms of number of DBs. The fourth cutoff for MCLUST is close to the cutoffs for PAM and MixAll, but the minima identification and the HBDSCAN methods do not find a cutoff at that value. The HDBSCAN only settled on one cutoff, which is different from all other methods: MCLUST's last cutoff comes the closest to it. The minima identification cutoffs seem to all be concentrated almost within the second group of MCLUST cutoffs. Whether they are statistically equal or similar is a different question, whose answer cannot be answered with a visual assessment. It makes sense that PAM and MixAll find similar cutoffs, as the algorithms are similar.

Overall, we can see how the proportions of DBs and proportions of total population in each cluster differs across methods. We see that MixAll and PAM have similar proportions of both in their two clusters. We can also see that in the Quintiles method, there are equal number of DBs per cluster, as that is what the method is by definition. Despite that, we see that the proportions of population are not equal, but each subsequent cluster contains greater proportions of population. This is not exactly the case for every amenity. This suggests that for Employment, the DBs that have higher proximity values also have higher populations on average than the DBs with lower proximity values to employment. This trend is similar for all the methods: groups with larger proximities hold generally a larger proportion of the population than the proportion of DBs.

Table 10 shows the summary statistics for every cluster. We see that the summary statistics differ across groups. For the most part, the median population of each cluster increases as the median proximity score increases, suggesting that areas with higher employment proximity scores generally also have higher populations. The median Index of Remoteness generally decreases as the median proximity score increases.

Table 9 shows the validation metric values for each clustering approach. The best Silhouette coefficient is from the HDBSCAN method, which also has the best Davies Bouldin index. The best Dunn index is from the PAM (followed closely by the MixAll), and the best Calinski Harabasz is from the MCLUST method.

Overall, it seems that for the employment amenity, the cutoff values are not similar across methods and the validation metrics don't have a consensus, suggesting that the proximity measures are not easily grouped algorithmically. However, the groups are somewhat characteristically distinct.

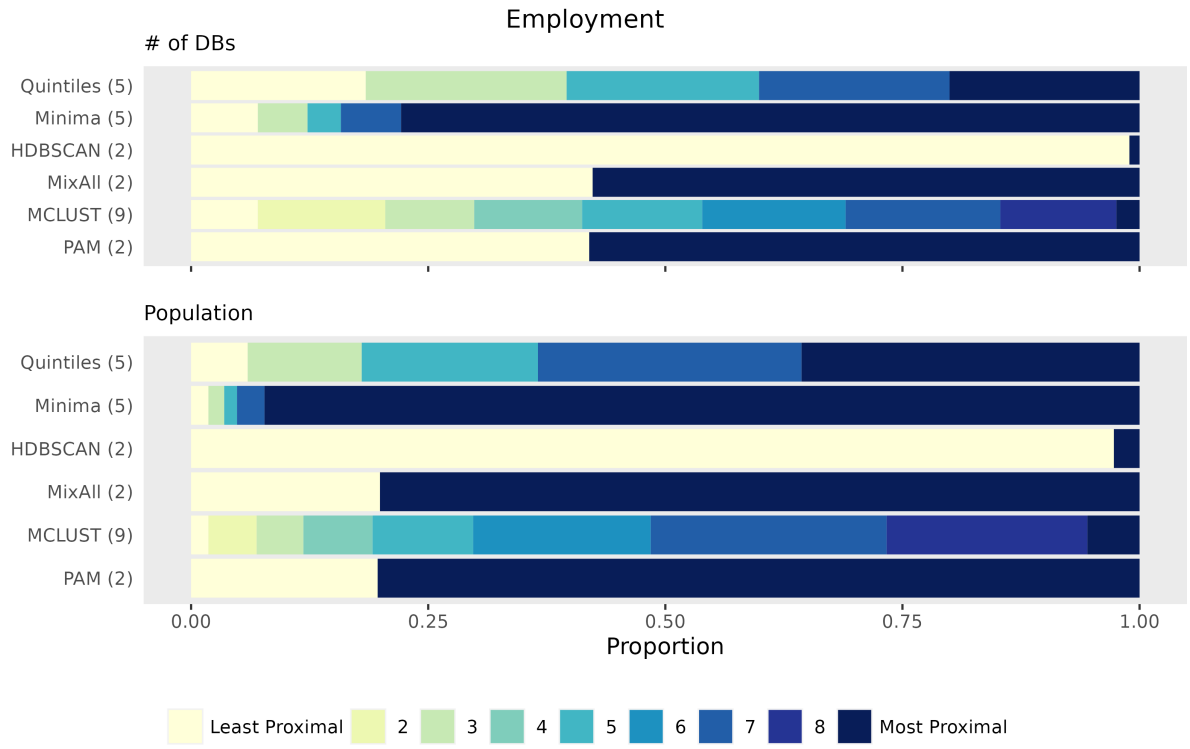

Figure 11: Proportion of DBs and population in each cluster for all approaches for the employment amenity.

Table 9: The validation metric values for each clustering approach for the employment amenity.

	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	0.63	0.00498	36372	0.59
MCLUST	0.59	0.00126	98539	0.56
Minima identification	0.60	0.00014	256	1.01

Table 10: 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.

	# 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
Quintiles C2	89,705 (21.2%)	23	0.24	None (69.6%)	Ontario (9.2%)	Low (99.9%)	0.001	4e-04 - 0.0030
Quintiles C3	85,928 (20.3%)	41	0.20	CMA (34.2%)	Ontario (12.7%)	Low (97.8%)	0.006	0.0030 - 0.0127
Quintiles C4	85,096 (20.1%)	65	0.11	CMA (79.7%)	Ontario (26.9%)	Low (89.9%)	0.022	0.0127 - 0.0368
Quintiles C5	84,859 (20.0%)	83	0.06	CMA (99.5%)	Ontario (37.2%)	Low (62.8%)	0.072	0.0368 - 1
Minima identification C1	29,831 (7.0%)	5	0.32	None (83.6%)	NovaScotia (8.8%)	Low (100.0%)	0.000	0 - 0.0000
Minima identification C2	22,179 (5.2%)	10	0.30	None (81.2%)	NovaScotia (5.6%)	Low (100.0%)	0.000	0.0000 - 2e-04
Minima identification C3	14,893 (3.5%)	10	0.27	None (78.3%)	Ontario (6.7%)	Low (100.0%)	0.000	2e-04 - 3e-04
Minima identification C4	26,887 (6.3%)	17	0.23	None (75.4%)	Ontario (8.6%)	Low (100.0%)	0.000	3e-04 - 5e-04
Minima identification C5	329,812 (77.9%)	50	0.14	CMA (59.1%)	Ontario (21.9%)	Low (87.2%)	0.014	5e-04 - 1
HDBSCAN C1	419,062 (98.9%)	38	0.16	CMA (47.7%)	Ontario (17.9%)	Low (90.9%)	0.006	0 - 0.2298
HDBSCAN C2	4,540 (1.1%)	122	0.03	CMA (100.0%)	Ontario (44.9%)	Med (45.9%)	0.292	0.2298 - 1
MixAll C1	179,334 (42.3%)	16	0.27	None (73.6%)	Ontario (7.2%)	Low (99.9%)	0.000	0 - 0.0036
MixAll C2	244,268 (57.7%)	63	0.11	CMA (73.3%)	Ontario (26.3%)	Low (82.8%)	0.023	0.0036 - 1
MCLUST C1	29,831 (7.0%)	5	0.32	None (83.6%)	NovaScotia (8.8%)	Low (100.0%)	0.000	0 - 0.0000
MCLUST C2	56,902 (13.4%)	10	0.27	None (78.6%)	Ontario (6.3%)	Low (100.0%)	0.000	0.0000 - 4e-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
MCLUST C4	48,188 (11.4%)	27	0.25	None (64.2%)	Ontario (9.3%)	Low (99.7%)	0.002	0.0012 - 0.0033
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
PAM C1	177,804 (42.0%)	16	0.27	None (73.8%)	Ontario (7.1%)	Low (99.9%)	0.000	0 - 0.0035
PAM C2	245,798 (58.0%)	62	0.11	CMA (73.0%)	Ontario (26.2%)	Low (82.9%)	0.023	0.0035 - 1

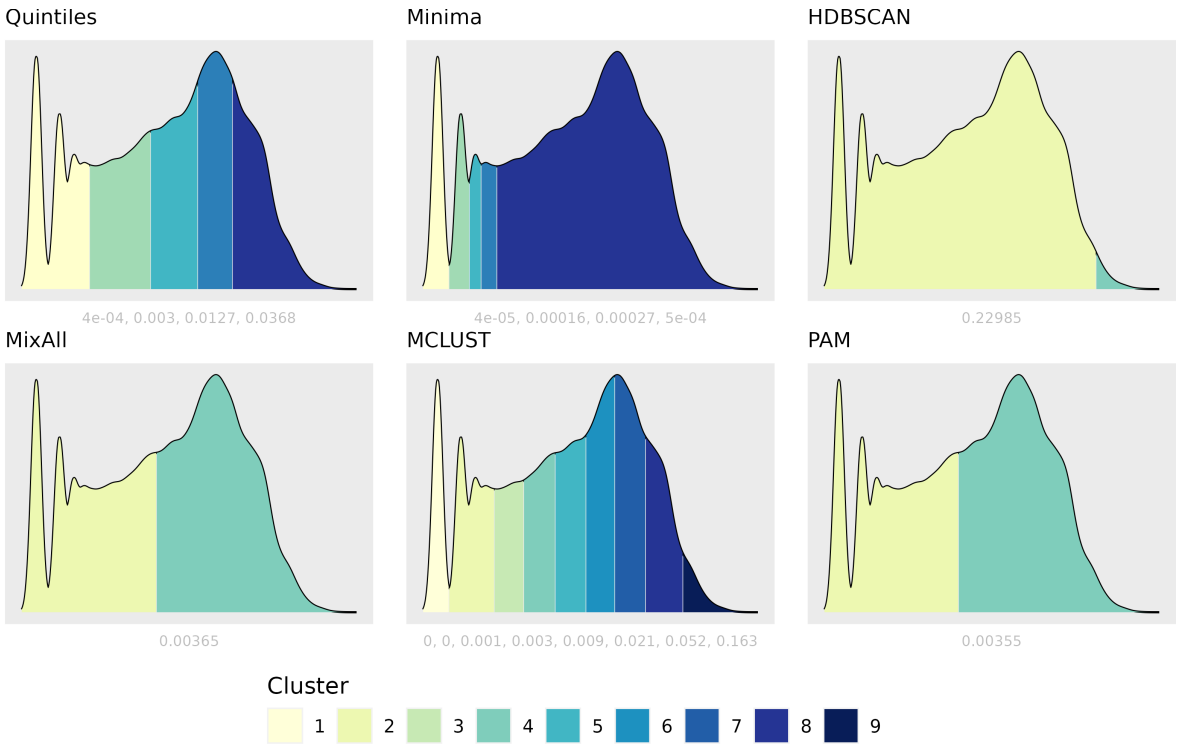

Figure 12: Cut-offs values shown on the log-transformed density plots for all clustering approaches for the employment amenity.

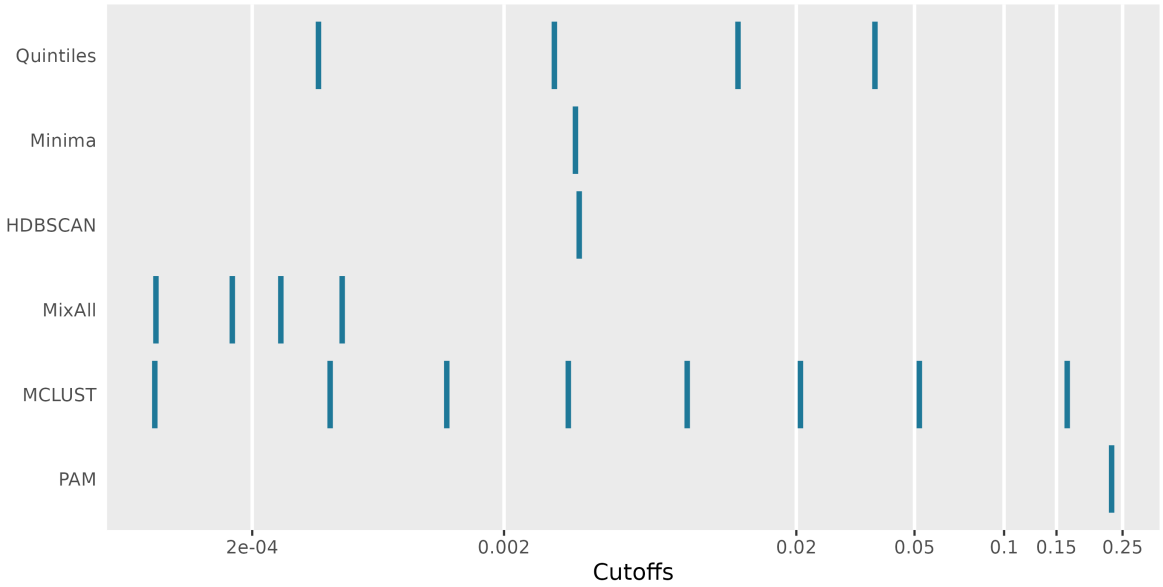

Figure 13: Cutoff values compared for the employment amenity for all clustering approaches.

7.2 Pharmacy

In Figure 14, we see again how for the most part, cutoff values across methods do not align. It seems like the first minima identification and HDBSCAN cutoffs are close with the third MCLUST cutoff, which may indicate robustness. Again, the MixAll cutoff is nearly identical to the PAM cutoff. HDBSCAN's second cutoff is somewhat close to MCLUST's 6th cutoff. In this case, MCLUST settles on some very small groups amongst other more equi-sized. We see that overall, the proportions of population in each cluster are similar to the proportions of DBs in each cluster.

In Table 12, we see that the median IoR for each group is more constant, especially relative to the Employment median IoR. The case is similar for the median population. We also see that in every group, the majority CMA type is CMA. This suggests that the proximity measures for the Pharmacy amenity are not as correlated to the populations or remoteness, and the groups are not characteristically distinct from each other.

Table 11 shows the validation metric values for each clustering approach. The best Silhouette coefficient is tied amongst the MixAll and the PAM methods. MixAll performs the best according to all the other metrics, although the PAM values are pretty similar, which was expected given the results from Table 12 and Figure 14.

Overall, the cutoff values are not similar to each other apart from a few and the groups are not characteristically distinct, suggesting that the pharmacy proximity measures are not distinctly groupable. The method with the best validation metrics is the MixAll algorithm, which only provides 2 groups.

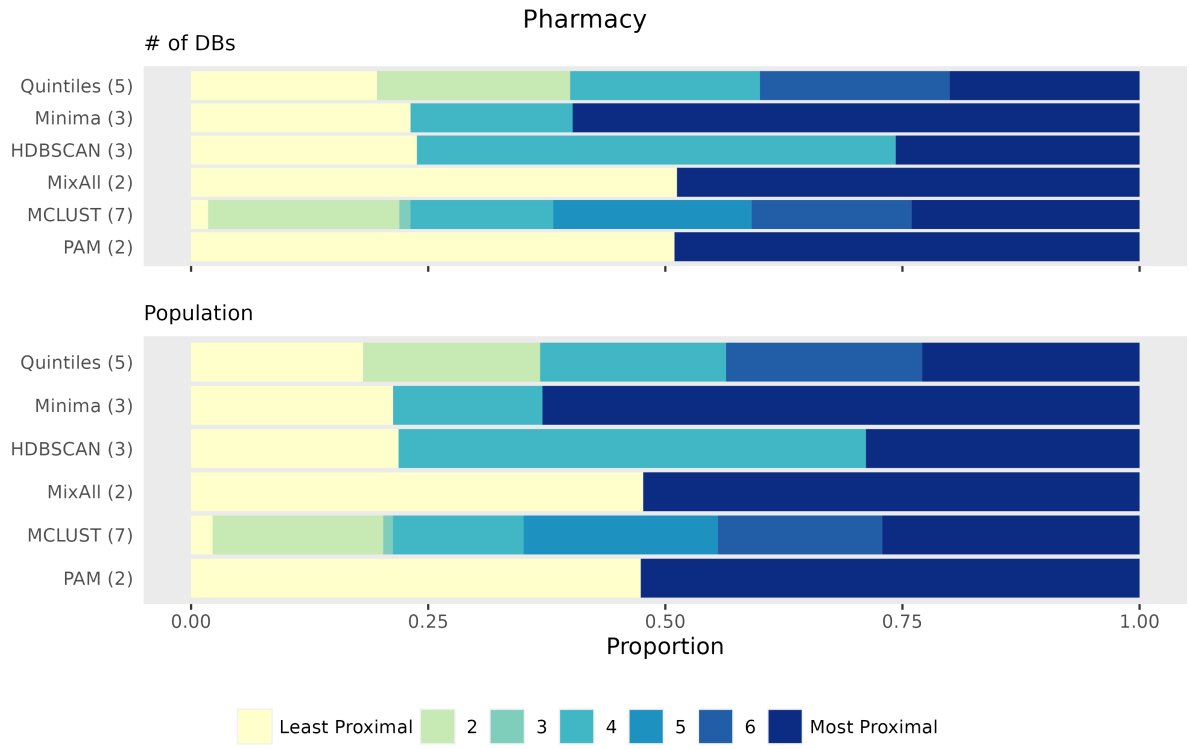

Figure 14: Proportion of DBs and population in each cluster for all approaches for the pharmacy amenity.

Table 11: The validation metric values for each clustering approach for the pharmacy amenity.

	Silhouette	Dunn	Calinski Harabasz	Davies Bouldin
Quintiles	0.43	0.00000	1409	1.01
MixAll	0.59	0.00105	12007	0.66
HDBSCAN	0.44	0.00000	4571	0.80
PAM	0.59	0.00084	11854	0.67
MCLUST	0.48	0.00020	4928	25.17
Minima identification	0.38	0.00010	639	0.80

Table 12: Summary statistics for each cluster found by all approaches for the pharmacy amenity. DB Population, IoR and proximity value show the median, while CMA Type, Province and Amenity Dense show the mode.

	# of DBs	DB Population	Median IoR	CMA Type	Province	Amenity Dense	Pharmacy	Range
Entire Population	178,521 (100.0%)	63	0.11	CMA (71.7%)	Ontario (27.4%)	Low (76.4%)	0.026	0 - 1
Quintiles C1	34,980 (19.6%)	60	0.13	CMA (64.1%)	Ontario (21.8%)	Low (91.5%)	0.007	0 - 0.0098
Quintiles C2	36,365 (20.4%)	60	0.12	CMA (66.8%)	Ontario (24.4%)	Low (88.7%)	0.014	0.0098 - 0.0193
Quintiles C3	35,730 (20.0%)	63	0.11	CMA (72.1%)	Ontario (28.5%)	Low (81.3%)	0.026	0.0193 - 0.0341
Quintiles C4	35,697 (20.0%)	66	0.11	CMA (75.0%)	Ontario (30.9%)	Low (72.0%)	0.046	0.0341 - 0.0641
Quintiles C5	35,749 (20.0%)	71	0.08	CMA (80.3%)	Ontario (31.6%)	Low (48.5%)	0.098	0.0641 - 1
Minima identification C1	41,305 (23.1%)	59	0.13	CMA (63.9%)	Ontario (22.0%)	Low (91.6%)	0.008	0 - 0.0114
Minima identification C2	30,505 (17.1%)	60	0.11	CMA (67.7%)	Ontario (24.7%)	Low (87.9%)	0.015	0.0114 - 0.0195
Minima identification C3	106,711 (59.8%)	66	0.11	CMA (75.8%)	Ontario (30.3%)	Low (67.2%)	0.046	0.0195 - 1
HDBSCAN C1	42,510 (23.8%)	59	0.13	CMA (63.9%)	Ontario (22.0%)	Low (91.6%)	0.008	0 - 0.0118
HDBSCAN C2	90,111 (50.5%)	63	0.11	CMA (71.5%)	Ontario (27.9%)	Low (81.2%)	0.025	0.0118 - 0.0525
HDBSCAN C3	45,900 (25.7%)	70	0.09	CMA (79.2%)	Ontario (31.5%)	Low (52.9%)	0.085	0.0525 - 1
MixAll C1	91,454 (51.2%)	60	0.12	CMA (66.7%)	Ontario (24.1%)	Low (88.6%)	0.013	0 - 0.0265
MixAll C2	87,067 (48.8%)	67	0.10	CMA (77.0%)	Ontario (30.9%)	Low (63.6%)	0.055	0.0265 - 1
MCLUST C1	3,222 (1.8%)	70	0.14	CMA (60.5%)	Ontario (20.7%)	Low (93.5%)	0.006	0 - 0.0064
MCLUST C2	35,979 (20.2%)	59	0.13	CMA (64.3%)	Ontario (22.0%)	Low (91.4%)	0.008	0.0064 - 0.0108
MCLUST C3	2,075 (1.2%)	56	0.13	CMA (61.8%)	Ontario (23.1%)	Low (92.8%)	0.011	0.0108 - 0.0114
MCLUST C4	26,864 (15.0%)	60	0.11	CMA (67.5%)	Ontario (24.7%)	Low (88.1%)	0.015	0.0114 - 0.0181
MCLUST C5	37,376 (20.9%)	62	0.11	CMA (71.8%)	Ontario (28.1%)	Low (82.0%)	0.025	0.0181 - 0.0332
MCLUST C6	30,077 (16.8%)	65	0.11	CMA (74.9%)	Ontario (30.8%)	Low (73.1%)	0.042	0.0332 - 0.0554
MCLUST C7	42,928 (24.0%)	70	0.09	CMA (79.5%)	Ontario (31.5%)	Low (51.8%)	0.088	0.0554 - 1
PAM C1	90,986 (51.0%)	60	0.12	CMA (66.6%)	Ontario (24.1%)	Low (88.7%)	0.013	0 - 0.0263
PAM C2	87,535 (49.0%)	67	0.10	CMA (77.0%)	Ontario (30.9%)	Low (63.7%)	0.055	0.0263 - 1

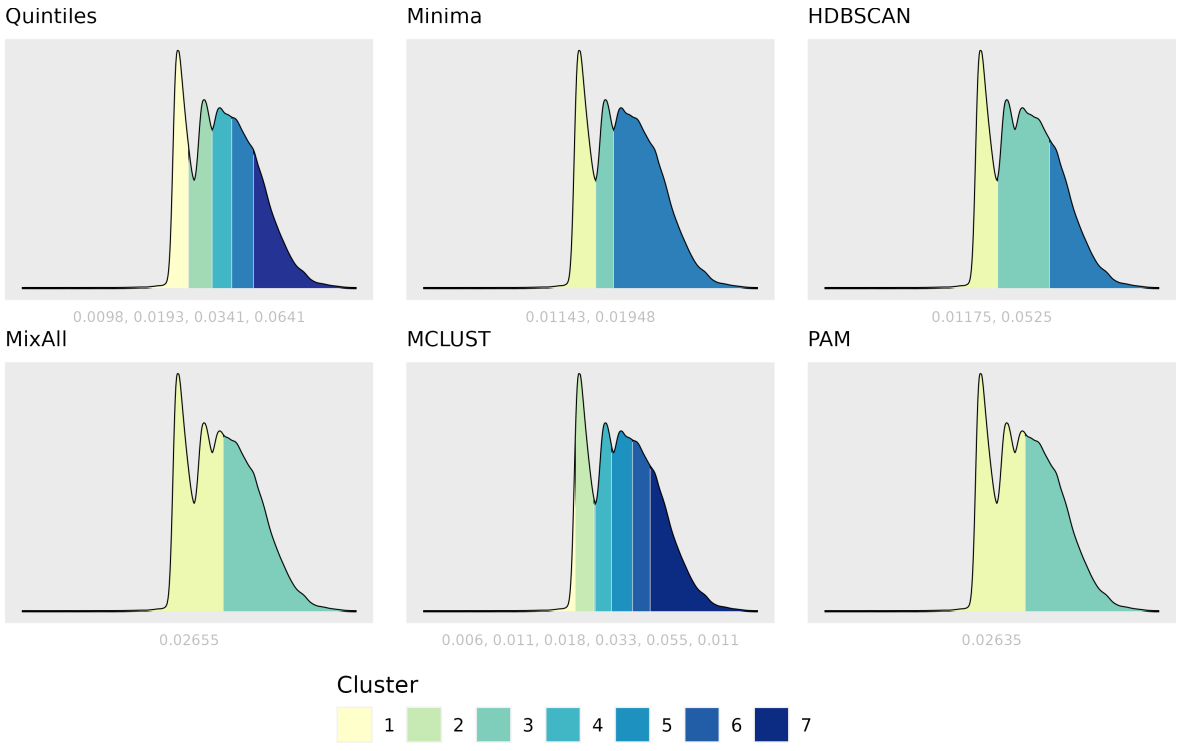

Figure 15: Cut-offs values shown on the log-transformed density plots for all clustering approaches for the pharmacy amenity.

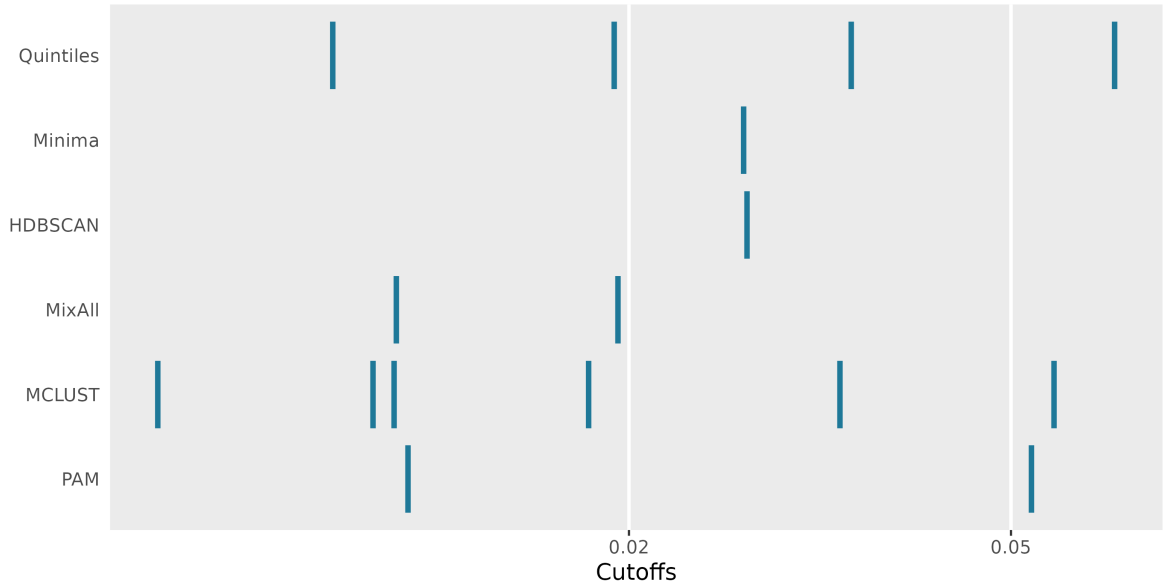

Figure 16: Cutoff values compared for the pharmacy amenity for all clustering approaches.

7.3 Child care

In Figure 17 we see the proportion of DBs and population in each cluster for each method. We again can observe that most cutoff values don't align with each other. Again, the MixAll and the PAM are matching. The first HBSCAN and minima identification cutoffs are otherwise the only ones somewhat aligned. The MCLUST cutoff aligns with the third Quintile cutoff, but since the Quintile method is blind to the data, it isn't of any significance. We see that the proportions of population are somewhat shifted relative to the proportion of DBs, suggesting that there may be slight differences in populations correlated with differences in proximity values.

Table 14 shows the summary statistics for each method and cluster for the Childcare amenity. An anomaly lies in the MCLUST C1: the median population is much larger. The number of DBs is also very small; it may be indicative that other cluster's medians are affected by a large number of DBs with very small populations.The median IoR seem to be dissimilar in different groups as the median proximity value increases. The CMA types of clusters with lower proximity values are of majority type not CMA, whereas those with higher proximity values in majority CMA types.

Table 13 shows the validation metric values for each clustering approach. MCLUST has the best Silhouette coefficient (MixAll runner up) and the best Davies Bouldin index (MixAll runner up). PAM-means has the best Dunn index (MixAll runner up), and MixAll the best Calinski Harabasz value (PAM runner up). These results suggest that since MixAll is consistently in the top 2 relative to the other methods, it may be the best, even though it only finds 2 groups.

Overall, the cutoff values are mostly dissimilar from each other, but the groups do hold different characteristics from each other, suggesting that the proximity values may be clusterable using different methods and/or in cohort with additional variables.

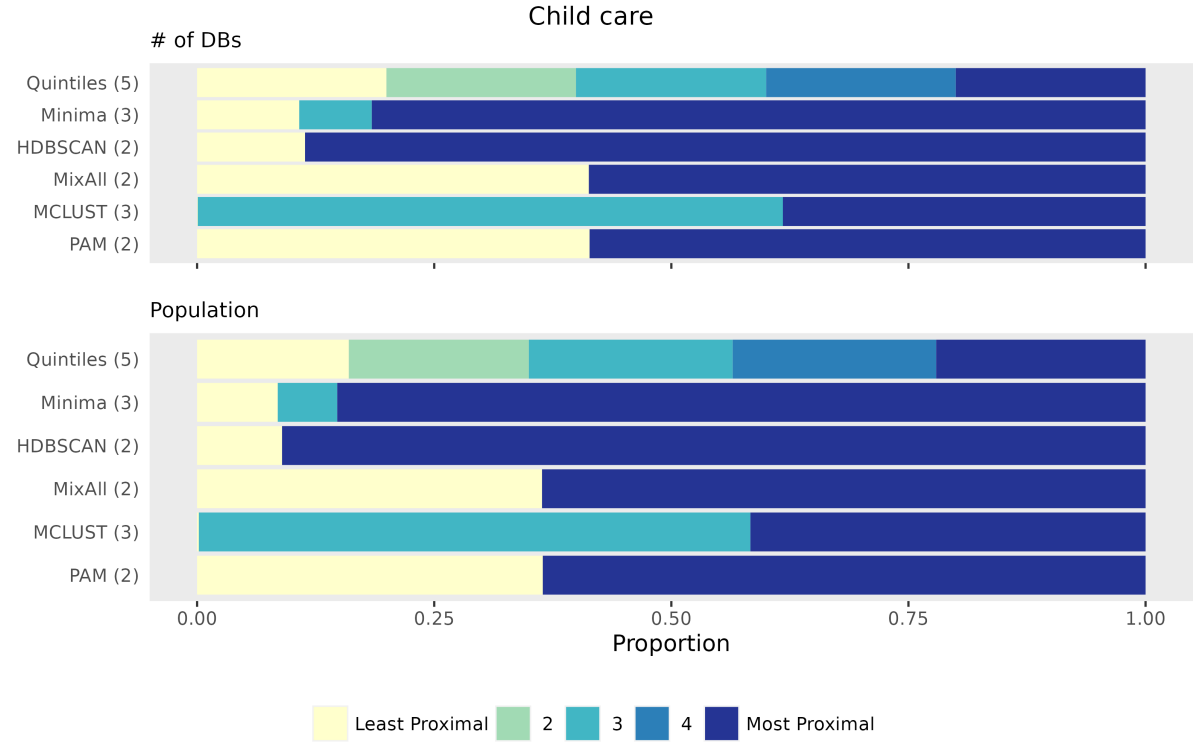

Figure 17: Proportion of DBs and population in each cluster for all approaches for the child care amenity.

Table 13: The validation metric values for each clustering approach for the child care amenity.

	Silhouette	Dunn	Calinski Harabasz	Davies Bouldin
Quintiles	0.44	0.00000	3696	0.79
MixAll	0.58	0.00067	15949	0.67
HDBSCAN	0.44	0.00000	3854	1.77
PAM	0.57	0.00072	15190	0.69
MCLUST	0.60	0.00032	9951	0.64
Minima identification	0.40	0.00011	543	0.76

Table 14: Summary statistics for each cluster found by all approaches for the child care amenity. DB Population, IoR and proximity value show the median, while CMA Type, Province and Amenity Dense show the mode.

	# of DBs	DB Population	Median IoR	CMA Type	Province	Amenity Dense	Childcare	Range
Entire Population	243,964 (100.0%)	62	0.11	CMA (68.3%)	Ontario (23.9%)	Low (82.7%)	0.048	0 - 1
Quintiles C1	48,703 (20.0%)	41	0.18	CMA (46.5%)	Ontario (20.9%)	Low (96.2%)	0.008	0 - 0.0152
Quintiles C2	48,757 (20.0%)	55	0.13	CMA (60.8%)	Ontario (29.4%)	Low (91.3%)	0.024	0.0152 - 0.0348
Quintiles C3	48,909 (20.0%)	66	0.11	CMA (71.5%)	Ontario (28.3%)	Low (84.4%)	0.048	0.0348 - 0.0636
Quintiles C4	48,776 (20.0%)	69	0.11	CMA (77.0%)	Ontario (23.8%)	Low (78.4%)	0.085	0.0636 - 0.1167
Quintiles C5	48,819 (20.0%)	80	0.08	CMA (85.7%)	Quebec (35.3%)	Low (63.4%)	0.175	0.1167 - 1
Minima identification C1	26,274 (10.8%)	40	0.19	CMA (43.6%)	Ontario (18.9%)	Low (96.9%)	0.006	0 - 0.0084
Minima identification C2	18,663 (7.6%)	43	0.16	CMA (49.7%)	Ontario (23.2%)	Low (95.4%)	0.011	0.0084 - 0.0139
Minima identification C3	199,027 (81.6%)	67	0.11	CMA (73.3%)	Ontario (24.6%)	Low (79.7%)	0.062	0.0139 - 1
HDBSCAN C1	27,765 (11.4%)	40	0.19	CMA (43.5%)	Ontario (18.9%)	Low (96.9%)	0.006	0 - 0.0090
HDBSCAN C2	216,199 (88.6%)	65	0.11	CMA (71.5%)	Ontario (24.5%)	Low (80.9%)	0.056	0.0090 - 1
MixAll C1	100,768 (41.3%)	49	0.15	CMA (54.1%)	Ontario (25.3%)	Low (93.5%)	0.016	0 - 0.0363
MixAll C2	143,196 (58.7%)	71	0.11	CMA (78.3%)	Ontario (22.9%)	Low (75.2%)	0.086	0.0363 - 1
MCLUST C1	172 (0.1%)	231	0.29	None (66.9%)	Quebec (8.7%)	Low (100.0%)	0.001	0 - 0.0019
MCLUST C2	150,501 (61.7%)	55	0.14	CMA (60.1%)	Ontario (26.2%)	Low (90.3%)	0.025	0.0019 - 0.0669
MCLUST C3	93,291 (38.2%)	74	0.10	CMA (81.6%)	Quebec (22.7%)	Low (70.5%)	0.120	0.0669 - 1
PAM C1	100,954 (41.4%)	49	0.15	CMA (54.1%)	Ontario (25.3%)	Low (93.5%)	0.016	0 - 0.0364
PAM C2	143,010 (58.6%)	71	0.11	CMA (78.3%)	Ontario (22.8%)	Low (75.1%)	0.086	0.0364 - 1

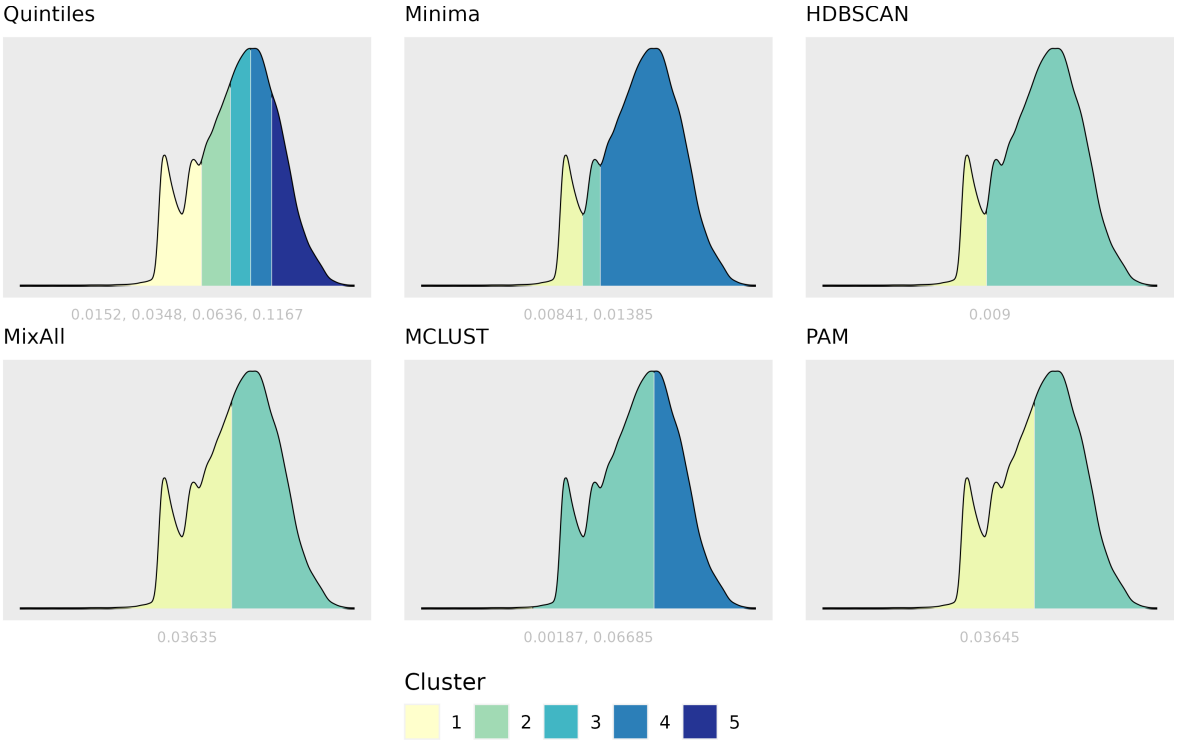

Figure 18: Cut-offs values shown on the log-transformed density plots for all clustering approaches for the child care amenity.

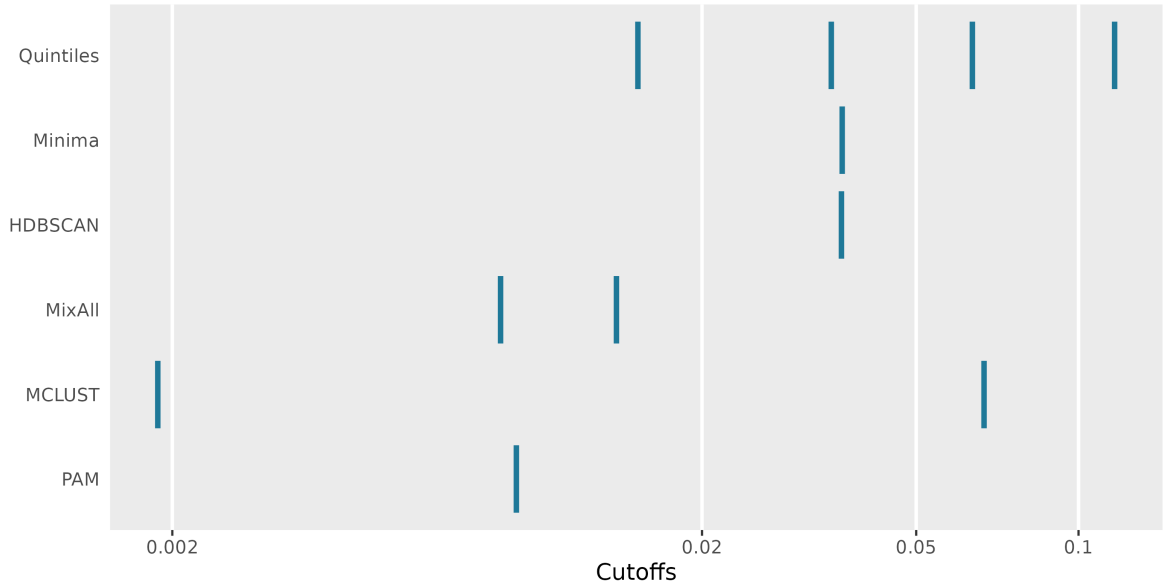

Figure 19: Cutoff values compared for the child care amenity for all clustering approaches.

7.4 Health care

In Figure 20 we see the proportion of DBs and population in each cluster for each method for the Healthcare amenity. In this case, it seems like none of the cutoffs align across methods: even the MixAll and PAM don't exactly agree, although they are still close to each other. The minima identification method finds most clusters at lower proximity values, whereas the HDBSCAN's cutoff is at a high proximity value. It seems like cutoffs for the population proportions are shifted left relative to the proportions of DBs, suggesting a trend with the population values and proximity values.

Table 16 shows the summary statistics for each cluster. In this case, there are many cases where the majority CMA type is not CMA. The median population seem to differ across groups, as well as the median IoR. The proportion of DBs within groups is not constant, as we had seen in Figure 20.

Table 15 shows the validation metric values for each clustering approach. HDBSCAN has the best silhouette coefficient by over 9 points as well as the best Davies-Bouldin result. The PAM has the best Dunn index, although the MixAll method's value follows closely. The MCLUST algorithm has the best Calinski Harabasz measure.

Overall, the cutoff values are mostly dissimilar from each other and the validation metrics mostly don't agree with each other, but the groups do hold different characteristics from each other, suggesting that the proximity values may be clusterable using different methods and/or in cohort with additional variables.

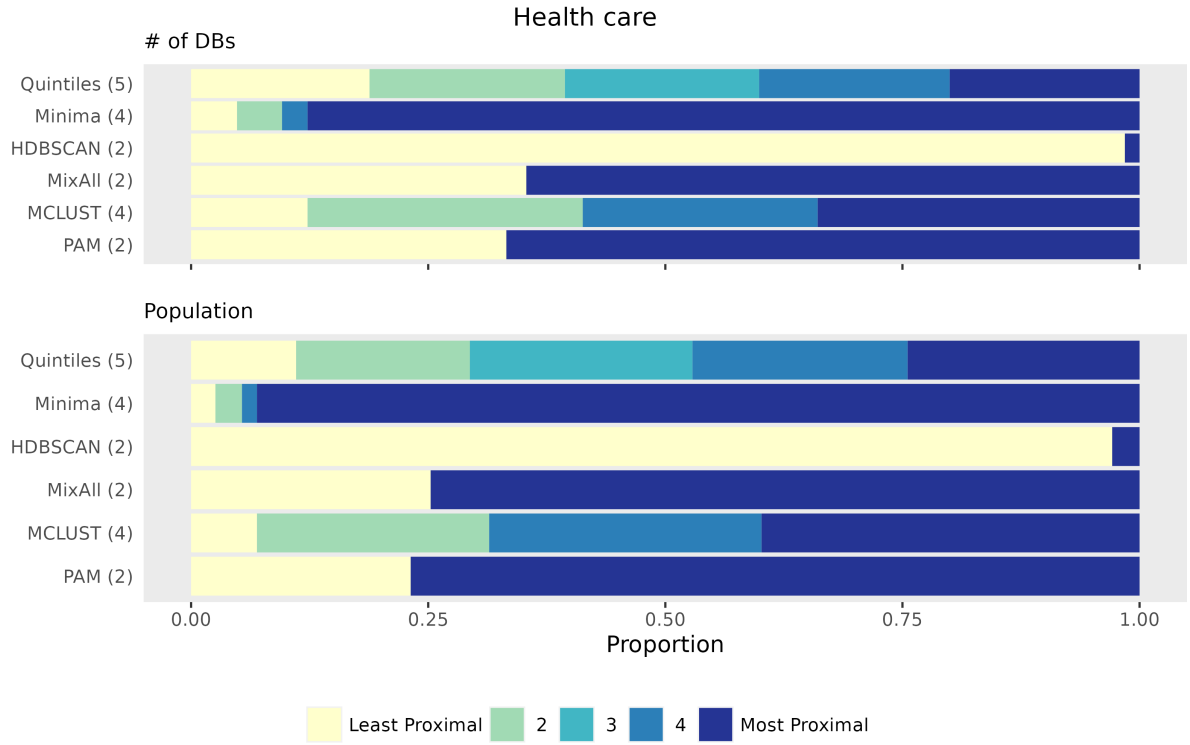

Figure 20: Proportion of DBs and population in each cluster for all approaches for the health care amenity.

Table 15: The validation metric values for each clustering approach for the health care amenity.

	Silhouette	Dunn	Calinski Harabasz	Davies Bouldin
Quintiles	0.39	0.00000	1724	1.13
MixAll	0.58	0.00707	18546	0.68
HDBSCAN	0.73	0.00291	2260	0.35
PAM	0.59	0.00779	18858	0.66
MCLUST	0.52	0.00234	23477	0.64
Minima identification	0.64	0.00015	103	1.01

Table 16: Summary statistics for each cluster found by all approaches for the health care amenity. DB Population, IoR and proximity value show the median, while CMA Type, Province and Amenity Dense show the mode.

	# of DBs	DB Population	Median IoR	CMA Type	Province	Amenity Dense	Healthcare	Range
Entire Population	300,465 (100.0%)	55	0.13	CMA (62.2%)	Ontario (22.8%)	Low (86.0%)	0.005	0 - 1
Quintiles C1	56,525 (18.8%)	26	0.20	None (48.0%)	Ontario (12.1%)	Low (99.7%)	0.000	0 - 7e-04
Quintiles C2	61,910 (20.6%)	48	0.15	CMA (50.9%)	Ontario (15.4%)	Low (97.1%)	0.002	7e-04 - 0.0032
Quintiles C3	61,500 (20.5%)	66	0.11	CMA (69.5%)	Ontario (25.6%)	Low (91.7%)	0.005	0.0032 - 0.0074
Quintiles C4	60,378 (20.1%)	65	0.11	CMA (71.4%)	Ontario (28.7%)	Low (83.4%)	0.011	0.0074 - 0.0184
Quintiles C5	60,152 (20.0%)	71	0.10	CMA (85.2%)	Ontario (31.7%)	Low (58.5%)	0.034	0.0184 - 1
Minima identification C1	14,556 (4.8%)	23	0.19	None (56.5%)	Ontario (14.7%)	Low (99.8%)	0.000	0 - 0.0000
Minima identification C2	14,259 (4.7%)	26	0.20	None (48.7%)	Ontario (12.6%)	Low (99.8%)	0.000	0.0000 - 2e-04
Minima identification C3	8,086 (2.7%)	26	0.20	None (44.4%)	Ontario (11.4%)	Low (99.6%)	0.000	2e-04 - 3e-04
Minima identification C4	263,564 (87.7%)	60	0.12	CMA (66.6%)	Ontario (24.2%)	Low (84.1%)	0.006	3e-04 - 1
HDBSCAN C1	295,804 (98.4%)	54	0.13	CMA (61.7%)	Ontario (22.6%)	Low (87.0%)	0.005	0 - 0.1052
HDBSCAN C2	4,661 (1.6%)	102	0.03	CMA (95.4%)	Ontario (38.3%)	Med (43.5%)	0.142	0.1052 - 1
MixAll C1	106,196 (35.3%)	35	0.19	None (40.2%)	Ontario (13.2%)	Low (98.7%)	0.001	0 - 0.0025
MixAll C2	194,269 (64.7%)	67	0.11	CMA (74.4%)	Ontario (28.1%)	Low (79.0%)	0.010	0.0025 - 1
MCLUST C1	36,901 (12.3%)	25	0.20	None (50.8%)	Ontario (13.2%)	Low (99.7%)	0.000	0 - 2e-04
MCLUST C2	87,182 (29.0%)	44	0.16	CMA (48.3%)	Ontario (14.6%)	Low (97.4%)	0.001	2e-04 - 0.0034
MCLUST C3	74,409 (24.8%)	67	0.11	CMA (70.5%)	Ontario (26.7%)	Low (90.2%)	0.006	0.0034 - 0.0093
MCLUST C4	101,973 (33.9%)	68	0.10	CMA (79.4%)	Ontario (30.5%)	Low (68.1%)	0.022	0.0093 - 1
PAM C1	99,871 (33.2%)	34	0.19	None (41.2%)	Ontario (12.9%)	Low (98.9%)	0.000	0 - 0.0022
PAM C2	200,594 (66.8%)	66	0.11	CMA (73.8%)	Ontario (27.8%)	Low (79.6%)	0.010	0.0022 - 1

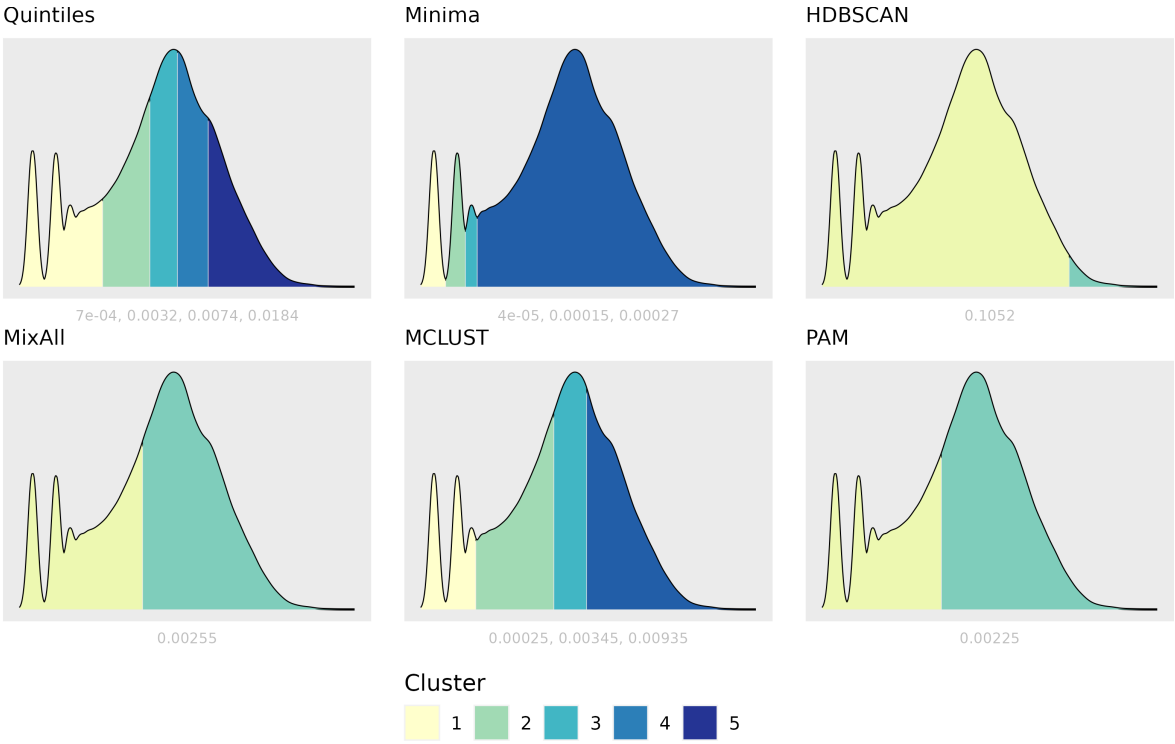

Figure 21: Cut-offs values shown on the log-transformed density plots for all clustering approaches for the health care amenity.

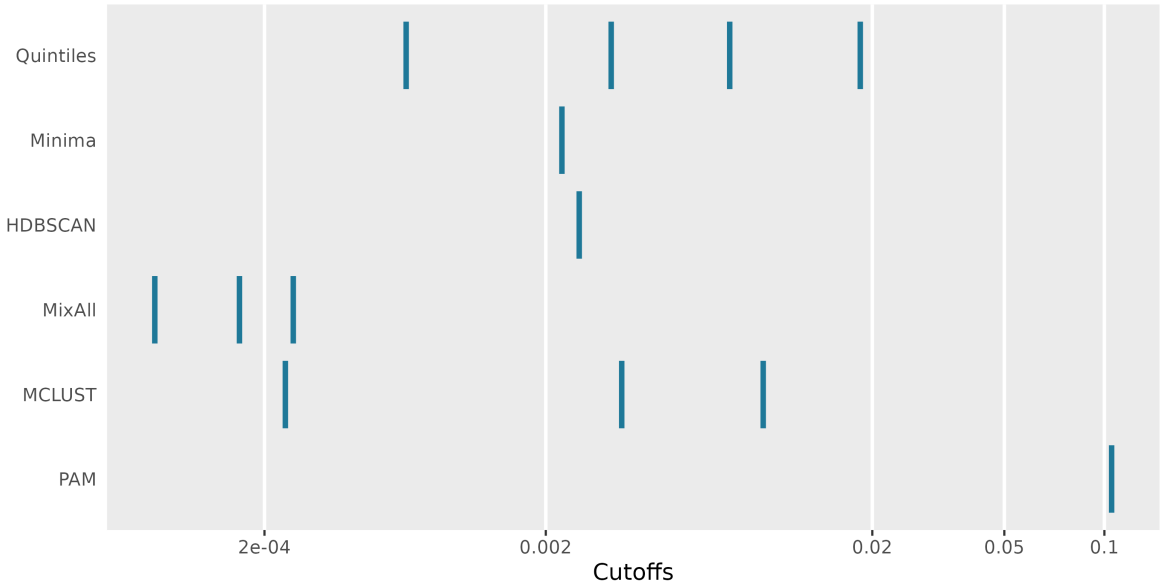

Figure 22: Cutoff values compared for the health care amenity for all clustering approaches.

7.5 Grocery

Looking at the summary statistics in Table 18 for the grocery amenity, we observe that the first cluster cutoffs are quite similar among the minima identification, HDBSCAN, and PAM clustering approaches. The cutoffs range from 0 to 0.0121, 0 to 0.0124, and 0 to 0.0113, respectively. Consequently, these clusters also exhibit similar numbers of DBs and DB population, as shown in Figure 23. However, these techniques do not agree on the cutoffs for the remaining data.

On the other hand, the quintiles technique simply divides the data into five equal parts, which does not align with the cutoffs obtained from any other clustering techniques. Similarly, the MixAll clustering cutoffs do not match with those of any other technique for any of the clusters.

Table 17 provides insights into the performance metrics, such as the Silhouette coefficient and Calinski Harabasz. Silhouette coefficient and Dunn index suggest that the MCLUST algorithm clusters the grocery amenity data better, dividing it into three groups. On the other hand, the Calinski Harabasz and Davies Bouldin indices favor PAM as the better performer, clustering the grocery amenity into eight groups. Interestingly, the first cluster identified by PAM is further divided into two clusters by MCLUST, while the rest of the data, where MCLUST identifies only one cluster, is separated into seven clusters by PAM.

Based on this analysis, it strongly suggests that cluster 1 should have a cutoff range of 0 to 0.0113, as suggested by the PAM technique. This suggestion is supported by the validation from two metrics indicating its better performance and also almost matches with the cutoff range of the first cluster identified by the other two clustering techniques (minima identification and HDBSCAN). However, for the remaining data, none of the techniques agree on the cutoffs, indicating a lack of consensus.

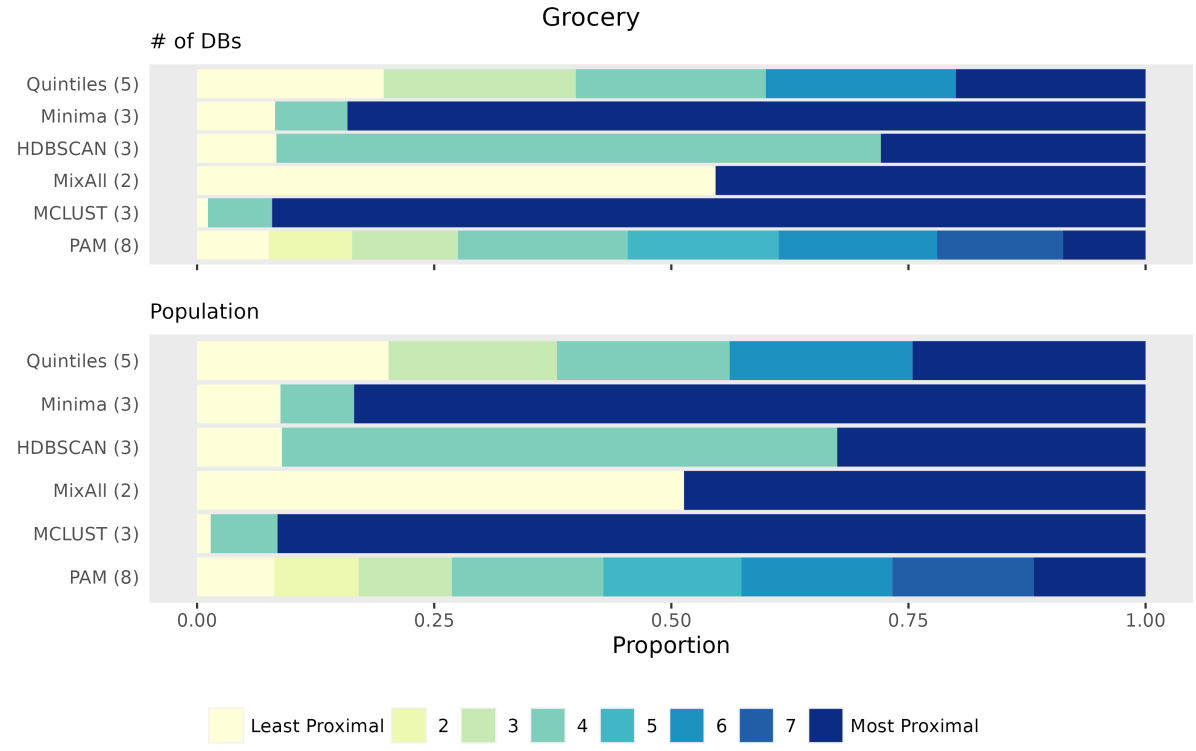

Figure 23: Proportion of DBs and population in each cluster for all approaches for the grocery amenity.

Table 17: The validation metric values for each clustering approach for the grocery amenity.

	Silhouette	Dunn	Calinski Harabasz	Davies Bouldin
Quintiles	0.40	0.00000	1787	0.83
MixAll	0.55	0.00071	7461	0.76
HDBSCAN	0.49	0.00000	1953	1.16
PAM	0.56	0.00070	19255	0.58
MCLUST	0.59	0.00115	1960	0.69
Minima identification	0.38	0.00013	220	0.82

Table 18: Summary statistics for each cluster found by all approaches for the grocery amenity. DB Population, IoR and proximity value show the median, while CMA Type, Province and Amenity Dense show the mode.

	# of DBs	DB Population	Median IoR	CMA Type	Province	Amenity Dense	Grocery	Range
Entire Population	141,063 (100.0%)	61	0.11	CMA (69.3%)	Ontario (25.1%)	Low (70.1%)	0.043	0 - 1
Quintiles C1	27,762 (19.7%)	65	0.11	CMA (71.9%)	Ontario (28.1%)	Low (81.9%)	0.014	0 - 0.0221
Quintiles C2	28,569 (20.3%)	56	0.13	CMA (61.3%)	Ontario (23.2%)	Low (80.6%)	0.029	0.0221 - 0.0348
Quintiles C3	28,266 (20.0%)	58	0.12	CMA (64.6%)	Ontario (25.1%)	Low (74.7%)	0.043	0.0348 - 0.0555
Quintiles C4	28,248 (20.0%)	58	0.11	CMA (68.5%)	Ontario (24.9%)	Low (67.3%)	0.072	0.0555 - 0.0985
Quintiles C5	28,218 (20.0%)	74	0.08	CMA (80.3%)	Ontario (24.5%)	Low (46.2%)	0.154	0.0985 - 1
Minima identification C1	11,600 (8.2%)	67	0.11	CMA (73.5%)	Ontario (28.6%)	Low (82.5%)	0.009	0 - 0.0121
Minima identification C2	10,766 (7.6%)	66	0.11	CMA (72.7%)	Ontario (28.9%)	Low (81.6%)	0.016	0.0121 - 0.0185
Minima identification C3	118,697 (84.1%)	60	0.11	CMA (68.6%)	Ontario (24.5%)	Low (67.9%)	0.053	0.0185 - 1
HDBSCAN C1	11,799 (8.4%)	67	0.11	CMA (73.4%)	Ontario (28.5%)	Low (82.6%)	0.009	0 - 0.0124
HDBSCAN C2	89,898 (63.7%)	58	0.12	CMA (65.2%)	Ontario (25.0%)	Low (76.8%)	0.035	0.0124 - 0.0763
HDBSCAN C3	39,366 (27.9%)	69	0.10	CMA (77.4%)	Ontario (24.5%)	Low (51.2%)	0.126	0.0763 - 1
MixAll C1	77,110 (54.7%)	60	0.12	CMA (66.0%)	Ontario (25.4%)	Low (79.9%)	0.027	0 - 0.0484
MixAll C2	63,953 (45.3%)	64	0.11	CMA (73.3%)	Ontario (24.8%)	Low (58.4%)	0.090	0.0484 - 1
MCLUST C1	1,615 (1.1%)	77	0.11	CMA (66.8%)	Ontario (27.1%)	Low (85.8%)	0.007	0 - 0.0072
MCLUST C2	9,542 (6.8%)	66	0.11	CMA (74.8%)	Ontario (28.9%)	Low (81.9%)	0.009	0.0072 - 0.0117
MCLUST C3	129,906 (92.1%)	61	0.11	CMA (68.9%)	Ontario (24.9%)	Low (69.1%)	0.048	0.0117 - 1
PAM C1	10,674 (7.6%)	68	0.11	CMA (73.8%)	Ontario (28.6%)	Low (82.6%)	0.008	0 - 0.0113
PAM C2	12,384 (8.8%)	66	0.11	CMA (72.6%)	Ontario (28.8%)	Low (81.5%)	0.016	0.0113 - 0.0189
PAM C3	15,758 (11.2%)	55	0.14	CMA (60.9%)	Ontario (23.2%)	Low (81.8%)	0.023	0.0189 - 0.0275
PAM C4	25,218 (17.9%)	57	0.13	CMA (63.1%)	Ontario (24.0%)	Low (79.0%)	0.033	0.0275 - 0.0389
PAM C5	22,468 (15.9%)	58	0.12	CMA (65.2%)	Ontario (25.2%)	Low (73.4%)	0.047	0.0389 - 0.0574
PAM C6	23,558 (16.7%)	58	0.11	CMA (68.2%)	Ontario (24.8%)	Low (67.8%)	0.071	0.0574 - 0.0919
PAM C7	18,749 (13.3%)	67	0.10	CMA (75.1%)	Ontario (25.7%)	Low (56.1%)	0.119	0.0919 - 0.1674
PAM C8	12,254 (8.7%)	85	0.06	CMA (86.1%)	Quebec (31.8%)	Med (40.8%)	0.232	0.1674 - 1

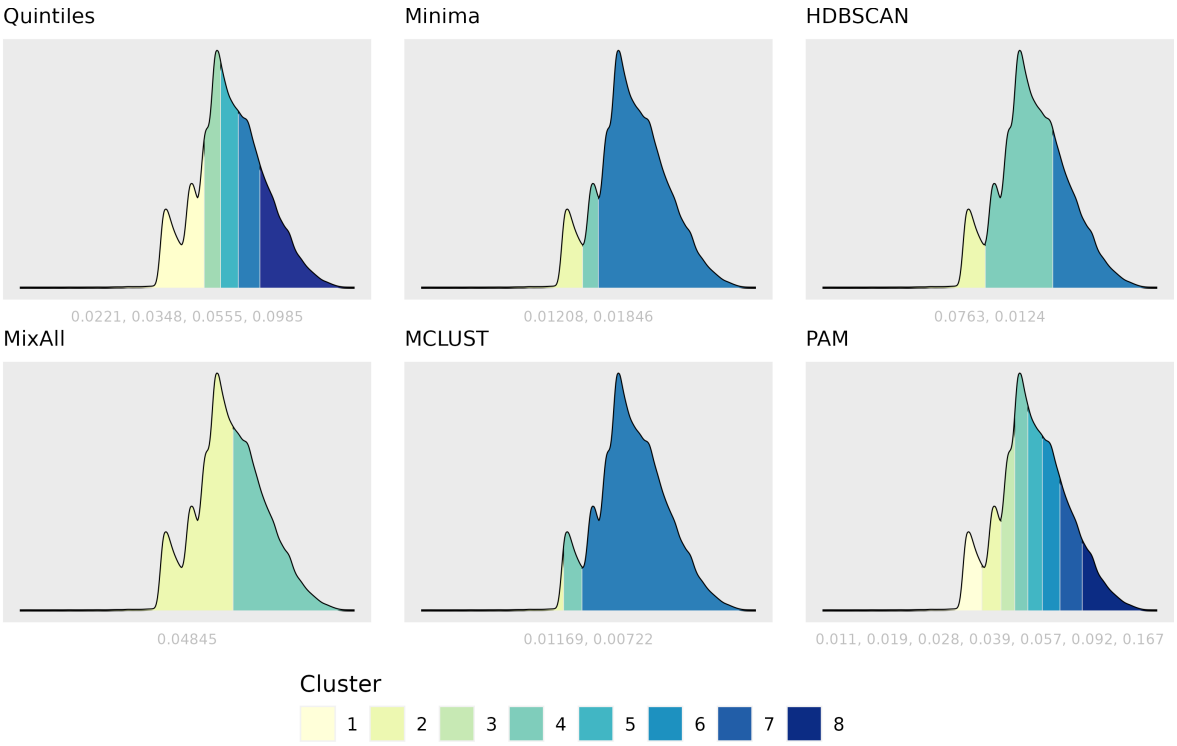

Figure 24: Cut-offs values shown on the log-transformed density plots for all clustering approaches for the grocery amenity.

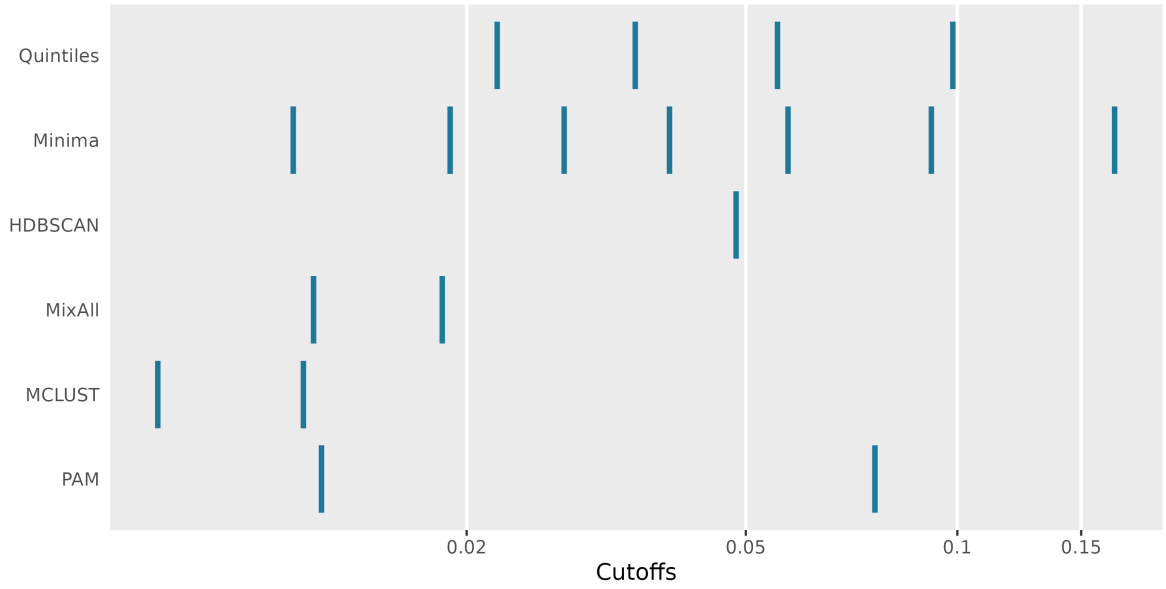

Figure 25: Cutoff values compared for the grocery amenity for all clustering approaches.

7.6 Primary Education

The number of clusters identified for the primary education amenity varies considerably between algorithms. While MixAll and PAM find only two clusters, MCLUST finds seven. While the majority of the cutoff values also vary between algorithms, there is some consistency. For example, the Minima identification method and HDBSCAN seem to agree as to the cutoff value between clusters 1 and 2. Additionally, Minima identification, MixAll and PAM are all able to find the density sparse region around 0.082 (figure 8). Despite the overall variation in cutoff values between approaches, there is not one algorithm that clearly outcompetes the others. This can be seen by comparing the different approaches using common clustering validation metrics (table 7). MixAll and PAM maximize the silhouette coefficient, whereas MCLUST maximizes the Dunn and Calinski Harabasz indices. Finally, the Minima identification method minimizes the Davies Bouldin index. Therefore, it is unclear which algorithm produces the 'best' cutoff values.

Table 8 shows that proximity to primary education is highest in densely populated cities. This is supported by the fact that clusters with higher proximity to primary education also have: a higher percentage of DBs in CMAs, a lower percentage of low amenity dense DBs, and a lower median IoR. Another interesting trend is that clusters with increased proximity to primary education also have more DBs in Ontario. This is likely because Ontario has a higher percentage of DBs in CMAs. Indeed, Ottawa and Toronto are found in Ontario.

While the mode of the categorical variables remains the same for most clusters, MCLUST is able to identify a unique cluster of DBs that does not follow the consensus. MCLUST's cluster 1 seems to be finding a small number of DBs that are rural (non-CMA), decently populated, and spread out fairly evenly across all of Canada. 100% of these DBs have low amenity density, and their proximity to primary education is the lowest of any cluster identified. While this cluster consists of only a very small percentage of the total number of DBs, the profile of this cluster differs significantly enough to be considered a valid, unique cluster.

7.7 Secondary Education

By analyzing the summary statistics in Table 20 for the secondary education amenity, we find that HDBSCAN and PAM show similar cutoffs for cluster 1, ranging from 0 to 0.0576 and 0 to 0.0557, respectively. Consequently, both approaches exhibit similar numbers of DBs and population in Figure 26 for this cluster. However, for the remaining data, the cutoffs do not align. While HDBSCAN identifies two additional clusters, PAM finds three clusters, and the cutoffs for these clusters are different.

Examining the performance metrics in Table 19 for this amenity, the Silhouette coefficient suggests that the MixAll method performs better in clustering. However, the Dunn index and Calinski Harabasz index favor PAM, while the Davies Bouldin index suggests MCLUST.

Based on this analysis, it suggests that cluster 1 for the secondary education amenity should have a cutoff range of approximately 0 to 0.0557, as suggested by PAM. This suggestion is supported by the validation from two metrics and is also consistent with HDBSCAN. However, for the remaining data, there is no consensus among the techniques regarding the cutoffs.

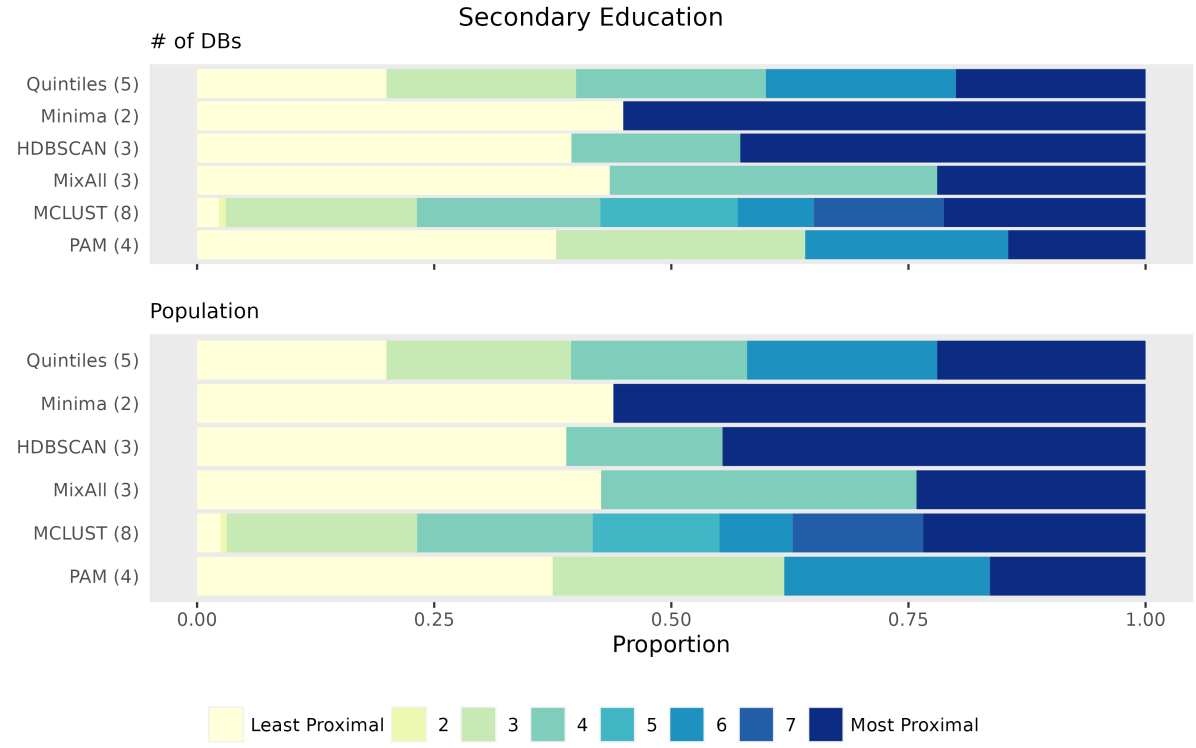

Figure 26: Proportion of DBs and population in each cluster for all approaches for the secondary education amenity.

Table 19: The validation metric values for each clustering approach for the secondary education amenity.

	Silhouette	Dunn	Calinski Harabasz	Davies Bouldin
Quintiles	0.44	0.00000	2686	0.75
MixAll	0.58	0.00028	13920	0.63
HDBSCAN	0.41	0.00018	2710	1.37
PAM	0.56	0.00178	16406	0.62
MCLUST	0.48	0.00040	7936	0.59
Minima identification	0.44	0.00052	2306	0.71

Table 20: Summary statistics for each cluster found by all approaches for the secondary education amenity. DB Population, IoR and proximity value show the median, while CMA Type, Province and Amenity Dense show the mode.

	# of DBs	DB Population	Median IoR	CMA Type	Province	Amenity Dense	Sec. Educ.	Range
Entire Population	141,213 (100.0%)	58	0.14	CMA (62.5%)	Ontario (18.2%)	Low (77.2%)	0.074	0 - 1
Quintiles C1	28,202 (20.0%)	58	0.13	CMA (63.4%)	Ontario (25.6%)	Low (82.8%)	0.037	0 - 0.0421
Quintiles C2	28,226 (20.0%)	55	0.14	CMA (60.1%)	Ontario (23.5%)	Low (82.7%)	0.048	0.0421 - 0.0586
Quintiles C3	28,260 (20.0%)	53	0.15	CMA (58.4%)	Ontario (18.6%)	Low (80.7%)	0.074	0.0586 - 0.0910
Quintiles C4	28,273 (20.0%)	58	0.14	CMA (62.7%)	Ontario (14.2%)	Low (74.0%)	0.114	0.0910 - 0.1492
Quintiles C5	28,252 (20.0%)	67	0.11	CMA (67.6%)	BritishColumbia (23.8%)	Low (66.0%)	0.213	0.1492 - 1
Minima identification C1	63,449 (44.9%)	56	0.14	CMA (61.0%)	Ontario (24.2%)	Low (82.9%)	0.043	0 - 0.0661
Minima identification C2	77,764 (55.1%)	60	0.14	CMA (63.7%)	BritishColumbia (15.5%)	Low (72.6%)	0.122	0.0661 - 1
HDBSCAN C1	55,741 (39.5%)	57	0.13	CMA (61.8%)	Ontario (24.6%)	Low (82.8%)	0.042	0 - 0.0576
HDBSCAN C2	25,135 (17.8%)	53	0.15	CMA (58.0%)	Ontario (19.1%)	Low (81.3%)	0.072	0.0576 - 0.0863
HDBSCAN C3	60,337 (42.7%)	62	0.14	CMA (64.9%)	BritishColumbia (17.8%)	Low (70.4%)	0.143	0.0863 - 1
MixAll C1	61,440 (43.5%)	56	0.13	CMA (61.2%)	Ontario (24.3%)	Low (82.9%)	0.043	0 - 0.0632
MixAll C2	48,748 (34.5%)	56	0.14	CMA (60.8%)	Ontario (16.2%)	Low (77.0%)	0.092	0.0632 - 0.1409
MixAll C3	31,025 (22.0%)	66	0.11	CMA (67.4%)	BritishColumbia (23.2%)	Low (66.5%)	0.204	0.1409 - 1
MCLUST C1	3,273 (2.3%)	62	0.14	CMA (61.3%)	Ontario (24.4%)	Low (84.8%)	0.034	0 - 0.0346
MCLUST C2	996 (0.7%)	62	0.12	CMA (65.4%)	Ontario (28.7%)	Low (80.9%)	0.035	0.0346 - 0.0347
MCLUST C3	28,454 (20.1%)	57	0.13	CMA (63.4%)	Ontario (25.5%)	Low (82.5%)	0.039	0.0347 - 0.0438
MCLUST C4	27,303 (19.3%)	54	0.14	CMA (59.1%)	Ontario (23.0%)	Low (83.0%)	0.051	0.0438 - 0.0618
MCLUST C5	20,457 (14.5%)	53	0.15	CMA (58.3%)	Ontario (18.6%)	Low (81.0%)	0.074	0.0618 - 0.0855
MCLUST C6	11,341 (8.0%)	56	0.15	CMA (61.1%)	Ontario (16.3%)	Low (76.5%)	0.093	0.0855 - 0.1011
MCLUST C7	19,329 (13.7%)	59	0.14	CMA (63.1%)	Ontario (13.7%)	Low (73.3%)	0.120	0.1011 - 0.1434
MCLUST C8	30,060 (21.3%)	67	0.11	CMA (67.5%)	BritishColumbia (23.4%)	Low (66.3%)	0.207	0.1434 - 1
PAM C1	53,475 (37.9%)	57	0.13	CMA (62.0%)	Ontario (24.8%)	Low (82.7%)	0.041	0 - 0.0557
PAM C2	37,062 (26.2%)	54	0.15	CMA (58.7%)	Ontario (18.5%)	Low (80.3%)	0.076	0.0557 - 0.0990
PAM C3	30,221 (21.4%)	59	0.14	CMA (63.5%)	BritishColumbia (15.4%)	Low (72.5%)	0.129	0.0990 - 0.1783
PAM C4	20,455 (14.5%)	69	0.11	CMA (68.9%)	BritishColumbia (25.2%)	Low (64.3%)	0.243	0.1783 - 1

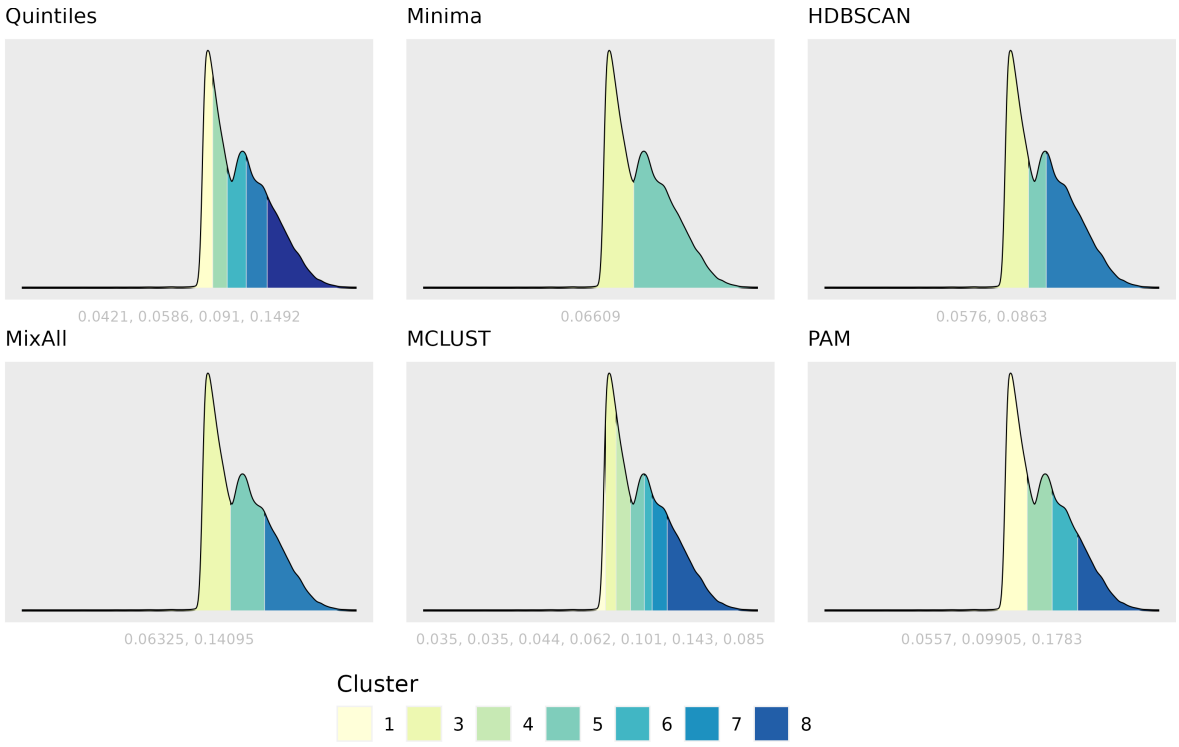

Figure 27: Cut-offs values shown on the log-transformed density plots for all clustering approaches for the secondary education amenity.

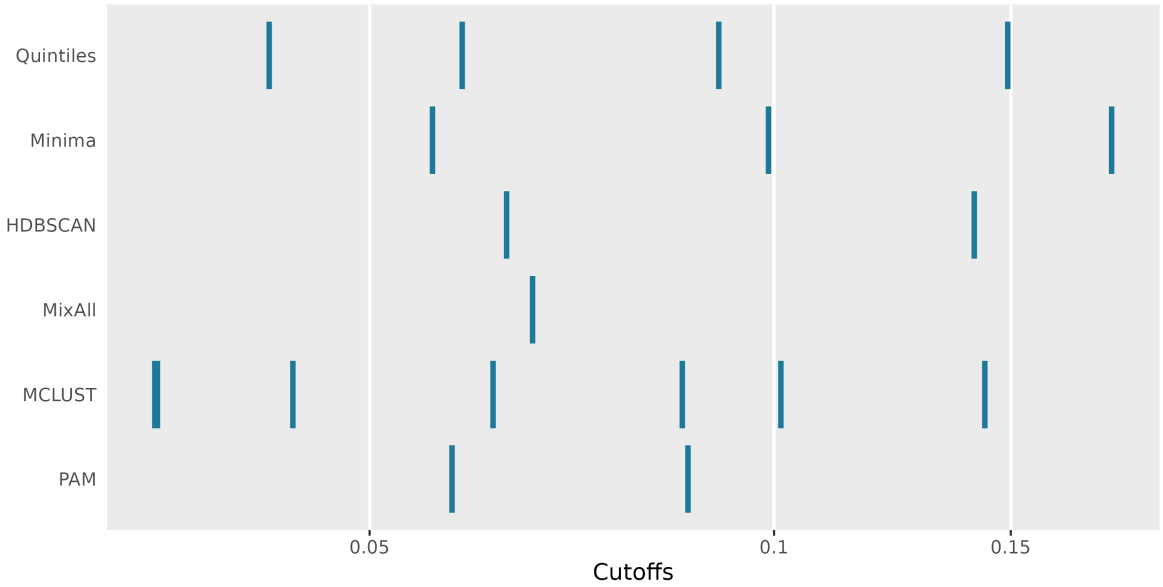

Figure 28: Cutoff values compared for the secondary education amenity for all clustering approaches.

7.8 Library

Analyzing the summary statistics in Table 22 for the library amenity, we observe that MixAll and PAM have similar cutoffs for the first cluster, ranging from 0 to 0.0993 and 0 to 0.0943, respectively. Additionally, both clustering techniques suggest the presence of 2 clusters, indicating a similar cutoff for cluster 2 as well. On the other hand, HDBSCAN identifies the first cluster in the range of 0 to 0.0546, which aligns with MCLUST if we combine the first three clusters of MCLUST with a cutoff range of 0 to 0.0538. The fourth cluster from MCLUST is similar to HDBSCAN if we combine HDBSCAN's cluster 2 and 3. For the remaining data, HDBSCAN identifies only one cluster, while MCLUST finds 3 additional clusters. Apart from MixAll and PAM, none of the other techniques agree with each other in terms of the number of clusters. Furthermore, the cutoffs are not the same, although some of them may be similar by default or when combining multiple clusters into one for comparison with other techniques.

Upon examining the validation metrics in Table 21, we find that the Silhouette coefficient, Dunn index, and Davies Bouldin index suggest that the Minima identification algorithm performs better in clustering this amenity, dividing the data into 2 groups. However, the Calinski Harabasz index suggests that MixAll performs better, also clustering the data into 2 groups, but with significantly different cutoffs compared to the Minima identification algorithm.

Based on this analysis, it suggests that the cutoff for the first cluster may be around 0.0538, and the cutoff for the second cluster may be near 0.0682, as both MCLUST and HDBSCAN find cutoffs in close proximity to these values. The cutoff for the third cluster may be near 0.0993, as PAM, MCLUST, and MixAll identify cutoffs in the vicinity of this value.

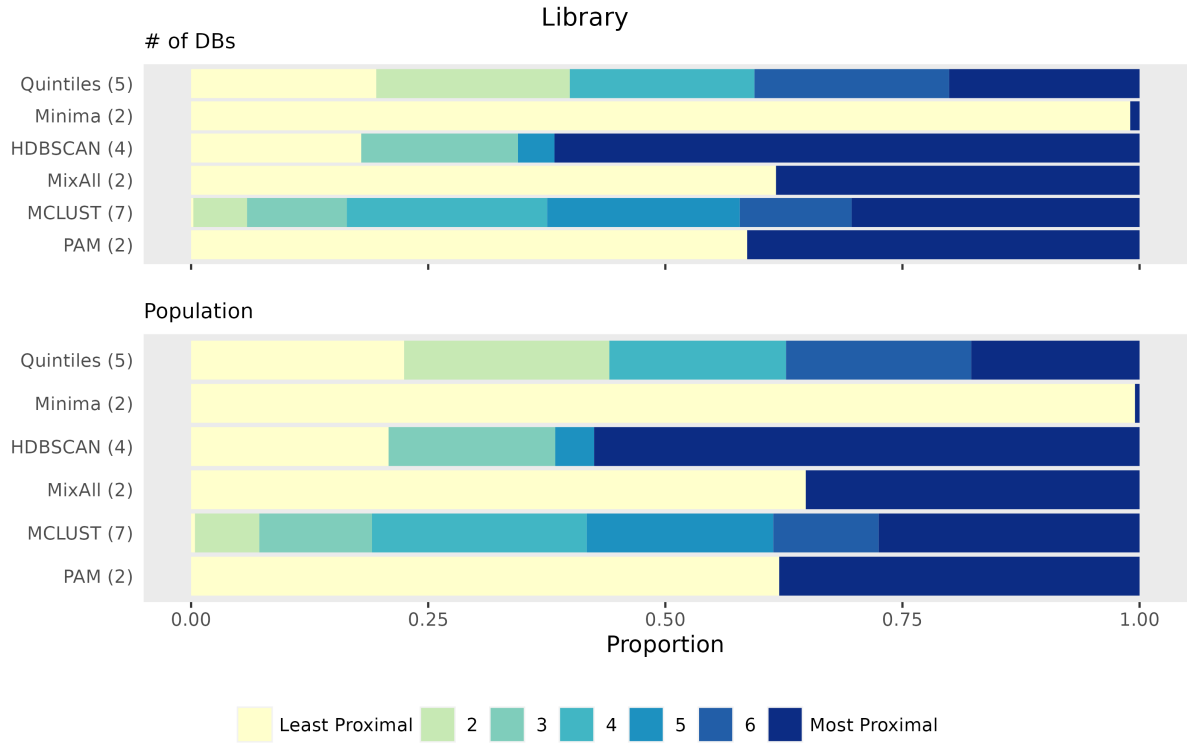

Figure 29: Proportion of DBs and population in each cluster for all approaches for the library amenity.

Table 21: The validation metric values for each clustering approach for the library amenity.

	Silhouette	Dunn	Calinski Harabasz	Davies Bouldin
Quintiles	0.43	0.00000	1386	0.84
MixAll	0.58	0.00243	7174	0.70
HDBSCAN	0.47	0.00028	1395	0.69
PAM	0.57	0.00320	6138	0.73
MCLUST	0.49	0.00167	4323	0.68
Minima identification	0.88	0.01046	1546	0.16

Table 22: Summary statistics for each cluster found by all approaches for the library amenity. DB Population, IoR and proximity value show the median, while CMA Type, Province and Amenity Dense show the mode.

	# of DBs	DB Population	Median IoR	CMA Type	Province	Amenity Dense	Library	Range
Entire Population	112,655 (100.0%)	48	0.14	CMA (54.4%)	Ontario (21.4%)	Low (62.6%)	0.081	0 - 1
Quintiles C1	21,995 (19.5%)	61	0.12	CMA (63.2%)	Ontario (24.0%)	Low (64.7%)	0.050	0 - 0.0558
Quintiles C2	22,988 (20.4%)	56	0.13	CMA (59.6%)	Ontario (23.3%)	Low (63.8%)	0.062	0.0558 - 0.0707
Quintiles C3	21,932 (19.5%)	48	0.15	CMA (52.1%)	Ontario (20.6%)	Low (63.2%)	0.080	0.0707 - 0.0960
Quintiles C4	23,117 (20.5%)	44	0.15	CMA (51.1%)	Ontario (20.4%)	Low (60.1%)	0.116	0.0960 - 0.1488
Quintiles C5	22,623 (20.1%)	33	0.17	CMA (45.8%)	Ontario (18.8%)	Low (61.3%)	0.211	0.1488 - 1
Minima identification C1	111,546 (99.0%)	48	0.14	CMA (54.6%)	Ontario (21.5%)	Low (62.5%)	0.080	0 - 0.6149
Minima identification C2	1,109 (1.0%)	19	0.22	None (49.4%)	Ontario (14.1%)	Low (71.2%)	0.719	0.6149 - 1
HDBSCAN C1	20,209 (17.9%)	62	0.12	CMA (63.3%)	Ontario (24.1%)	Low (64.7%)	0.050	0 - 0.0546
HDBSCAN C2	18,620 (16.5%)	56	0.13	CMA (60.3%)	Ontario (23.5%)	Low (63.9%)	0.060	0.0546 - 0.0658
HDBSCAN C3	4,331 (3.8%)	55	0.13	CMA (58.9%)	Ontario (22.7%)	Low (63.9%)	0.067	0.0658 - 0.0691
HDBSCAN C4	69,495 (61.7%)	42	0.15	CMA (49.9%)	Ontario (20.0%)	Low (61.6%)	0.115	0.0691 - 1
MixAll C1	69,474 (61.7%)	55	0.14	CMA (58.0%)	Ontario (22.5%)	Low (63.8%)	0.063	0 - 0.0993
MixAll C2	43,181 (38.3%)	38	0.15	CMA (48.4%)	Ontario (19.6%)	Low (60.6%)	0.152	0.0993 - 1
MCLUST C1	268 (0.2%)	114	0.30	None (83.6%)	NovaScotia (3.4%)	Low (100.0%)	0.029	0 - 0.0417
MCLUST C2	6,381 (5.7%)	63	0.11	CMA (65.7%)	Ontario (25.2%)	Low (64.4%)	0.048	0.0417 - 0.0488
MCLUST C3	11,854 (10.5%)	60	0.12	CMA (63.3%)	Ontario (23.9%)	Low (64.1%)	0.051	0.0488 - 0.0538
MCLUST C4	23,791 (21.1%)	56	0.13	CMA (60.2%)	Ontario (23.5%)	Low (63.9%)	0.060	0.0538 - 0.0682
MCLUST C5	22,881 (20.3%)	49	0.14	CMA (53.2%)	Ontario (20.9%)	Low (63.1%)	0.079	0.0682 - 0.0927
MCLUST C6	13,289 (11.8%)	45	0.15	CMA (50.7%)	Ontario (19.9%)	Low (61.5%)	0.103	0.0927 - 0.1163
MCLUST C7	34,191 (30.4%)	36	0.16	CMA (47.6%)	Ontario (19.4%)	Low (60.6%)	0.170	0.1163 - 1
PAM C1	66,047 (58.6%)	55	0.13	CMA (58.5%)	Ontario (22.7%)	Low (63.9%)	0.062	0 - 0.0943
PAM C2	46,608 (41.4%)	38	0.15	CMA (48.5%)	Ontario (19.6%)	Low (60.8%)	0.146	0.0943 - 1

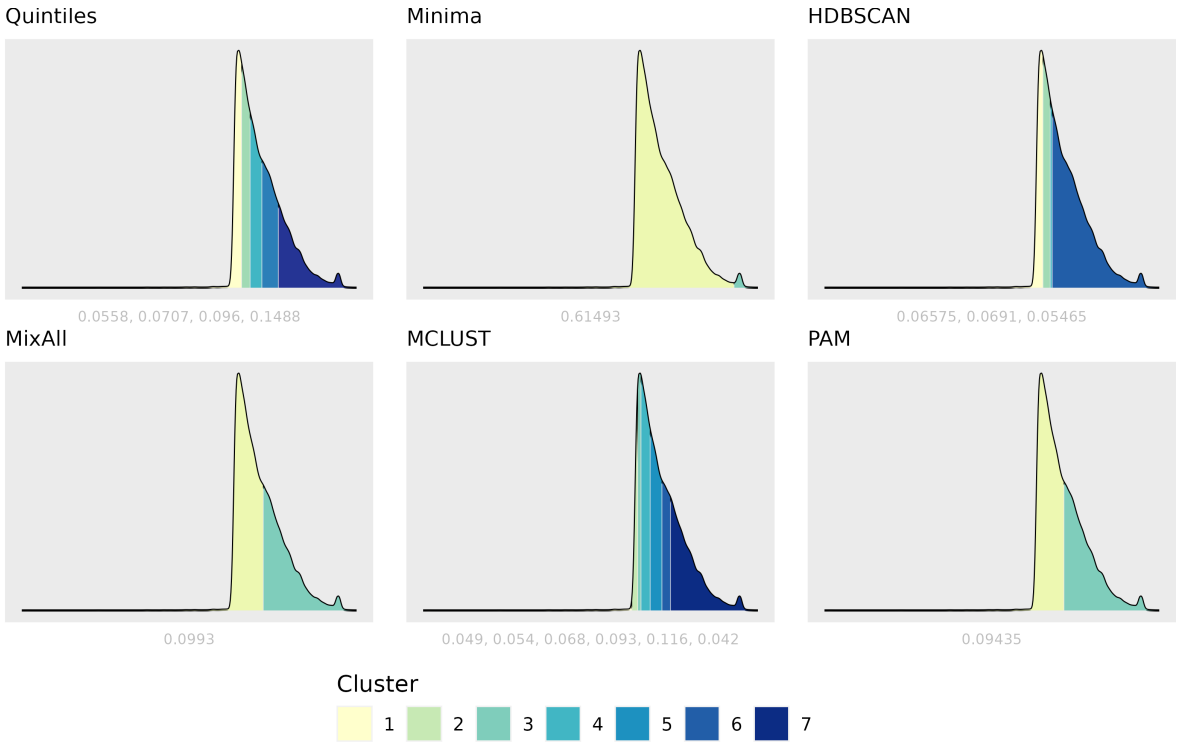

Figure 30: Cut-offs values shown on the log-transformed density plots for all clustering approaches for the library amenity.

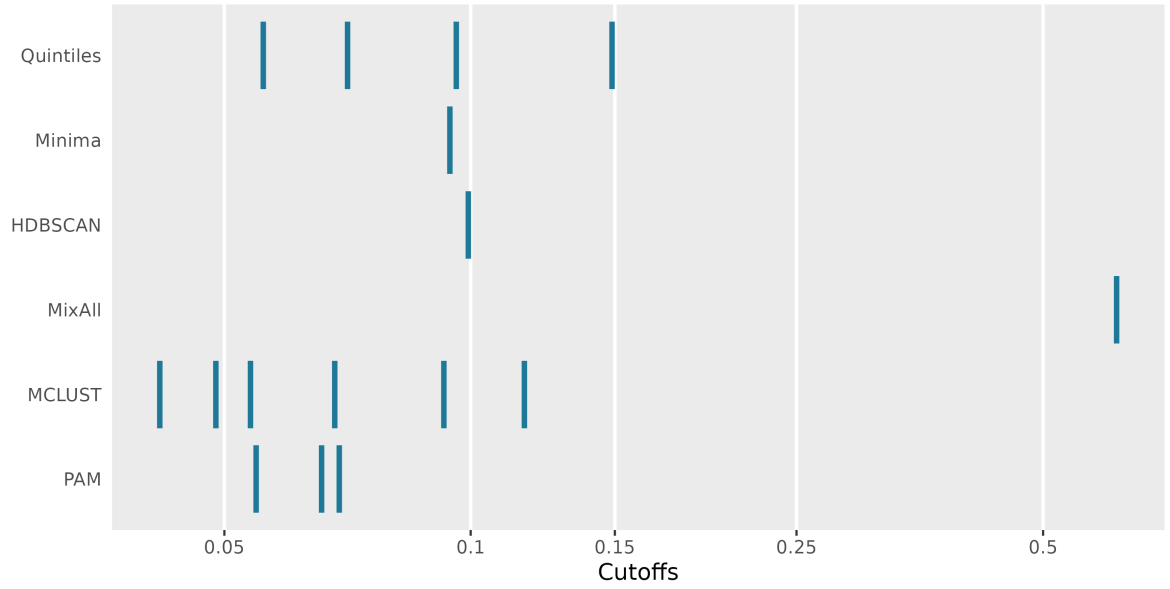

Figure 31: Cutoff values compared for the library amenity for all clustering approaches.

7.9 Parks

Examining Table 24 for the parks amenity, we observe that MixAll and PAM have similar cutoffs for the first cluster, ranging from 0 to 0.0447 and 0 to 0.0450, respectively. Additionally, both techniques find only 2 clusters, indicating a similar result for cluster 2 as well. MCLUST also identifies a cutoff near this range, but it consists of 3 clusters within the 0 to 0.0463 range. MCLUST further finds 5 other clusters for the remaining data. Apart from these findings, none of the other cutoffs match across all the approaches for the park's amenity. Figure 32 demonstrates that the number of DBs and DB population aligns with these cutoff combinations.

Analyzing the validation metrics in Table 23 for clustering techniques applied to parks, we find that the Silhouette coefficient and Davies Bouldin index suggest that PAM performs better in clustering this amenity, while the Dunn index favors MixAll and the Calinski Harabasz index suggests MCLUST.

Based on this analysis, we can conclude that although none of these methods agree exactly on the cutoffs, and the validation metrics also do not unanimously support one technique, the first cluster cutoff may be around 0.0183, as three techniques (Minima identification, HDBSCAN, and MCLUST) find cutoffs near this value. The second cluster cutoff may be around 0.0450, as MixAll, MCLUST, and PAM identify a cutoff point close to this value, MixAll and PAM supported by the validation metrics. The remaining data may belong to a single cluster, as suggested by MixAll and PAM.

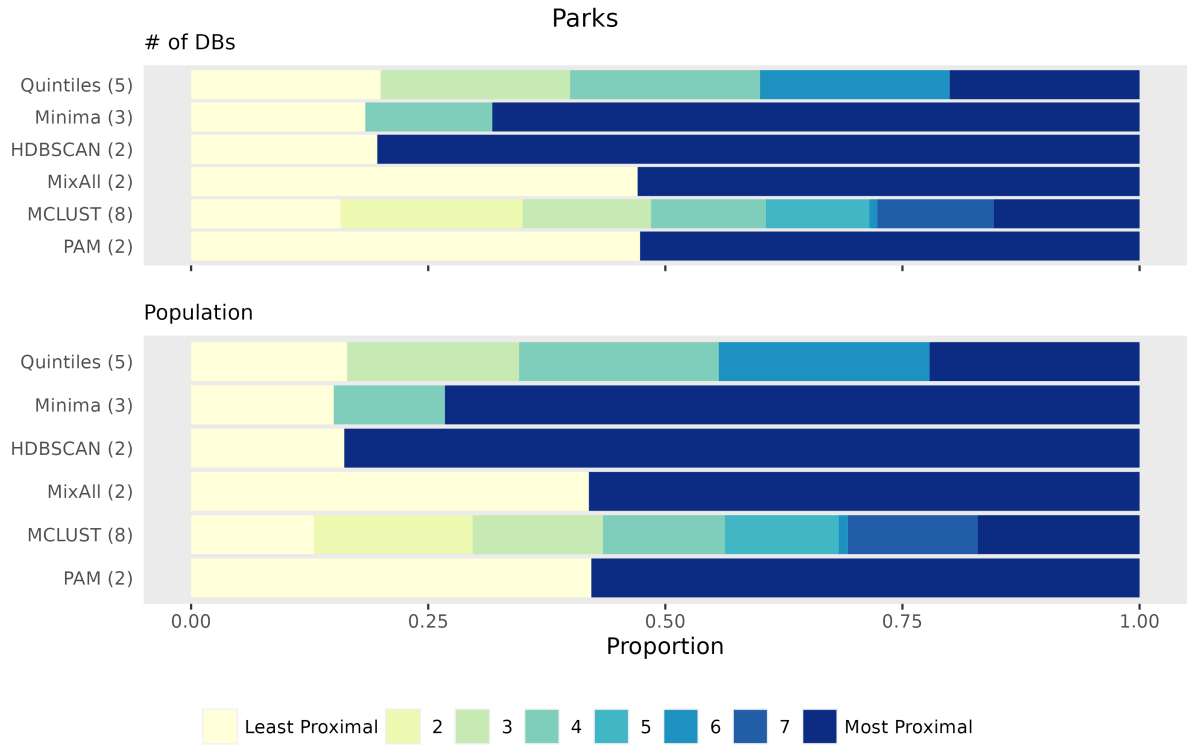

Figure 32: Proportion of DBs and population in each cluster for all approaches for the parks amenity.

Table 23: The validation metric values for each clustering approach for the parks amenity.

	Silhouette	Dunn	Calinski Harabasz	Davies Bouldin
Quintiles	0.48	0.00000	4676	0.74
MixAll	0.57	0.00052	14414	0.70
HDBSCAN	0.36	0.00000	4008	4.06
PAM	0.58	0.00044	14512	0.69
MCLUST	0.46	0.00011	17244	0.96
Minima identification	0.43	0.00013	1342	0.70

Table 24: Summary statistics for each cluster found by all approaches for the parks amenity. DB Population, IoR and proximity value show the median, while CMA Type, Province and Amenity Dense show the mode.

	# of DBs	DB Population	Median IoR	CMA Type	Province	Amenity Dense	Parks	Range
Entire Population	234,068 (100.0%)	62	0.11	CMA (68.4%)	Ontario (25.3%)	Low (82.3%)	0.048	0 - 1
Quintiles C1	46,782 (20.0%)	45	0.16	CMA (47.2%)	Ontario (16.1%)	Low (95.7%)	0.013	0 - 0.0203
Quintiles C2	46,761 (20.0%)	55	0.13	CMA (61.0%)	Ontario (22.4%)	Low (90.1%)	0.028	0.0203 - 0.0372
Quintiles C3	46,859 (20.0%)	65	0.11	CMA (70.1%)	Ontario (28.0%)	Low (83.7%)	0.048	0.0372 - 0.0614
Quintiles C4	46,808 (20.0%)	71	0.11	CMA (77.3%)	Ontario (30.3%)	Low (77.3%)	0.079	0.0614 - 0.1050
Quintiles C5	46,858 (20.0%)	72	0.11	CMA (86.1%)	Ontario (29.4%)	Low (65.1%)	0.149	0.1050 - 1
Minima identification C1	42,995 (18.4%)	45	0.16	CMA (47.1%)	Ontario (16.0%)	Low (95.7%)	0.012	0 - 0.0183
Minima identification C2	31,335 (13.4%)	52	0.14	CMA (58.4%)	Ontario (21.0%)	Low (91.7%)	0.024	0.0183 - 0.0294
Minima identification C3	159,738 (68.2%)	68	0.11	CMA (76.0%)	Ontario (28.6%)	Low (76.9%)	0.071	0.0294 - 1
HDBSCAN C1	45,949 (19.6%)	45	0.16	CMA (47.1%)	Ontario (16.1%)	Low (95.7%)	0.013	0 - 0.0200
HDBSCAN C2	188,119 (80.4%)	66	0.11	CMA (73.5%)	Ontario (27.5%)	Low (79.1%)	0.061	0.0200 - 1
MixAll C1	110,198 (47.1%)	52	0.14	CMA (56.1%)	Ontario (20.3%)	Low (91.8%)	0.023	0 - 0.0447
MixAll C2	123,870 (52.9%)	70	0.11	CMA (79.3%)	Ontario (29.6%)	Low (74.0%)	0.087	0.0447 - 1
MCLUST C1	36,926 (15.8%)	45	0.16	CMA (46.9%)	Ontario (15.9%)	Low (95.7%)	0.012	0 - 0.0159
MCLUST C2	44,867 (19.2%)	52	0.14	CMA (57.4%)	Ontario (20.7%)	Low (91.9%)	0.024	0.0159 - 0.0324
MCLUST C3	31,713 (13.5%)	62	0.11	CMA (66.4%)	Ontario (25.9%)	Low (86.1%)	0.039	0.0324 - 0.0463
MCLUST C4	28,343 (12.1%)	67	0.11	CMA (72.0%)	Ontario (28.9%)	Low (82.3%)	0.054	0.0463 - 0.0624
MCLUST C5	25,512 (10.9%)	71	0.11	CMA (76.6%)	Ontario (30.8%)	Low (78.6%)	0.072	0.0624 - 0.0825
MCLUST C6	1,974 (0.8%)	71	0.10	CMA (79.2%)	Ontario (31.0%)	Low (74.7%)	0.084	0.0825 - 0.0845
MCLUST C7	28,802 (12.3%)	72	0.11	CMA (79.3%)	Ontario (29.6%)	Low (74.2%)	0.100	0.0845 - 0.1219
MCLUST C8	35,931 (15.4%)	72	0.11	CMA (87.8%)	Ontario (29.3%)	Low (62.9%)	0.168	0.1219 - 1
PAM C1	110,802 (47.3%)	52	0.14	CMA (56.2%)	Ontario (20.4%)	Low (91.7%)	0.023	0 - 0.0450
PAM C2	123,266 (52.7%)	70	0.11	CMA (79.3%)	Ontario (29.6%)	Low (73.9%)	0.088	0.0450 - 1

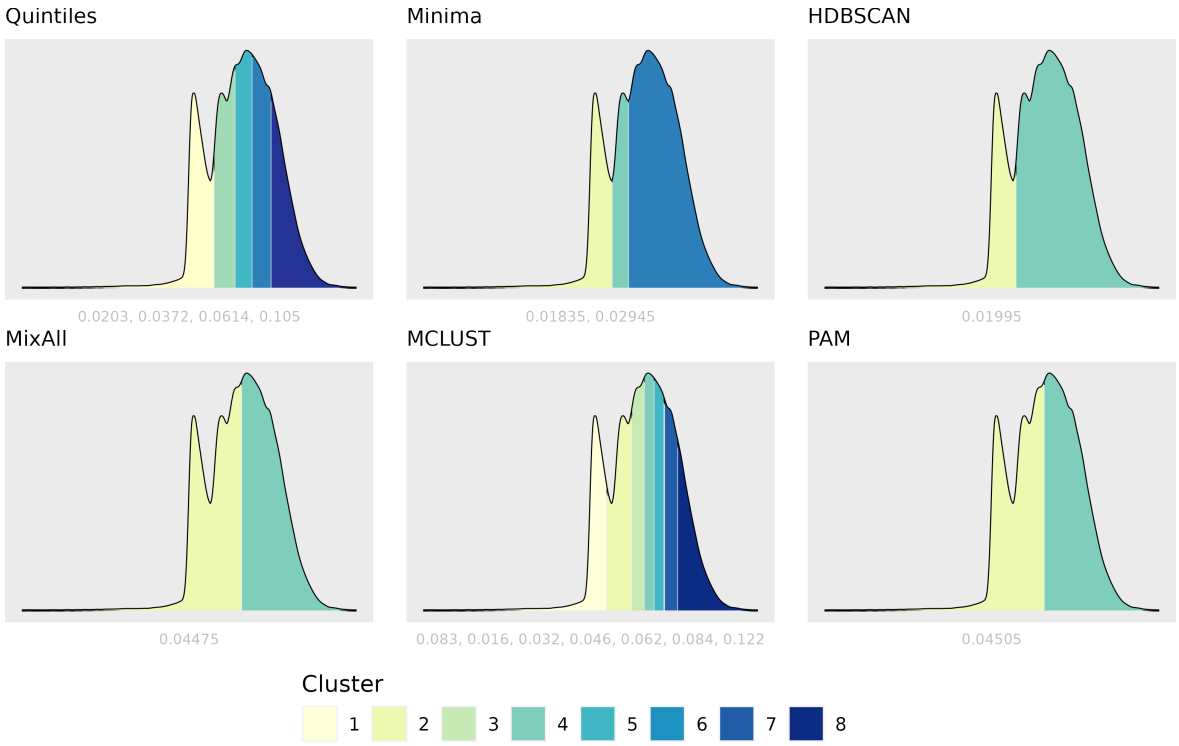

Figure 33: Cut-offs values shown on the log-transformed density plots for all clustering approaches for the parks amenity.

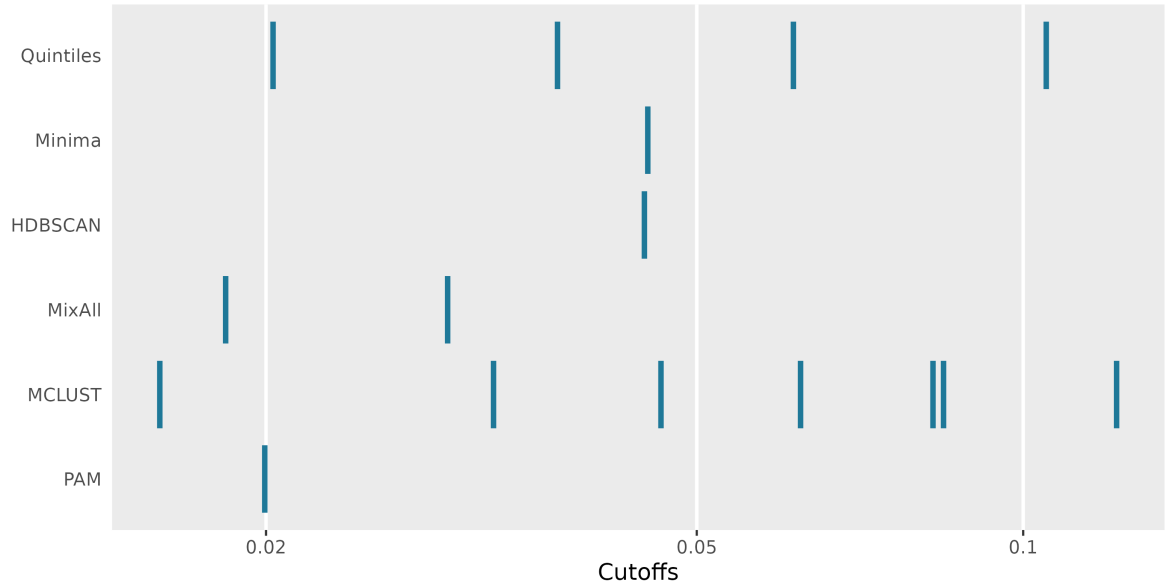

Figure 34: Cutoff values compared for the parks amenity for all clustering approaches.

7.10 Transit

Examining Table 26 for the transit amenity, we observe that although the number of clusters matches in HDBSCAN, MixAll, and PAM, none of the cutoffs align across all these clustering techniques. This trend is also reflected in the number of DBs and DB population, as shown in Figure 35. While the combination of the first two clusters from the quintile method matches the cutoff for the first cluster in MixAll, we should not consider it since the quintile method does not determine cutoffs based on the underlying data.

Analyzing the validation metrics in Table 25 for clustering techniques applied to transit, we find that MCLUST performs better in clustering the transit amenity, as suggested by the Calinski Harabasz and Davies Bouldin indices. However, the Dunn index favors MixAll, and the Silhouette index suggests Minima identification as the better performers.

Based on this analysis, it is evident that different clustering techniques yield different cutoffs, and the validation metrics also suggest different techniques without any common cutoffs. Therefore, it is not possible to emphasize any particular cutoffs for clustering in this case.

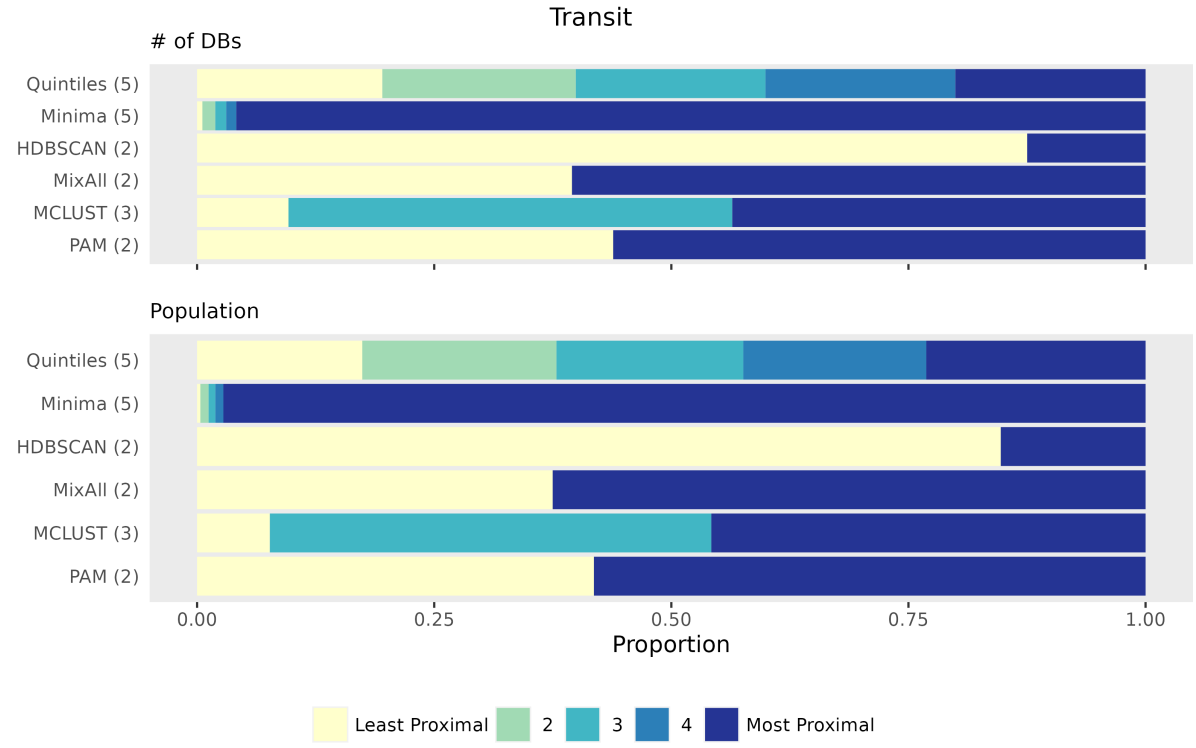

Figure 35: Proportion of DBs and population in each cluster for all approaches for the transit amenity.

Table 25: The validation metric values for each clustering approach for the transit amenity.

	Silhouette	Dunn	Calinski Harabasz	Davies Bouldin
Quintiles	0.42	0.00000	1519	1.00
MixAll	0.55	0.00355	9466	0.76
HDBSCAN	0.27	0.00000	958	2.46
PAM	0.54	0.00297	8940	0.78
MCLUST	0.58	0.00249	11502	0.64
Minima identification	0.74	0.00017	26	0.87

Table 26: Summary statistics for each cluster found by all approaches for the transit amenity. DB Population, IoR and proximity value show the median, while CMA Type, Province and Amenity Dense show the mode.

	# of DBs	DB Population	Median IoR	CMA Type	Province	Amenity Dense	Transit	Range
Entire Population	181,305 (100.0%)	73	0.10	CMA (89.8%)	Ontario (31.9%)	Low (76.9%)	0.009	0 - 1
Quintiles C1	35,411 (19.5%)	58	0.11	CMA (73.7%)	Ontario (31.6%)	Low (94.3%)	0.001	0 - 0.0026
Quintiles C2	36,983 (20.4%)	75	0.10	CMA (86.6%)	Ontario (33.3%)	Low (91.7%)	0.004	0.0026 - 0.0067
Quintiles C3	36,255 (20.0%)	74	0.11	CMA (92.0%)	Ontario (30.5%)	Low (85.8%)	0.009	0.0067 - 0.0131
Quintiles C4	36,300 (20.0%)	73	0.10	CMA (96.8%)	Ontario (30.2%)	Low (72.3%)	0.018	0.0131 - 0.0272
Quintiles C5	36,356 (20.1%)	85	0.06	CMA (99.4%)	Ontario (34.1%)	Med (46.7%)	0.044	0.0272 - 1
Minima identification C1	1,014 (0.6%)	37	0.15	CMA (48.3%)	Ontario (29.7%)	Low (95.6%)	0.000	0 - 0.0000
Minima identification C2	2,474 (1.4%)	38	0.13	CMA (56.8%)	Ontario (26.6%)	Low (93.9%)	0.000	0.0000 - 1e-04
Minima identification C3	2,082 (1.1%)	36	0.13	CMA (59.5%)	Ontario (27.2%)	Low (92.5%)	0.000	1e-04 - 3e-04
Minima identification C4	1,923 (1.1%)	43	0.13	CMA (66.8%)	Ontario (30.8%)	Low (95.3%)	0.000	3e-04 - 4e-04
Minima identification C5	173,812 (95.9%)	74	0.10	CMA (91.1%)	Ontario (32.1%)	Low (76.2%)	0.010	4e-04 - 1
HDBSCAN C1	158,674 (87.5%)	71	0.10	CMA (88.4%)	Ontario (31.4%)	Low (83.2%)	0.008	0 - 0.0388
HDBSCAN C2	22,631 (12.5%)	90	0.06	CMA (99.8%)	Ontario (35.8%)	Med (49.3%)	0.056	0.0388 - 1
MixAll C1	71,659 (39.5%)	67	0.11	CMA (80.2%)	Ontario (32.5%)	Low (93.0%)	0.003	0 - 0.0065
MixAll C2	109,646 (60.5%)	76	0.09	CMA (96.1%)	Ontario (31.6%)	Low (66.4%)	0.018	0.0065 - 1
MCLUST C1	17,469 (9.6%)	47	0.13	CMA (66.9%)	Ontario (30.6%)	Low (94.8%)	0.000	0 - 0.0010
MCLUST C2	84,840 (46.8%)	73	0.10	CMA (87.0%)	Ontario (32.1%)	Low (90.5%)	0.005	0.0010 - 0.0116
MCLUST C3	78,996 (43.6%)	78	0.07	CMA (97.8%)	Ontario (32.0%)	Low (58.5%)	0.025	0.0116 - 1
PAM C1	79,536 (43.9%)	68	0.11	CMA (81.2%)	Ontario (32.3%)	Low (92.7%)	0.003	0 - 0.0076
PAM C2	101,769 (56.1%)	76	0.09	CMA (96.5%)	Ontario (31.7%)	Low (64.6%)	0.020	0.0076 - 1

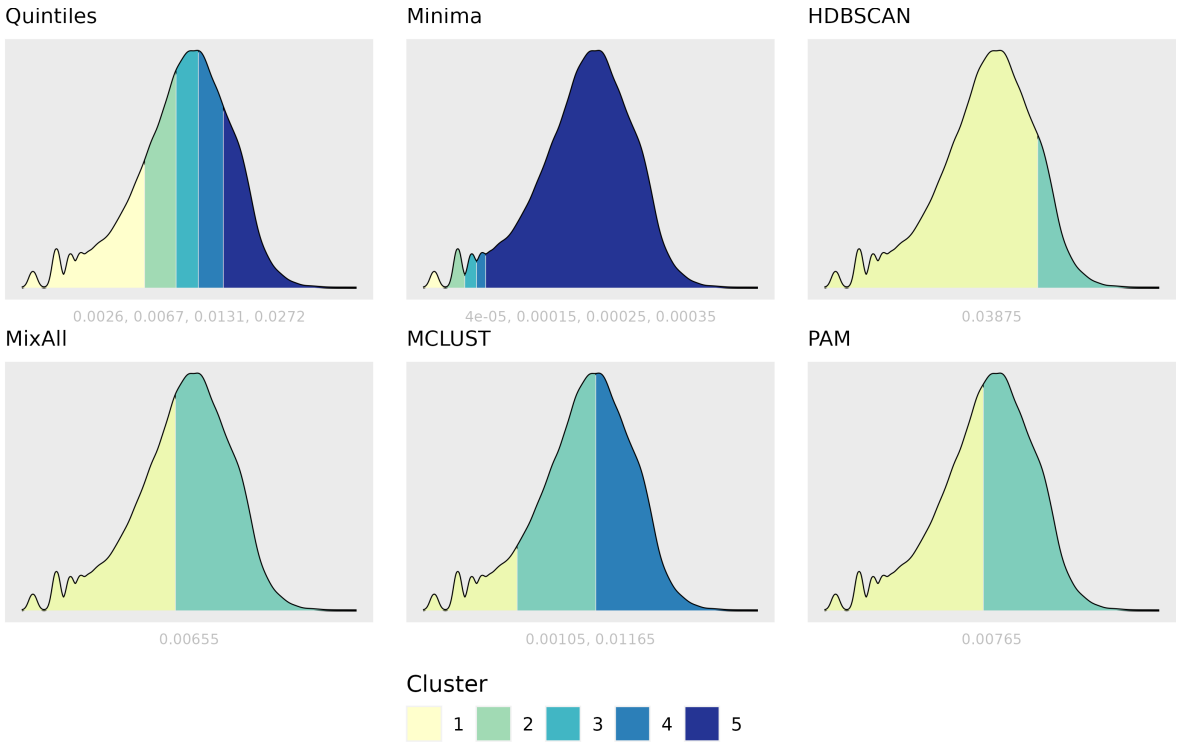

Figure 36: Cut-offs values shown on the log-transformed density plots for all clustering approaches for the transit amenity.

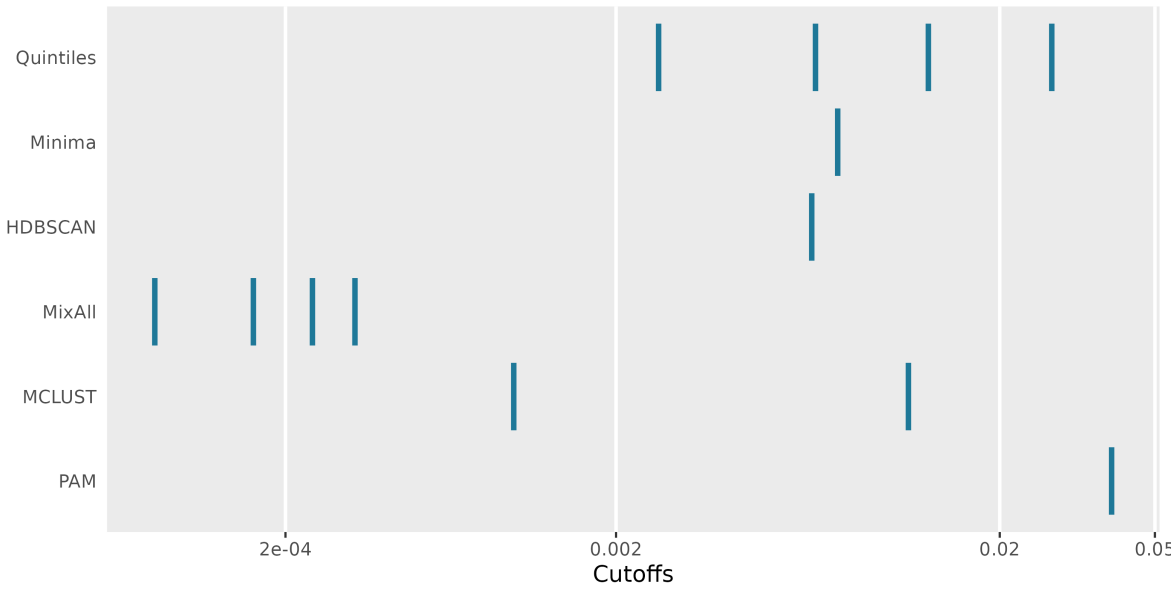

Figure 37: Cutoff values compared for the transit amenity for all clustering approaches.

8 Discussion

The most significant takeaway from the current investigation is the lack of clear-cut segments in the early release of the PMD. While it is true that log-transforming the proximity measures did reveal certain density-sparse regions, the clustering algorithms utilized were not able to 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 in Canada is a relatively smooth gradient without any obvious clusters.

Not only were results inconsistent between approaches for the same amenity, but results between amenities using the same algorithm were not always comparable. For example, the PAM algorithm consistently found a very low number of clusters, except for the grocery amenity, for which it found many more. The MCLUST algorithm demonstrates a similar inconsistency: for the transit amenity it only finds three clusters, but for the employment amenity it finds nine. There are two explanations for this behaviour. The first is that these inconsistencies are due to the same underlying problem that causes inconsistencies between algorithms for the same amenity: namely that the data is not particularly clusterable. Therefore, the algorithms are dividing the measures somewhat arbitrarily at chance fluctuations. The second reason is that the 'true' number of clusters likely differs significantly between amenities. Therefore, the number and type of clusters for one amenity would not be expected to appear similar to another amenity. In other words, this explanation suggests that just because a DB can be easily classified as either low or high access to amenity A, this does not mean that this same DB can be easily classified into one of only two categories for amenity B. Instead, one amenity may be distributed very simply, while others may be distributed in a more complex manner.

The current investigation is an informative first glance into the clusterability of the early release of the PMD. We were able to attempt many different clustering algorithms, which reinforces our finding that PMD segments are not robust and reproducible, but are instead sensitive to a variety of factors. This broad scope of approaches will help to guide and refine the endeavors of future researchers. While we did take extensive care to ensure the validity and reproducibility of our results, we were constrained in some aspects of our methodology. The most major concern is that of computational constraints. Due to the complexity of several of the algorithms we implemented, subsampling was required in order to avoid running out memory. 3% subsampling was most commonly used. In the future, researchers not subjected to similar computational restraints should seek to run their algorithms on the entire dataset, rather than a subsample. Additionally, it is worth noting that since the PMD was only recently released as "experimental statistics", it is possible that better, more comprehensive ways of calculating the proximity index using additional/different data sources may be developed in the future, which may render our methodology obsolete.

There are many other potential avenues for future research in the clustering of the PMD. These include: guiding the number of clusters to be the same between all algorithms, trying different combinations of variables in multidimensional clustering (as opposed to clustering only on the proximity measure in question), trying additional clustering algorithms, clustering on different data transformations, attempting sub-clustering (ie. some way of dealing with outliers aside from selective models), as well as soft assignments (i.e., using overlapping ranges where the cut-off is not a single point but, ideally, a narrow continuous interval).

9 Conclusion

The current project aimed to explore segmentation of continuous proximity measures in the Proximity Measure Database (PMD) developed by Statistics Canada. The goal was to create intuitive and understandable categorical measures for amenities, which could inform decision making processes for policymakers and urban planners. By categorizing the proximity measures, it becomes easier to prioritize efforts in enhancing access and promoting social and economic sustainability within communities.

The project employed various clustering methods, including minima identification, HDBSCAN, MixAll, MCLUST, and PAM algorithms to determine optimal cutoff values and cluster boundaries for each amenity. Additionally, cluster validation metrics such as the Silhouette coefficient, Dunn index, Calinski-Harabasz, and Davies-Bouldin were used to evaluate the performance of each clustering technique and determine the appropriate number of clusters.

The results showed that the PMD had a low clustering tendency, even after log-transformation. Clustering techniques produced diverse outcomes, and there was no single algorithm that consistently outperformed others. The overall lack of consistency serves to demonstrate the lack of obvious clusters within the proximity measures of the PMD.

Although there was some overlap between algorithms, cluster profiling revealed that the clusters identified by different algorithms were mostly distinct. One common trend that held true for the majority of clusters was that as amenity proximity increases, median IoR decreases, population increases, percentage of CMA DBs increases, and the percentage of low amenity dense DBs decreases. The Pharmacy amenity was the only one that did not follow this trend: all of the clusters, as proximity measures increased, had near constant summary statistics.

Overall, this project explored a variety of clustering methods in segmenting continuous proximity measures and generating meaningful categorical measures. In light of our original research goal to find the optimal cut-off values and cluster boundaries, these findings were not conclusive. The overall lack of clustering consistency and distinct characteristics of clusters identified by different algorithms within this limited data underscore the need for further refinement and exploration. We theorize that an "exhaustive" PMD which is not limited by travel radii will have more identifiable groups, and so the methods outlined in this report will be useful to determine the clusters. Future research can build upon these findings to refine the clustering techniques and explore additional factors that contribute to the definition of clusters in the PMD.

10 References

1 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.

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.* 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.

3 Bernard, D. (2018). Clustering Indices. https://cran.r-hub.io/web/packages/clusterCrit/clusterCrit.pdf 1.2.7.

4 Caliński, T., & Harabasz, J. (1974). *A Dendrite Method for Cluster Analysis.* Communications in Statistics-theory and Methods 3: 1-27. doi:10.1080/03610927408827101.

5 de Smith, M. J., Goodchild, M. F., Longley, P. A., & Colleagues. (2021). *Geospatial Analysis* 6th Edition, 2021 update. Retrieved June 12, 2023, from https://www.spatialanalysisonline.com/HTML/index.html.

6 Davies, D.L. & Bouldin, D.W. (1979). *A Cluster Separation Measure.* IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-1 (2): 224–227. doi:10.1109/TPAMI.1979.4766909.

7 Dunn, J. C. (1974). "Well-Separated Clusters and Optimal Fuzzy Partitions". Journal of Cybernetics. 4 (1): 95–104. doi:10.1080/01969727408546059.

8 Hashmi, F. (2021, November 27). *Data Science Interview Questions for IT Industry Part-4: Unsupervised ML - Thinking Neuron.* Thinking Neuron. https://thinkingneuron.com/data-science-interview-questions-for-it-industry-part-4-unsupervised-ml/#DBSCAN.

9 Hahsler, M., Piekenbrock, M., & Doran, D. (2019). *dbscan: Fast Density-Based Clustering with R.* Journal of Statistical Software, 91(1), https://doi.org/10.18637/jss.v091.i01.

10 Iovleff, S. (2019, September 12). *MixAll: Clustering Mixed data with Missing Values.* Retrieved June 12, 2023, from https://cran.r-project.org/web/packages/MixAll/vignettes/Introduction-Mixtures.pdf.

11 *Jenks Natural Breaks Classification* - GIS Wiki — The GIS Encyclopedia (2018). Retrieved June 12, 2023, from http://wiki.gis.com/wiki/index.php/Jenks_Natural_Breaks_Classification.

12 Kassambara, A. (2017). *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning.* STHDA.

13 Kassambara, A. (2018). *K-Medoids in R: Algorithm and Practical Examples.* Retrieved June 12, 2023, from https://www.datanovia.com/en/lessons/k-medoids-in-r-algorithm-and-practical-examples/.

14 Kassambara A, Mundt F (2020). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.7, https://CRAN.R-project.org/package=factoextra.

15 Kenton, W. (2023). *Kurtosis Definition, Types, and Importance.* Investopedia. https://www.investopedia.com/terms/k/kurtosis.asp.

16 Marbac, M. M., & Sedki, M. S. (2017). *Variable Selection for Model-Based Clustering of Continuous, Count, Categorical or Mixed-Type Data Set with Missing Values* [Software]. In CRAN (2.0.1). http://cran.nexr.com/web/packages/VarSelLCM/.

17 OECD, Statistics Canada. (2018). *Workshop on Modernising Statistical Systems for Better Data on Regions and Cities.* Retrieved May 4, 2023, from https://www.oecd.org/cfe/regionaldevelopment/modernising-statistical-systems.htm.

18 Rousseeuw, P.J. (1987) *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis.* Journal of Computational and Applied Mathematics, Volume 20, Pages 53-65,ISSN 0377-0427. https://doi.org/10.1016/0377-0427(87)90125-7.

19 Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). *MCLUST 5: clustering, classification, and density estimation using Gaussian finite mixture models.* The R Journal, 8(1), 289-317. https://doi.org/10.32614/RJ-2016-021.

20 Statistics Canada (2020a). *Proximity Measures Data Viewer.* https://www150.statcan.gc.ca/n1/pub/71-607-x/71-607-x2020011-eng.htm.

21 Statistics Canada (2020b). *Proximity Measures Database – Early release.* https://www150.statcan.gc.ca/n1/pub/17-26-0002/172600022020001-eng.htm.

22 Statistics Canada. (2021). *Dictionary, Census of Population, 2021 Dissemination block (DB).* https://www12.statcan.gc.ca/census-recensement/2021/ref/dict/az/definition-eng.cfm.

23 Subedi,R., Roshanafshar, S., & Greenberg, T.L. (2020). *Developing Meaningful Categories for Distinguishing Levels of Remoteness in Canada.* Analytical Studies: Methods and References, Statistics Canada. https://www150.statcan.gc.ca/n1/pub/11-633-x/11-633-x2020002-eng.htm.

24 Wickham, H., Averick, M., Bryan, J., Chang, W. W., McGowan, L. D., François, R., Grolemund, G., Hayes, A. G., Henry, L., Hester, J., Kuhn, M., Pedersen, T. G., Miller, E. W., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the Tidyverse. Journal of Open Source Software, 4(43), 1686. https://doi.org/10.21105/joss.01686.

A Appendix

A.1 Successful Methods

A.1.1 PAM (Partitioning Around Medoids - PAM)

K-means is a clustering algorithm that aims to partition a dataset into K clusters, where each data point belongs to the group with the closest mean. The Partitioning Around Medoids (PAM) variation replaces the concept of mean with medoids to handle noise and outliers more effectively (Kaufman and Rousseeuw 1990).

The PAM algorithm, an evolution of the K-means clustering method, operates by selecting K representative objects, or medoids, among the observations of the dataset. These medoids are the most centrally located data points in a cluster, which means the average dissimilarity between a medoid and all other objects within the same cluster is minimized. In contrast to the K-means algorithm, which uses means as cluster centers, PAM's utilization of medoids makes it a robust alternative, less sensitive to noise and outliers.

The PAM algorithm works in two phases: the 'build phase' and the 'swap phase'. During the build phase, K objects are selected to be the medoids, the dissimilarity matrix is calculated, and every object is assigned to its closest medoid. The swap phase attempts to improve the clustering quality by exchanging selected objects (medoids) and non-selected objects. If the sum of the dissimilarities of all objects to their nearest medoid (the objective function) can be reduced by this swapping, then the swap is carried out. The process continues until the objective function can no longer be decreased, resulting in a set of K representative objects which minimize the sum of the dissimilarities of the observations to their nearest representative object (Kaufman and Rousseeuw 1990).

Since PAM is an extension of the K-means algorithm, it operates under similar assumptions but with additional robustness due to the use of medoids. The following are the fundamental assumptions it makes about the data:

- Globular clusters: It assumes that the natural grouping of data forms globular or spherical clusters. This assumption helps separate clusters effectively when the algorithm operates on the data (Perceptive Analytics, 2017).
- Clusters of similar size: The algorithm works under the assumption that clusters contain approximately the same number of data points. It determines the boundaries of the cluster based on this assumption (Perceptive Analytics, 2017).
- Distance Measures: The algorithm employs distance measures, such as Euclidean distance, to compute similarity. This assumes that 'straight-line' distance is the appropriate measure of similarity, which may not hold true in all contexts (Perceptive Analytics, 2017)

We selected the PAM algorithm for its robustness to outliers, an attribute particularly beneficial given the nature of our dataset that comprises 10 distinct amenities. Emphasizing a univariate approach, we individually clustered each amenity to preserve their unique characteristics and to align with the algorithm's cluster homogeneity assumption. Although our proximity data was already normalized between 0 and 1, we further refined the distribution of each amenity through log transformation. This transformation was not a direct requirement of the algorithm, but it was implemented to create more normally distributed data and reduce outliers.

A.1.2 Gaussian Mixture Models (MCLUST)

MCLUST is an R package that provides a comprehensive approach to finite mixture models, providing functions for model-based clustering, classification, and density estimation based on

Gaussian Mixture Models (GMMs). GMMs are probabilistic models assuming that the data points in a given dataset are generated from a mixture of Gaussian distributions, with each Gaussian component representing a distinct cluster (Scrucca et al., 2016).

MCLUST uses the Expectation-Maximization (EM) algorithm for estimating the parameters. The EM algorithm operates iteratively in two steps:

1. Expectation (E) Step: Expected values of the component memberships are calculated based on the current parameters.

2. Maximization (M) Step: The log-likelihood function is maximized to update the parameter estimates based on these expected values.

The process is repeated until convergence, providing the parameter estimates for the mixture model. Furthermore, MCLUST automatically computes and selects the best model as per the Bayesian Information Criterion (BIC), considering different numbers of clusters and different parameterizations of the covariance matrix (Scrucca et al., 2016).

The Gaussian Mixture Models (MCLUST) operate on several assumptions. Firstly, MCLUST is formulated on the belief that the data is generated from a mixture of Gaussian distributions. This assumption provides a statistical framework that guides how the algorithm processes the dataset (Scrucca et al., 2016).

Another key presumption is that the variables within each component are normally distributed and independent. This is a common assumption in many statistical techniques and it impacts how the algorithm assesses relationships within the data.

Lastly, MCLUST relies on maximum likelihood estimation, which is influenced by initial values and can potentially converge to local, instead of global, maxima. This characteristic points to the algorithm's optimization strategy and its approach to finding the most probable parameters for the Gaussian distributions (Scrucca et al., 2016).

Gaussian Mixture Models (MCLUST) proved to be an effective choice for our project due to its robustness and flexibility in handling model-based clustering. With our dataset's characteristics - univariate proximity measures for ten different amenities, the capability of MCLUST to handle different Gaussian components was an instrumental feature for identifying unique clusters.

Prior to implementing MCLUST, we took strategic steps in our data preprocessing, such as applying appropriate transformations when necessary. While it's important to clarify that these transformations weren't specifically carried out for MCLUST, they naturally helped our data to align more closely with the Gaussian distribution, an assumption inherent in the model. This approach enhanced the performance potential of the algorithm, making it a more suitable fit for our data.

Another particularly appealing feature of MCLUST was its ability to autonomously compute the optimal model, taking into account varying numbers of clusters and different configurations of the covariance matrix. This made the algorithm more robust and efficient in managing the complexities inherent in our dataset, consequently improving the overall quality of the clustering results.

A.1.3 MixAll

MixAll is a clustering model that functions on the premise of mixture models. These models assume that data is generated from a combination of probability distributions, which is ideal for handling datasets with diverse distributions or missing values.

The MixAll model is basically a mixture model. Mixture models assume data is generated from a combination of probability distributions. Parameter estimation is achieved by maximizing the observed log-likelihood or integrated log-likelihood for data with missing values.

Estimation algorithms like expectation-maximization (EM), SEM, and CEM are used and the default is EM which is highlighted below, involving steps such as imputation, conditional probability calculation, and parameter updates. The EM algorithm iteratively performs these steps until convergence (Iovleff, 2019).

The EM algorithm consists of several iterative steps:

1. I step: Impute the missing values x_i^m using the current MAP value provided by the current parameter θ^{m-1}.
2. E step: Compute the current conditional probabilities t_{ik}^m for $i = 1, \ldots, n$ and $k = 1, \ldots, K$ using the current parameter θ^{m-1}.
3. M step: Update the maximum likelihood estimate θ^m of θ using the conditional probabilities t_{ik}^m as conditional mixing weights, aiming to maximize the log-likelihood function, where $t^m = (t_{ik}^m, i = 1, \ldots, n, k = 1, \ldots, K)$.
4. Parameter update: The updated expression of mixture proportions p_k^m for $k = 1, \ldots, K$ are computed. Detailed formulas for updating the parameters λ_k and α depend on the component parameterization (Iovleff, 2019).

It's important to note that the notation and steps described above are derived from the article "MixAll: Clustering Mixed data with Missing Values" by Serge Iovleff.

The MixAll algorithm operates based on several assumptions that could influence the implementation and results of the model. One such assumption pertains to the use of the `clusterDiagGaussian` function. This function is designed to work with multivariate data, treating each variable as independent during the clustering process. Given that our data is univariate, this aspect of the algorithm may offer results unique to our dataset.

Another assumption built into MixAll is that the data arises from a Gaussian mixture. This suggests that the model expects the underlying distribution of the data to resemble a blend of Gaussian distributions. Lastly, a third assumption pertains to the standard deviations within each component of the model, which the algorithm anticipates to be varied (Iovleff, 2019).

Initially, we chose to employ the MixAll algorithm for its capacity to effectively deal with missing data. However, this motivation was short lived as we shifted our attention to clustering in the univariate case. Furthermore, similar to MCLUST, MixAll's adeptness in handling a diverse range of distributions added to its appeal for our project.

The algorithm's assumption of data arising from a Gaussian mixture aligns well with the preprocessing measures we adopted for our data. Specifically, as previously stated, we applied a log transformation to our initially right-skewed data to achieve a distribution that more closely approximates a Gaussian one. This transformation assisted in leveraging MixAll's inherent strength in managing datasets with varied distributions, thereby enhancing the effectiveness of the clustering analysis.

A.1.4 Hierarchical Denisity-Based Spatial Clustering of Applications with Noise (HDBSCAN)

Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) is a flexible clustering algorithm that extends DBSCAN by converting it into a hierarchical clustering algorithm. The density-based algorithm can find clusters of varying densities and is designed to be more flexible than some of the other more prominent clustering techniques. This feature allows it to recognize and work with clusters of varying densities, adding to its versatility and applicability across diverse datasets (McInnes, Healy, & Astels, 2016).

HDBSCAN works on the concept of density-based clustering (DBSCAN) but goes a step further by introducing hierarchy, allowing it to discover clusters of varying densities. This algorithm operates in two main steps:

1. Transform the space according to the density/sparsity. This transformation ensures that sparse areas are more distant. It utilizes the core distance (defined by parameter MinPts) and mutual reachability distance to create an undirected weighted graph, and then applies the single-linkage clustering to the graph (Campello et al., 2015).

2. Create a hierarchy of clusters. The hierarchy produced by single-linkage clustering is then simplified by transforming it into a tree, which is then condensed by pruning branches not representing a cluster. The pruning process is guided by the stability of clusters, which is computed based on their persistence over the distance (Campello et al., 2015).

HDBSCAN, like other density-based clustering algorithms, assumes that clusters are dense regions in the data space, separated by regions of lower density. It does not require the clusters to be of a particular geometric shape, making it versatile for different datasets. However, it does expect the density within clusters to be relatively uniform, and it may struggle with clusters of widely varying densities. It also assumes that noise is present in the data, which it will not include in clusters, instead treating it as 'background noise' (Campello et al., 2015).

HDBSCAN was chosen because of its ability to detect clusters of varying densities, offering flexibility that aligned with the nature of our data. Additionally, HDBSCAN's assumption of density-based clusters proved suitable for our project, particularly because the proximity measures of the amenities in our dataset naturally lent themselves to such a density-based analysis, as our goal was to detect density sparse regions. Lastly, and probably the most enticing reason was the algorithm's tendency to handle noise. The algorithm helped ensure a robust clustering output, accommodating for potential outliers that were present in the data.

A.1.5 Multivariate - ClustImpute

ClustImpute algorithm on multi-dimensional, log-scaled proximity measures. Other variables used for this clustering along with the one amenity at a time include:

- "CSD_AREA"
- "PMS_CSDPOP"
- "PMS_DBPOP"
- "IOR_Index_of_remoteness"

These variables were scaled from 0-1 prior to clustering. This algorithm "draws the missing values iteratively based on the current cluster assignment so that correlations are considered on this level". Also, "penalizing weights are imposed on imputed values and successively decreased (to zero) as the missing data imputation gets better". The idea is that the missing value is imputed by those other observations that are more similar to it (ie. in the same cluster).

Algorithm Steps:

1. It replaces all NAs by random imputation, i.e., for each variable with missings, it draws from the marginal distribution of this variable not taking into account any correlations with other variables

2. Weights < 1 are used to adjust the scale of an observation that was generated in step 1. The weights are calculated by a (linear) weight function that starts near zero and converges to 1 at n_end.

3. A k-medioids clustering is performed with a number of c_steps steps starting with a random initialization.

4. The values from step 2 are replaced by new draws conditionally on the assigned cluster from step 3.

5. Steps 2-4 are repeated nr_iter times in total. The k-medioids clustering in step 3 uses the previous cluster centroids for initialization.

6. After the last draws a final k-medioids clustering is performed.

A.1.6 Multivariate - VarSelLCM

The varselLCM (Variable Selection in Latent Class Models) clustering algorithm is a method that combines latent class modeling with variable selection techniques to identify meaningful clusters in data (Marbac & Sedki, 2017). This method has been applied on all the amenity proximity measures together.

Due to the significant presence of NA values in the dataset, it is necessary to utilize an algorithm that can cluster the data without the need for imputing these NA values. Imputing the NA values in this case could have a substantial impact on the resulting clusters.

Moreover, it is not feasible to simply remove the NA values from all columns in the dataset. This approach would lead to a significant reduction in the amount of available data. Additionally, the presence of missing values in one column can affect the available values in other columns, making it impractical to remove NA values indiscriminately from the dataset.

1. Data Preparation: The algorithm takes as input a dataset consisting of categorical variables. It is assumed that the data is generated from an underlying latent class structure, where each observation belongs to a specific latent class.

2. Model Initialization: The algorithm begins by randomly assigning observations to different latent classes. It initializes the model parameters, including the class probabilities and the conditional probabilities of each variable within each class.

3. Expectation-Maximization (EM) Algorithm: The varselLCM algorithm employs an iterative process based on the EM algorithm. In the expectation step (E-step), the algorithm calculates the probability of each observation belonging to each class based on the current model parameters.

4. Variable Selection: In the maximization step (M-step), the algorithm selects a subset of relevant variables that contribute to the clustering process. It employs a variable selection criterion, such as the Bayesian Information Criterion (BIC), to identify the most informative variables for clustering.

5. Model Update: Once the relevant variables are selected, the algorithm updates the model parameters based on the observed data and the selected variables. It estimates the class probabilities and the conditional probabilities of the selected variables within each class.

6. Iterative Process: Steps 3-5 are repeated iteratively until convergence is achieved. The algorithm continues updating the model parameters and selecting variables until the clustering solution stabilizes.

7. Final Clustering Solution: Once convergence is reached, the algorithm assigns each observation to the latent class with the highest probability. The resulting clustering solution represents a partitioning of the data into distinct clusters based on the selected variables and their associated probabilities within each class (Marbac & Sedki, 2017).

Initially, VarselLCM was utilized for multivariate clustering. However, upon observing distinct cluster patterns in the data through log transformation, the focus shifted towards univariate clustering. Unfortunately, attempts to apply VarselLCM for univariate clustering were unsuccessful as it did not converge. Consequently, it was not possible to proceed with the technique.

A.2 Unsuccessful Methods

A.2.1 OPTICS

OPTICS stands for Ordering Points To Identify Clustering Structure. This algorithm can be seen as a generalization of DBSCAN. A major issue with DBSCAN is that it fails to find clusters of varying density due to fixed ϵ. This is solved in OPTICS by using an approach of finding reachability of each point from the core points and then deciding the clusters based on reachability plot (Hashmi, 2021).

Considering the log-transformed data, we observed multiple peaks and troughs, suggesting that the clusters may have varying densities. Therefore, aim to explore the applicability of OPTICS, a clustering technique adept at accommodating varying densities (Hahsler et al., 2019). Also there were a decent amount of outliers in the proximity measures which OPTICS can handle (2.3. Clustering, n.d.).

Relevant terminologies for OPTICS:

- ϵ, epsilon (eps): is the Maximum distance between to points that can be considered to form a group/cluster.
- MinPts: is the minimum number of points that must be present near each other within the epsilon (ϵ) range in order for them to all form a group or cluster.
- Core Point: A point in the data that has at least MinPts number of points nearby within the eps (ϵ) range.
- Border Point/Non-Core Point: A border point or non-core point is a data point in which there are fewer than the minimum number of points (MinPts) within reach of it (at a distance of eps).
- Noise: A noise point is a data point in which there isn't a single point within eps of it.
- Core Distance: Core distance can be less than the predetermined value of, epsilon (ϵ), which is the maximum allowed distance to find MinPts. Core distance denotes the minimum distance needed for a point to become a core point and denotes that the MinPts number of points can be found within this distance.
- Reachability distance: Reachability Distance is the minimum distance from the cluster's extreme point if the point is outside the core distance, and the core distance is the distance necessary to reach the point from the cluster if it is inside the core distance (Hashmi, 2021).

Algorithm Steps:

1. For the given values of MinPts and eps (ϵ), find out if a point is close to MinPts number of points within a distance less than or equal to eps. Tag it as a Core Point. Update the reachability distance = core distance for all the points within the cluster.
2. If it is not a core point then find out its density connected distance from the nearest cluster. Update the reachability distance.
3. Arrange the data in increasing order of reachability distance for each cluster. The smallest distances come first and represent the dense sections of data and the largest distances come next representing the noise section. This is a special type of dendrogram.
4. Find out the places where a sharp decline is happening in the reachability distance plot.
5. "Cut" the plot in the y-axis by a suitable distance to get the clusters (Hashmi, 2021).

The clustering process was applied solely to the employment variable without considering any supplementary explanatory variables. The resulting clusters overlap and intersect with other clusters. This overlapping and intersecting nature is not suitable for creating distinct profiles. For this reason, the decision was made not to continue with this technique.

A.2.2 Jenks Natural Break Classification

The Jenks Natural Breaks Classification (or Optimization) system is a data classification method designed to optimize the arrangement of a set of values into "natural" classes. A Natural class is the most optimal class range found "naturally" in a data set. Natural breaks are determined with a frequency histogram. Class boundaries are identified as troughs in the data. Many dataset will not have obvious natural breaks which means that this method would tend to show breaks where none really exists (Jenks Natural Breaks Classification - GIS Wiki - the GIS Encyclopedia, 2018.)

By attempting to minimize the average deviation of each class from the class mean while maximizing the average deviation of each class from the means of the other classes, the Jenks Natural Breaks Classification method attempts to reduce the variance within classes while enhancing the variance between classes (Wikipedia contributors, 2023).

Jenks Natural Breaks is chosen for application due to the limitation of proximity measures in representing data distribution, specifically when the distribution is not normal. Jenks Natural Breaks, being a non-parametric method, does not assume any specific data distribution and can be applied to a wide range of data types and distributions. This makes it a suitable choice in cases where the proximity measures' distribution deviates from normality. By considering the inherent characteristics of the data, Jenks Natural Breaks can identify natural groupings based on the actual data distribution, enhancing the clustering results. (Geospatial Analysis 6th Edition, 2021 Update - De Smith, Goodchild, Longley and Colleagues, 2021)

Algorithm Steps:

1. The user selects the attribute, x, to be classified and specifies the number of classes required, k.

2. A set of k-1 random or uniform values are generated in the range [min{x},max{x}]. These are used as initial class boundaries.

3. The mean values for each initial class are computed and the sum of squared deviations of class members from the mean values is computed. The total sum of squared deviations (TSSD) is recorded

4. Individual values in each class are then systematically assigned to adjacent classes by adjusting the class boundaries to see if the TSSD can be reduced. This is an iterative process, which ends when improvement in TSSD falls below a threshold level, i.e. when the within class variance is as small as possible and between class variance is as large as possible. True optimization is not assured. The entire process can be optionally repeated from Step 1 or 2 and TSSD values compared (Geospatial Analysis 6th Edition, 2021 Update - De Smith, Goodchild, Longley and Colleagues, 2021).

The results of the Jenks Natural Break classification are not useful for several reasons. Firstly, when considering employment and childcare, there were variations identified. However, for other amenities, the algorithm consistently suggested 2 or 3 clusters. The problem arises when we observe that the natural breaks for these clusters are within a very narrow range. For example, the first cluster has a range from 0 to 0.0095, and the second cluster has a range from 0.0095 to 0.7452. The remaining data points above this range are grouped into the third cluster. When plotting these clusters on a kernel density plot, we observe that only one cluster is visible. This is because the ranges for the other two clusters are so small that they cannot be effectively

visualized. This lack of visibility hinders the usefulness of the classification results. Moreover, these findings are not helpful for profiling purposes as they ignore the variations in the larger range. Focusing solely on the narrow ranges of the clusters neglects the valuable information and differences present in the broader range of data points.

A.3 Extra Plots and Tables

Table 27: Data Dictionary for the PMD.

Amenity	Definition
Employment	Measures the closeness of a dissemination block to any dissemination block with a source of employment within a driving distance of 10 km. This measure is derived from the employment counts of all businesses – that is, all North American Industry Classification (NAICS) codes in the Business Register.
Grocery	Measures the closeness of a dissemination block to any dissemination block with a grocery store within a walking distance of 1 km. This measure is derived from the total revenue of all NAICS 4451 businesses in the Business Register.
Pharmacy	Measures the closeness of a dissemination block to any dissemination block with a pharmacy or a drug store within a walking distance of 1 km. This measure is derived from the presence of all NAICS 446110 businesses in the Business Register.
Health care	Measures the closeness of a dissemination block to any dissemination block with a health care facility within a driving distance of 3 km. This measure is derived from the employment counts of all NAICS 6211, 6212, 6213, 621494, and 622 businesses in the Business Register.
Child care	Measures the closeness of a dissemination block to any dissemination block with a child care facility within a walking distance of 1.5 km. This measure is derived from the presence of all NAICS 624410 businesses in the Business Register.
Primary Education	Measures the proximity to primary education measures the closeness of a dissemination block to any dissemination block with a primary school within a walking distance of 1.5 km. Primary schools are classified as education facilities with an International Standard Classification of education (ISCED) level of 1. The data source is a conglomeration of the Open Database of Education Facilities and other sources of education facilities.
Secondary Education	Measures the closeness of a dissemination block to any dissemination block with a secondary school within a walking distance of 1.5 km. The data source is a conglomeration of the Open Database of Education Facilities and other sources of education facilities where secondary schools are classified as ISCED2 and/or ISCED3.
Transit	Measures the closeness of a dissemination block to any source of public transportation within a 1 km walking distance. This measure is derived from the number of all trips between 7:00 a.m. - 10:00 a.m. from a conglomeration of General Transit Feed Specification (GTFS) data sources.
Parks	Measures the closeness of a dissemination block to any dissemination block with a neighborhood park within a 1 km walking distance. This measure is derived from the presence of all parks from a conglomeration of authoritative open data sources and OpenStreetMap.
Libraries	Measures the closeness of a dissemination block to any dissemination block with a library within a 1.5 km walking distance. This measure is derived from the presence of all libraries from a conglomeration of open and publicly available data sources.
Amenity Dense	An aggregate measure was created to indicate neighbourhoods that have access to basic needs for a family with minors. A dissemination block with access to a grocery store, pharmacy, health care facility, child care facility, primary school, library, public transit stop, and source of employment is referred to as an amenity dense neighbourhood. A high amenity density neighbourhood is defined as an amenity dense neighbourhood that has proximity measure values in the top third of the distribution for each of the eight proximity measures.

Table 28: All acronyms used in this report.

Acronym	**Full Name**
PMD	Proximity Measures Database
DEIL	Data Exploration and Integration Lab
DB	Dissemination Block
StatCan	Statistics Canada
IoR	Index of Remoteness
CSD	Census Subdivision
EDA	Exploratory data analysis
VAT	Visual Assessment of Tendency
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
PAM	Partitioning Around Medoids
OPTICS	Ordering Points To Identify Clustering Structure
CMA	Census Metropolitan Area
CA	Census Agglomeration

Table 29: All mathematical symbols used in this report.

Symbol	**Meaning**
k	Number of clusters
θ	MixAll hyperparameter
t	MixAll probability
p	MixAll mixture proportion
ϵ	OPTICS maximum separation distance

Table 30: Summary statistics of log-transformed numerical variables from the PMD.

Moments	Employment	Pharmacy	Childcare	Healthcare	Grocery	Pri. Educ.	Sec. Educ.	Library	Parks	Transit
10% Dec.	-8.5172	-4.8796	-4.8283	-8.1117	-4.2336	-3.4420	-3.2834	-2.9779	-4.3583	-6.7254
20% Dec.	-7.6009	-4.6152	-4.1799	-7.1309	-3.8077	-3.1773	-3.1653	-2.8842	-3.8922	-5.9145
30% Dec.	-6.5713	-4.2199	-3.7214	-6.2659	-3.5405	-2.8422	-3.0241	-2.7726	-3.5791	-5.3817
40% Dec.	-5.7764	-3.9425	-3.3553	-5.7138	-3.3553	-2.6297	-2.8353	-2.6479	-3.2888	-4.9908
50% Dec.	-5.0207	-3.6613	-3.0428	-5.2785	-3.1350	-2.4068	-2.5956	-2.5072	-3.0324	-4.6565
60% Dec.	-4.3583	-3.3755	-2.7536	-4.8929	-2.8896	-2.2018	-2.3958	-2.3424	-2.7887	-4.3275
70% Dec.	-3.8258	-3.0835	-2.4686	-4.4918	-2.6311	-1.9900	-2.1698	-2.1464	-2.5333	-3.9900
80% Dec.	-3.2995	-2.7458	-2.1473	-3.9900	-2.3167	-1.7597	-1.9018	-1.9045	-2.2528	-3.6009
90% Dec.	-2.6214	-2.3187	-1.7418	-3.3697	-1.8702	-1.4563	-1.5469	-1.5573	-1.9005	-3.1168
Min.	-9.2103	-9.2103	-9.2103	-9.2103	-8.5172	-7.6009	-7.4186	-8.5172	-9.2103	-9.2103
Median	-5.0207	-3.6613	-3.0428	-5.2785	-3.1350	-2.4068	-2.5956	-2.5072	-3.0324	-4.6565
Mean	-5.3064	-3.6087	-3.1445	-5.5035	-3.0844	-2.4178	-2.5101	-2.3698	-3.0670	-4.8040
Max.	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
Std. Dev.	2.1556	0.9607	1.1298	1.7612	0.9114	0.7301	0.6680	0.5789	0.9174	1.4123
Skew	-0.2371	0.3630	-0.2076	-0.3443	0.1324	0.1125	0.6055	1.0242	-0.1541	-0.5501
Kurtosis	2.0032	2.5437	2.4316	2.5304	2.8460	2.3186	2.6266	4.1390	3.0366	3.2069

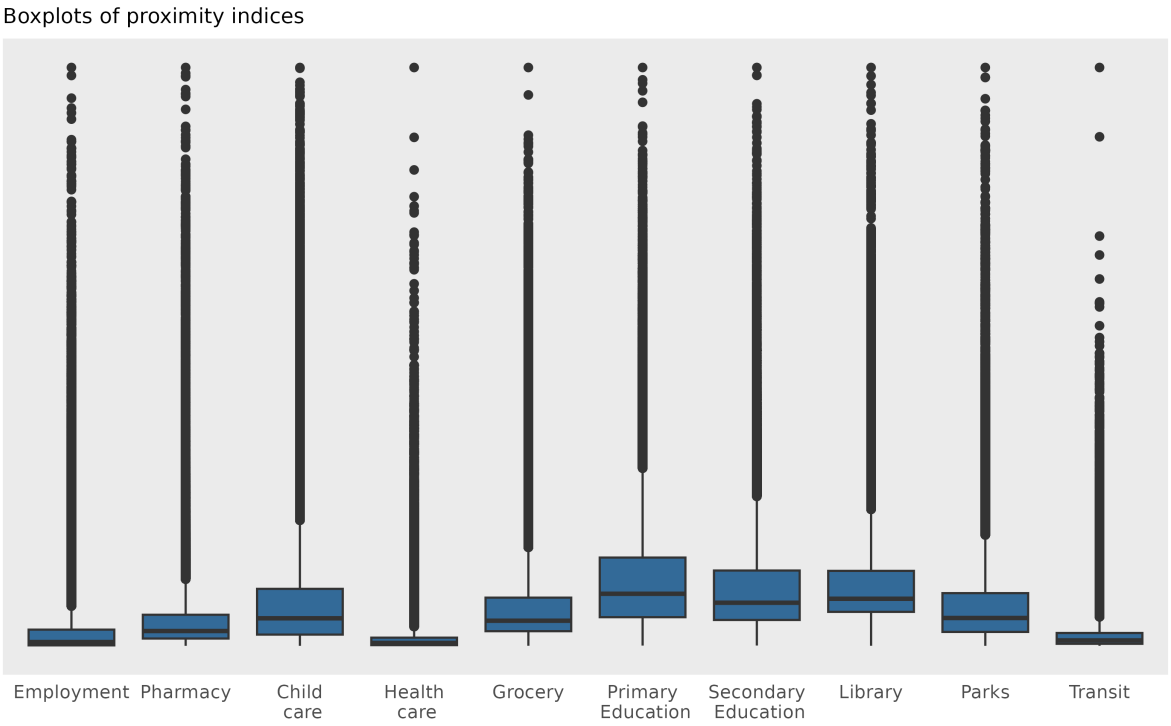

Figure 38: Boxplots showing outliers for all ten amenities of the PMD.

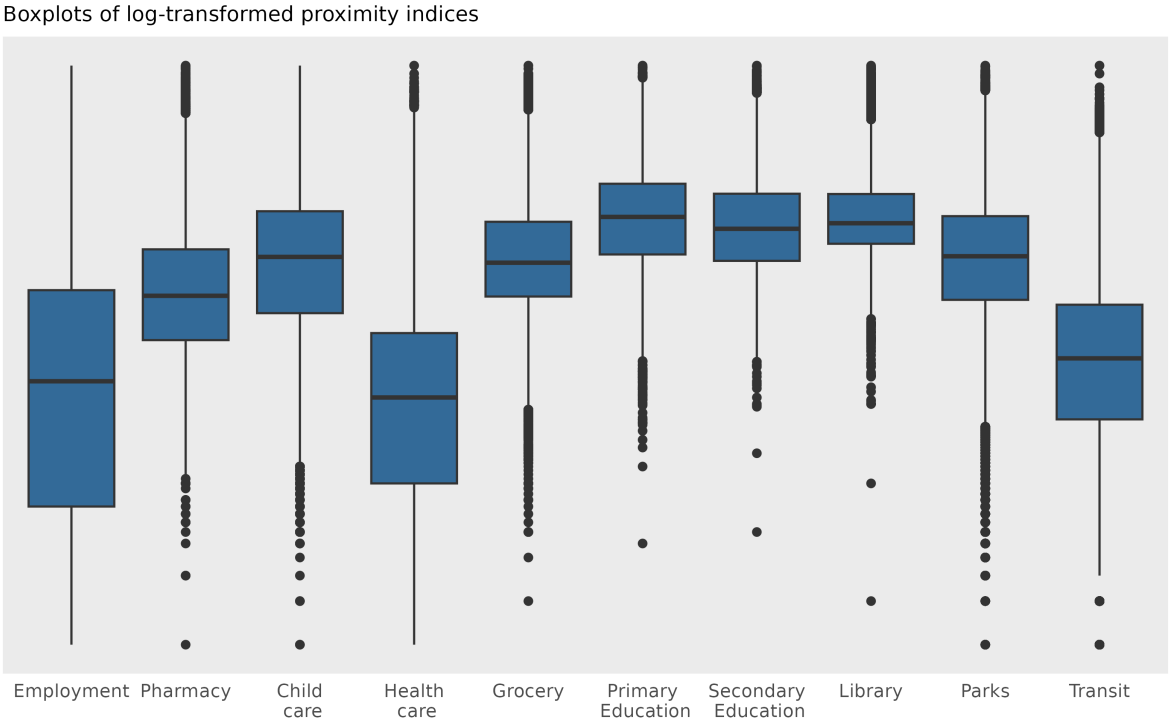

Figure 39: Boxplots showing outliers for all ten log-transformed amenities of the PMD.

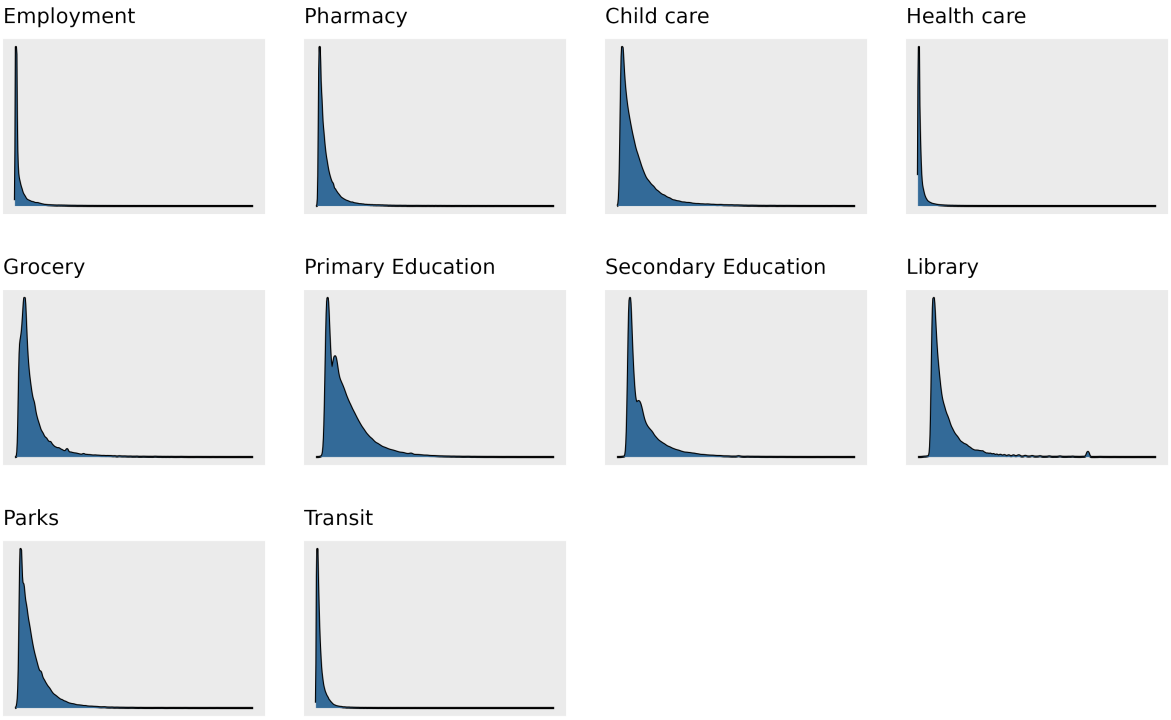

Figure 40: Density distributions for all ten amenities of the PMD.

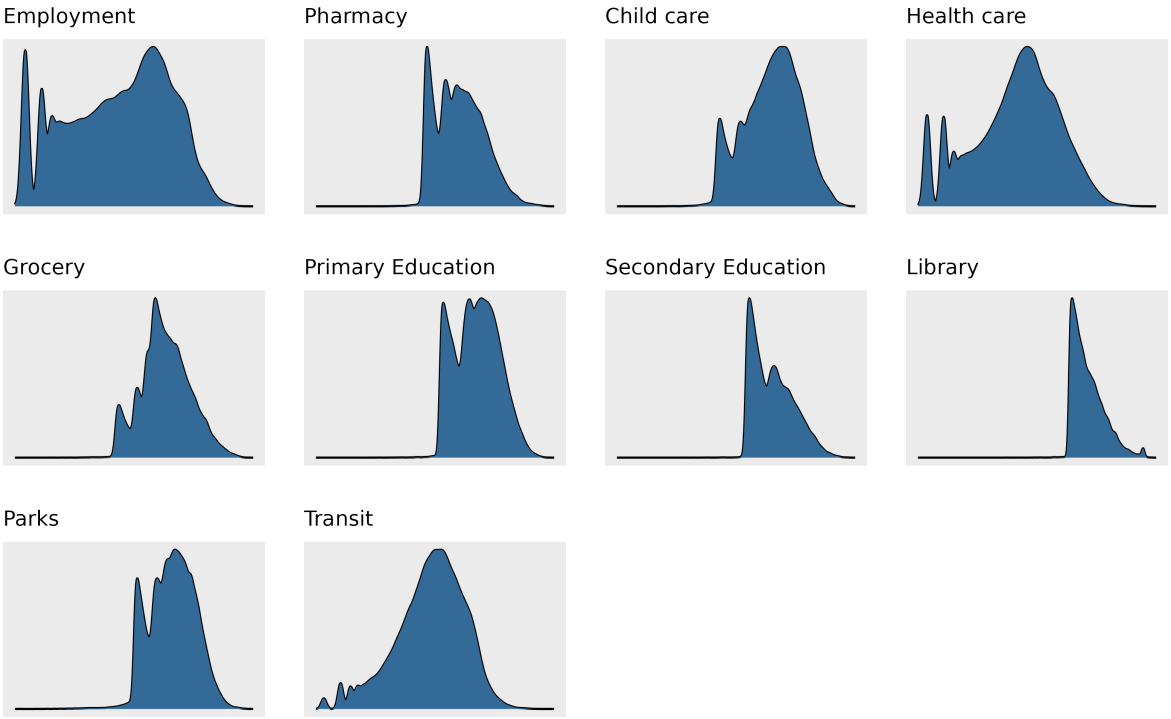

Figure 41: Log-transformed density distributions for all ten amenities of the PMD.

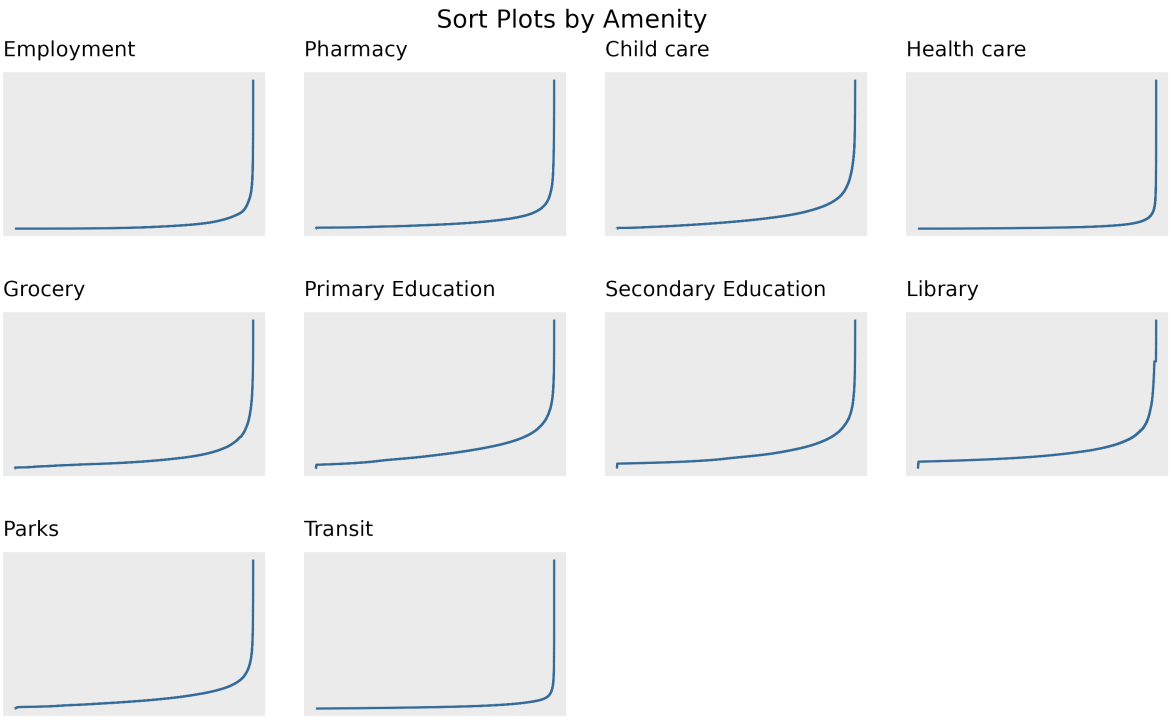

Figure 42: Sort plots for each amenity in the PMD.

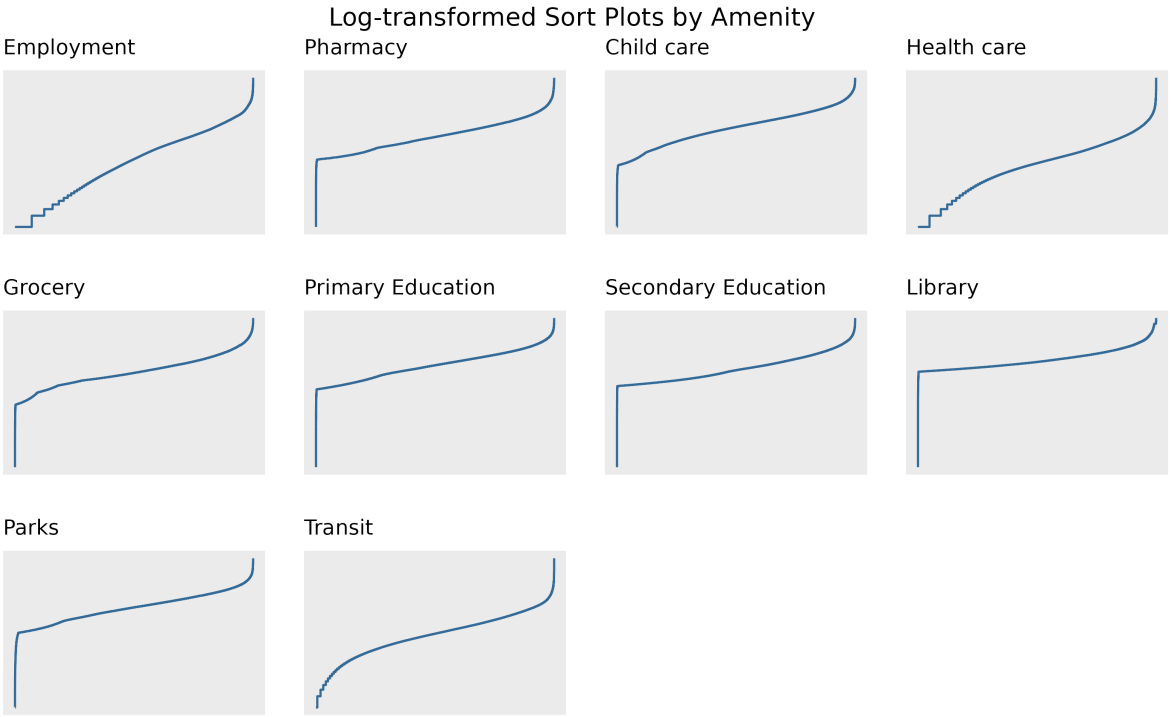

Figure 43: Log-transformed sort plots for each amenity in the PMD.