Spatial patterns of seasonal ridership, built environment and social demographics: A taxonomy of Bike Share Toronto stations

Zehui Yin [redacted]

1. Background

Cities are growing larger and attracting more residents with the impact of globalization and improvements in transportation. With higher demand in mobility needs, challenges toward existing transportation infrastructures arise. The City of Toronto nowadays is much larger than the previous Old Toronto before 1998. In the meantime, the enlargement of the city limit seems doesn't accompany sufficient improvement in mobility. Allen et al. (2022) discovered that Toronto suffers from the highest extreme commuter percentage in Canada, which is the percentage of one-way commuting over 60 minutes. This pose concerns over the traditional public transportation methods.

The bicycle-sharing program is a relatively new transportation mode compared to time-honoured street cars, buses, and subways. The City of Toronto established Bike Share Toronto in 2011 as part of the Toronto Bike Plan. Currently, Bike Share Toronto operates a system of 626 stations and 7185 bikes 24 hours per day and in all seasons. This program complements the historical existing public transportation system, reduces bike theft concerns for cyclists, and provides environmental benefits (El-Assi et al., 2017). However, there are also problems with the program. There are only 18.1% of the total population lives within the bicycle share service area in Toronto. Also, most bicycle share stations are located in or around the downtown core, although provide relatively equitable access compared with other major cities in Canada (Hosford & Winters,

2018). Many periphery areas in Toronto indeed don't have access to the bike share system, which indicates huge potential for future improvement for the system.

The Bike Share Toronto system has advantages and disadvantages. Therefore, more insights are needed to fully understand the station system characteristics to potentially identify stations or types of stations that need adjustment or have improvement potential. In the existing literature, when it comes to bike-share systems, more attention is given to demand prediction, accessibility, and equality (Cheng et al., 2022; El-Assi et al., 2017; Hosford & Winters, 2018). They are useful in providing general information about the performance and what impacts the performance regarding the whole bike share system. When it comes to specific cities or systems, these generalized insights are not sufficient to identify the characteristics of individuals or groups of stations within a certain system. Therefore, case studies are needed for individual cities to account for the unique geographies, cultures, and socio-demographic characteristics.

2. Research Questions

In this project, I fill the research gap by conducting a case study of the Bike Share Toronto system with a focus on two research questions below:

RQ1. What are the seasonal spatial pattern differences in Toronto bike-share trips at station levels?

RQ2. What are the similarities and differences between the locations, trip count percentages, built environments, and social demographics of Toronto bike-share stations?

3. Data

In this project, my study area is the City of Toronto. It is the capital of Ontario and the most populous city in Canada. Data used in the analysis are collected from four sources Toronto Open Data, Statistics Canada, CHESS, and Scholars GeoPortal. I used Toronto bike-share trip data from 2021 and the station data in 2022. Due to data availability constraints, data used are generally not collected in the same year. Several data sources are utilized to calculate some socio-demographic and built environment variables at the station level for analysis and comparison. Schools and places of interest in Toronto are considered to be built environment features of the bikeshare stations. Stations close to schools or tourist attractions could attract more student or traveller ridership, respectively. Table 1 below provides a summary of the data sources used in this paper.

Table 1: Data sources

Name	Description	Format	Source (URL)
Bikeshare ridership 2021	Toronto bike-share trip data in 2021	CSV	City of Toronto Open Data
Bikeshare station data	Toronto bike-share station information in 2022	JSON	Toronto Parking Authority through GBFS
2016 census	2016 census results by census tract	CSV	Statistics Canada through CHESS
2016 census tract	2016 census tract boundaries	Shapefile	Statistics Canada
2014 land use	2014 Land use data in Ontario	Shapefile	DMTI Spatial Inc through Scholars GeoPortal
Bikeways	Bikeways in Toronto	Shapefile	City of Toronto Open Data
Intersection	Road intersections in Toronto	Shapefile	City of Toronto Open Data
School	All types of school locations	Shapefile	City of Toronto Open Data
Place of interest	Places of Interest and Toronto Attractions	Shapefile	City of Toronto Open Data

4. Methods

All the Toronto bike share trips in 2021 are classified, based on the standard season interpretation, into four seasons: spring (March, April, May), summer (June, July, August), autumn (September, October, November), and winter (December, January, February).

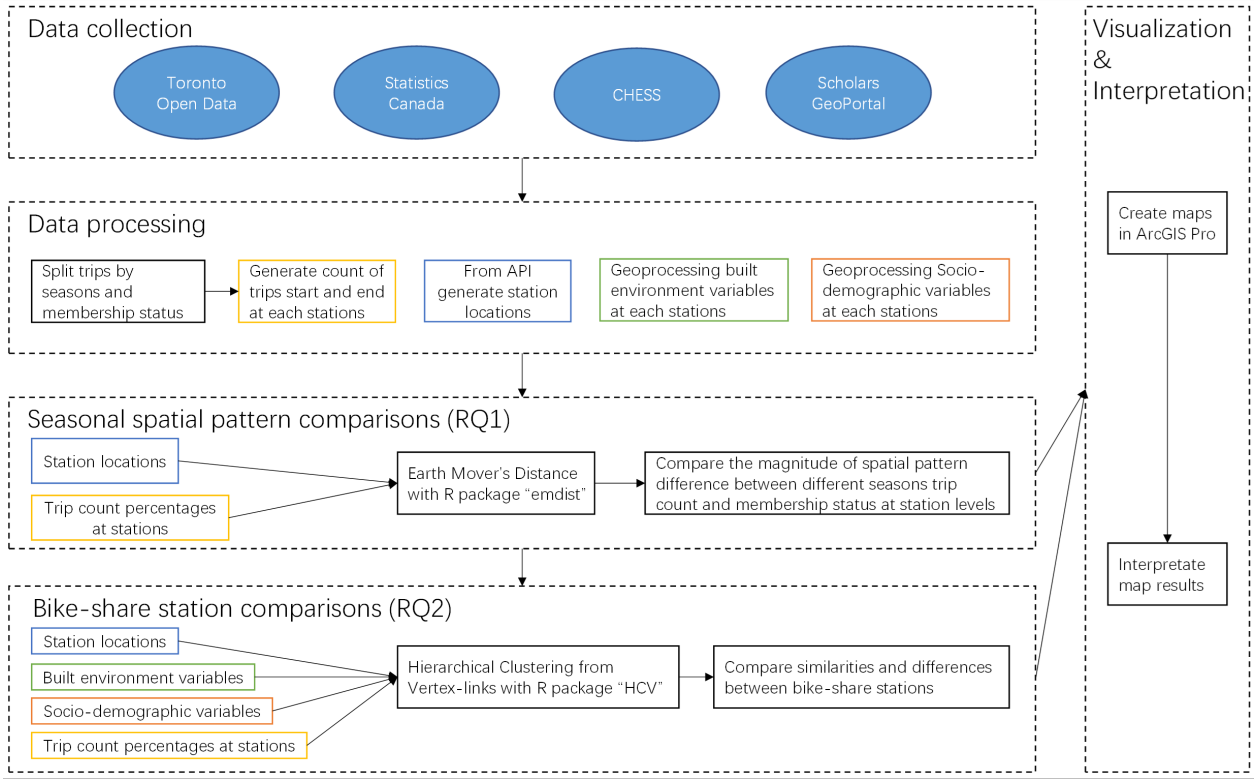

Figure 1: Workflow diagram

Since Toronto Bike Share is a station-based system, I aggregated trip counts to their origin and destination stations separated by membership status. To account for seasonal absolute trip number differences, as relative seasonal spatial patterns are more of interest, the trip counts at every station are standardized to proportions for each season. Also, some built environment and social demographic variables are calculated at station levels using areal weighting with 500-meter buffers. Social-demographic variables are based on 2016 Census data at the Census tract level. Table 2 provides descriptive statistics of these variables. In the existing literature, the buffer radiuses used in the analyses range from 50 to 500 meters (Cheng et al., 2022; El-Assi et al., 2017; Hosford & Winters, 2018). In this analysis, I decided to use relatively larger buffers, as it is normal and possible for regular people to walk for 500 meters. Figure 1 above illustrates the workflow for this essay.

4.1 Earth Mover's Distance

To address RQ1, I calculated the two-dimensional Earth Mover's Distance (EMD) between the spatial distributions of bike share trip count percentages at station levels (Rubner et al., 2000). The computation is performed in R with the package "emdist" (Urbanek & Rubner, 2022). EMD computes the cost of converting one distribution to the other and can be used as a metric to compare the similarity or differences between two equal-sized multi-dimensional matrices (McKenzie, 2019). Specifically, the coordinates used in the calculation are easting (x) and northing (y) of the stations from the local projection NAD83 / UTM zone 17N (ESPG: 26917) while the trip count percentages at stations are used as weightings.

It is worth mentioning that a higher EMD indicates larger costs needed to convert one distribution to the other or the two distributions are more different from each other. In the meantime, a smaller EMD implies that the two distributions are similar to each other.

4.2 Hierarchical Clustering from Vertex-links

The traditional k-means, single-link or complete-link agglomerative hierarchical clustering methods cannot account for both spatial proximity and attribute similarity (Liu, 2012). Omitting spatial distribution would jeopardize the clustering results, therefore, to answer RQ2, Hierarchical Clustering from Vertex-links is calculated based on distances between stations in both attribute space and geographic space in R with the package "HCV" (Tzeng & Hsu, 2022). Tzeng and Hsu (2022) pointed out that the fundamental agglomerative clustering accounting for the constraints of vertex links on the geometry domain approach guaranteed the resulting clusters to be spatially

continuous by the enforcement at each step of iterations. Therefore, the resulting clusters would not be fragmented or have multiple parts that are disconnected. To further account for the correlation between the attributes of stations, Mahalanobis distance, which is calculated using the inverse of the covariance matrix of the dataset of interest, is used to measure the differences between stations in attribute spaces (De Maesschalck et al., 2000).

5. Results

In this section, I first provide the results of the descriptive statistics for the calculated social-demographic and built environment variables and the bike share trips. Then, I answer RQ1 based on EMD results and the map visualizations. In the end, I summarize and interpret the clustering results to address RQ2.

5.1 Descriptive Statistics

Based on the results of the descriptive statistics, summer is the busiest season in the year when it comes to total trip counts, while in spring riders are more likely to travel longer times compared to any other time. The casual trips drop extremely in winter which is below 10% of counterparts in other seasons.

When it comes to bike share station level variables, these stations tend to locate more within residential land use type. Also, the average employment density is relatively lower than the population density. These stations situate less in the employment zones, although being mostly located in the downtown core areas.

Table 2: Descriptive statistics

Bike share trips			N = 3575182	
Seasons	*Member trips*	*Casual trips*	*n*	*Mean trip durations*
Spring	512981	292738	805719	1218.381
Summer	895422	518078	1413500	1087.625
Autumn	509649	555215	1064864	901.0014
Winter	222678	68421	291099	778.396
Bike share stations		N = 626		
Variable name	*Calculation method*	*Mean*	*Median*	*Standard deviation*
bikeway_length	Bikeway length falls within 500-meter buffers of stations	2404.005	2403.584	1077.892
landuse_commercial	Commercial land use percentage within 500-meter buffers of stations	4.102862	0	7.918675
landuse_institutional	Institutional land use percentage within 500-meter buffers of stations	10.03356	6.458006	12.46606
landuse_open_area	Open area land use percentage within 500-meter buffers of stations	3.858952	2.406328	4.15139
landuse_recreational	Recreational land use percentage within 500-meter buffers of stations	10.3438	6.314922	12.47046
landuse_residential	Residential land use percentage within 500-meter buffers of stations	56.78685	60.33764	23.71126
landuse_industrial	Industrial land use percentage within 500-meter buffers of stations	11.67621	3.234388	15.34019
landuse_waterbody	Waterbody land use percentage within 500-meter buffers of stations	3.194515	0	9.619545
landuse_entropy	$Entropy = -\sum_{k=1}^{n} P_k * \frac{\ln(P_k)}{\ln(n)}$	0.5120056	0.5227071	0.1692294
population_density	Average population density within 500-meter buffers of stations	10789.84	9542.587	5817.102
employment_density	Average employment density within 500-meter buffers of stations	6503.855	5517.437	3913.683
median_income	Median income from nearest census tract centroid to stations	73382.09	68978	31042
average_age	Average age within 500-meter buffers of stations	39.73445	39.85261	2.817437
street_connectivity	Number of intersections within 500-meter buffers of stations	136.9968	134	51.99037
school_presence	Whether a school falls within 500-meter buffers of stations	0.8306709	1	0.3753422
POI_presence	Whether a place of interest falls within 500-meter buffers of stations	0.5191693	1	0.5000319

5.2 Seasonal Spatial Pattern Comparisons

The EMD between the same season's origin and destination are all around 100. There are negligible seasonal variations in the distribution of trips' origins and destinations. Also, there are little changes observed in the EMD when change both comparing distributions from origins to destinations. Therefore, the same season's trip origin and destination distributions are very similar to each other throughout the year 2021. Table 1 below shows the EMD for different pairs of distributions. Note that the origin and destination columns denote comparisons between two origin distributions or between two destination distributions, respectively.

Table 3: Earth Mover's Distance results

Distribution 1	**Distribution 2**	**EMD**	
Compare origin and destination distributions in the same seasons			
Total trip origins in spring	Total trip destinations in spring	93.35248	
Total trip origins in summer	Total trip destinations in summer	102.9498	
Total trip origins in autumn	Total trip destinations in autumn	114.0404	
Total trip origins in winter	Total trip destinations in winter	101.9681	
Compare spatial distributions in different seasons		**Origin**	**Destination**
Total trips in spring	Total trips in summer	370.1424	377.5933
Total trips in spring	Total trips in autumn	817.3242	832.0513
Total trips in spring	Total trips in winter	978.1482	988.6324
Total trips in summer	Total trips in autumn	481.1	486.5856
Total trips in summer	Total trips in winter	651.4673	655.3649
Total trips in autumn	Total trips in winter	209.2595	215.855
Compare user-type distributions in the same seasons		**Origin**	**Destination**
Member trips in spring	Casual trips in spring	1630.353	1630.665
Member trips in summer	Casual trips in summer	1110.518	1106.718
Member trips in autumn	Casual trips in autumn	368.7628	362.5514
Member trips in winter	Casual trips in winter	335.955	333.3849

When it comes to different season comparisons, surprisingly, the highest EMD observed is between spring and winter. Although these two seasons are temporally next to each other and have relatively smaller temperature differences compared to summer and winter, there is a large

difference in trip spatial patterns. Also, the second largest EMD is between spring and autumn, while the smallest EMD is between autumn and winter.

For comparison between user-type distributions, there are huge EMD between members and non-member trip distributions in spring and summer, while little distance is observed in the counterparts between autumn and winter. There are relatively minor spatial pattern differences between member and non-member trips during autumn and winter compared to spring and summer.

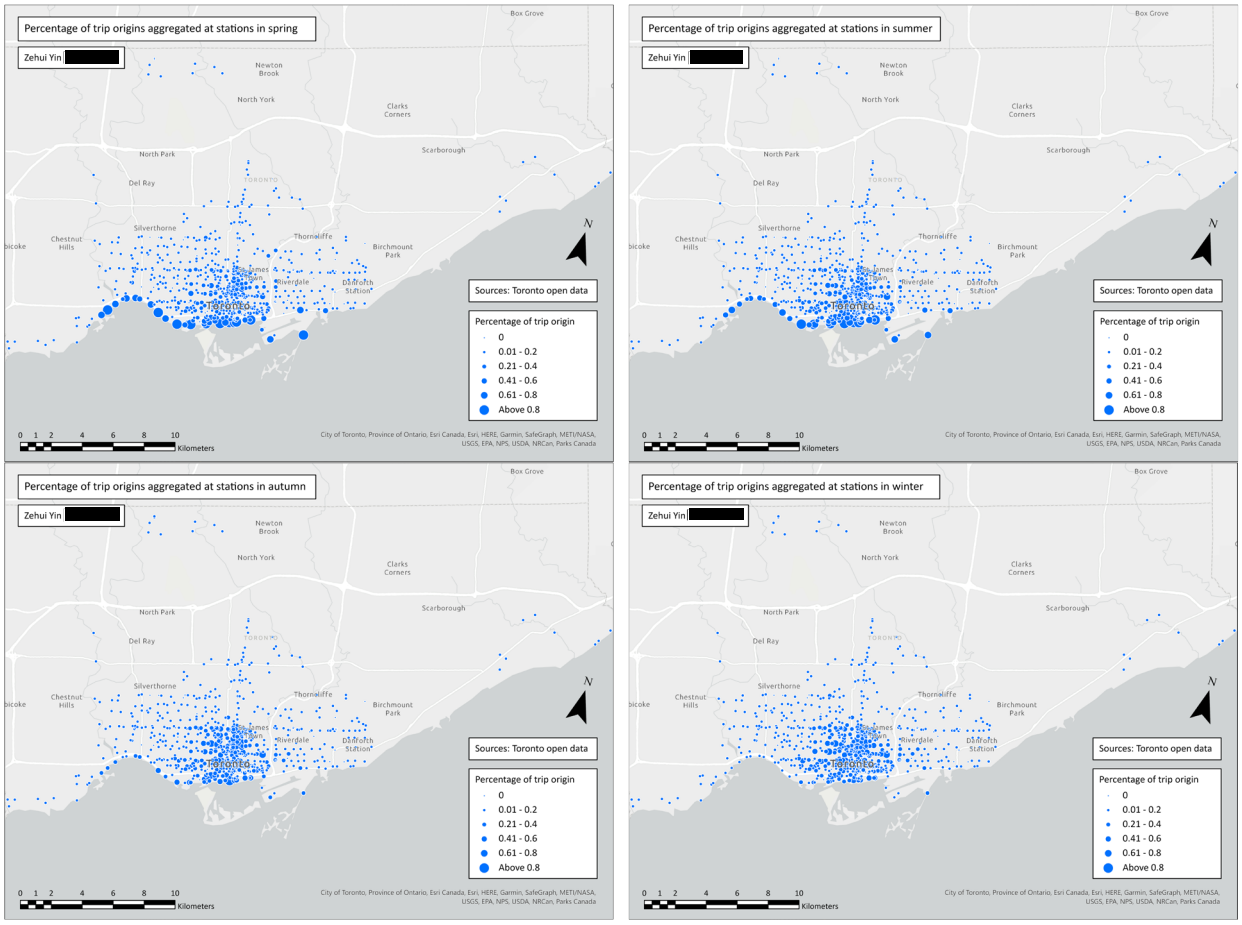

Figure 2: Seasonal trip origin distribution patterns

After examining the quantified differences, I visualized the percentage of trip origins aggregated at stations in different seasons to further investigate the spatial differences. Based on figure 2 above, the stations located along Lake Ontario have extremely high demand during spring

and summer, while the stations located in the downtown core areas have relatively stable high demand throughout the year. The high usage of lake-side stations in the warm seasons is most likely due to the casual riders' recreational activities, while the stable usage of downtown core stations is likely related to annual members who use this service more for commuting.

5.3 Bike Share Station Comparisons

Hierarchical clustering with vertex link based on Mahalanobis distance results in the dendrogram shown below. There is an outlier station (which is the only station forming cluster 6) named "Widmer St / Adelaide St W" which itself forms a cluster. This station, based on trip data, is active in 2021 and had a large number of trips start or end there. However, since the station data is collected in 2022, this particular station is disabled currently and has been replaced with another station at its location with a different name.

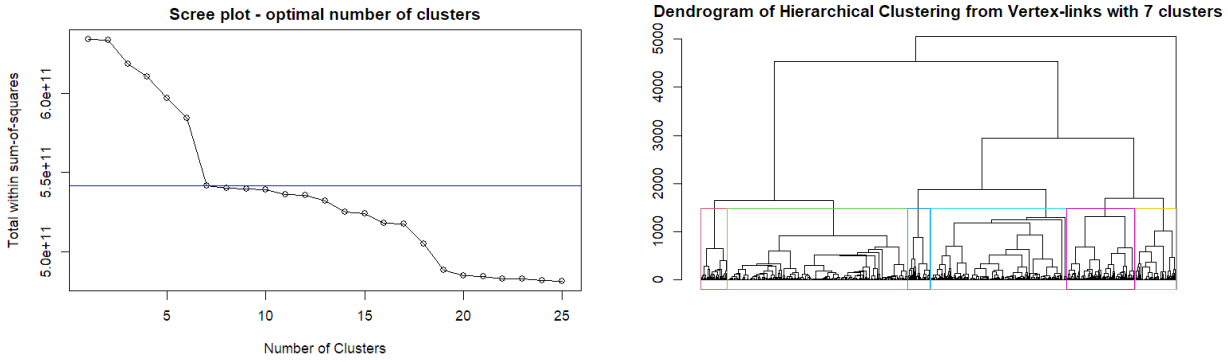

Figure 3: Scree plot and dendrogram of hierarchical clustering from vertex-link

The scree plot above is based on the total within sum-of-squares. Intuitively, since the clustering is based on Mahalanobis distance, the total within sum-of-squares should also be calculated based on this distance calculation approach. However, although the covariance matrix for the whole dataset is invertible, due to the much smaller number of stations within each cluster,

some clusters' covariance matrices are no longer invertible. Therefore, alternatively, the within sum-of-squares shown in the scree plot is an approximation calculated based on Euclidean distance between stations in attribute space. The Euclidean distance as an approximation facilitates the further interpretation of the attributes. As the interpretation is based purely on attributes, larger attribute differences without rescaling between groups are desired. Based on the scree plot, I choose to classify the stations into 7 clusters, since the total within sum-of-squares curve flattens at 7.

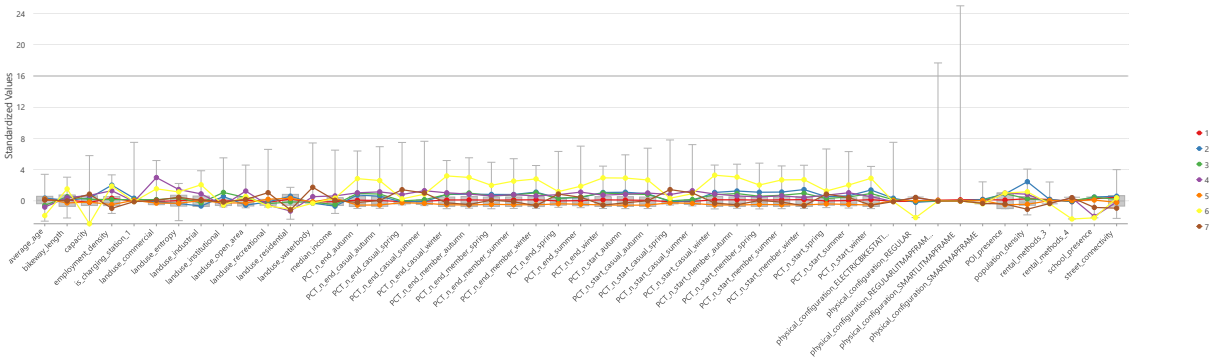

Figure 4: Boxplot of variables grouped by clusters

According to the figure 4 boxplot above, cluster 1 which consists of 179 stations identifies the average stations in the whole system. Their attribute averages are extremely close to the whole system medians at every attribute. They located in a relatively central part of Toronto being surrounded by other clusters. 35 stations located in high population and employment density zones consist of cluster 2. These stations are adjacent to relatively low-income residential areas around St James Town. Cluster 2 stations enjoy the highest usage among bike-share annual members in all seasons. Cluster 3 which has 90 stations are mostly located within institutional land-use areas and have more schools around them, which are mostly the University of Toronto Saint George

Campus. The usage pattern of these stations is very similar to the counterparts of cluster 2 stations, although the general usage level is slightly lower. Cluster 4 containing 30 stations is located within the most diverse land-use areas. These stations are mostly located in commercial, industrial and open space areas while being free from school presences. They have the highest usage in autumn compared to any other clusters.

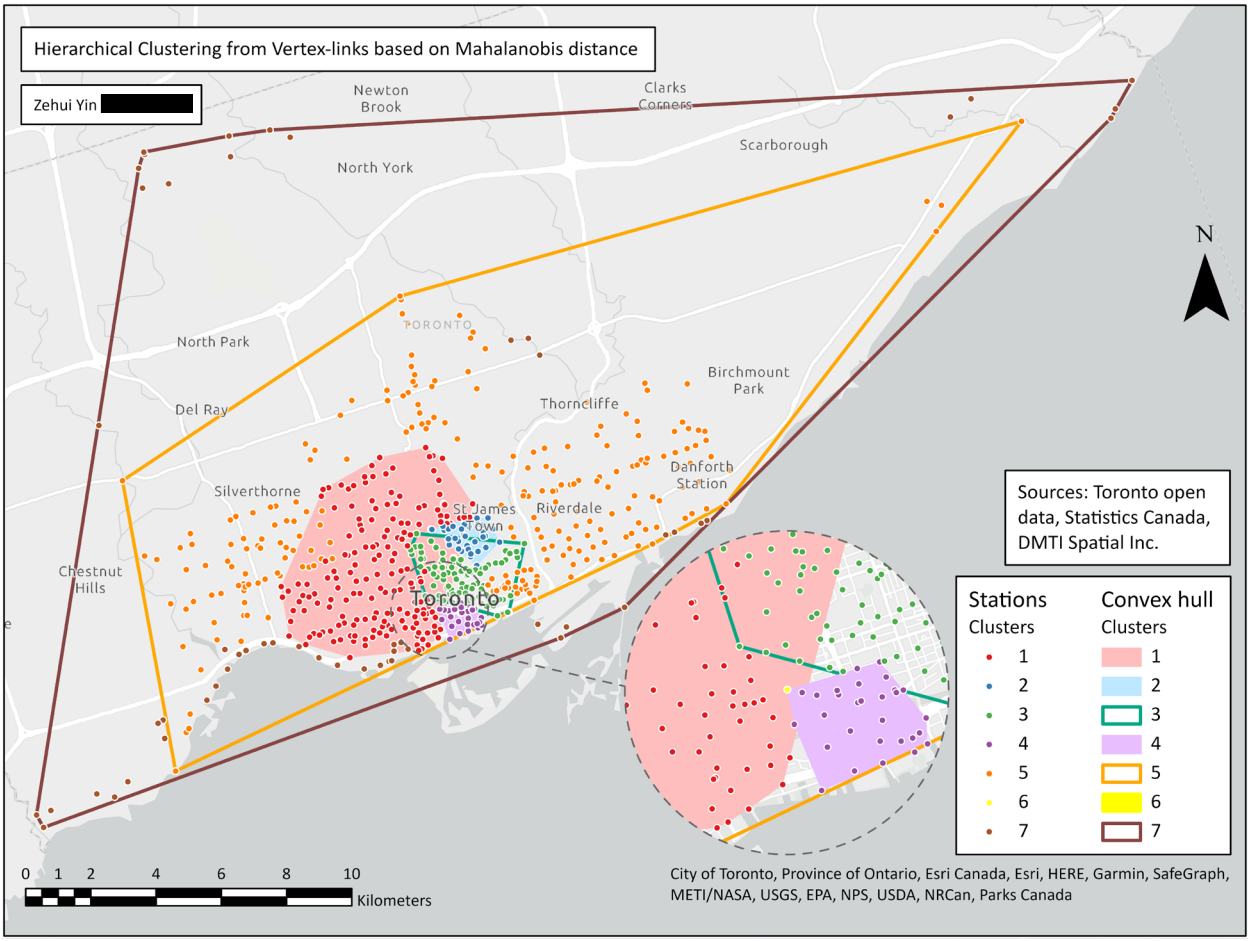

Figure 5: Map of clustering result

Moving away from the downtown areas, 237 stations forming cluster 5 suffer from the lowest demand in the whole system throughout the year. They are located at the transition ring between the downtown core and the peripheral areas of Toronto. Most of them are located within residential areas with low land use diversity whilst having low population and employment

densities. It is also worth mentioning that both the capacities are low, and the bikeway infrastructures are poor at these stations. There are 54 stations forming cluster 7 which are situated at the edges of the whole system. These stations have the highest station capacities and are mostly located in recreation areas and around water bodies. They enjoy the highest casual usage in spring and summer while suffering low demand from other seasons and types of riders.

6. Discussion

According to the seasonal spatial pattern comparisons, trip distributions are similar between spring and summer and between autumn and winter, while there is a huge difference between spring and winter trip patterns. There are significant temperature differences in different seasons in Toronto. El-Assi et al. (2017) also found that there is a positive correlation between warmer temperatures and bike share usage. Therefore, it is no surprise that the warm seasons are different from the cold seasons. However, apart from temperature, I found the difference could also arise from the temporal sequence, as spring and winter are temporally connected. The period of temperature change from winter to spring is arguably the most noticeable time, adding people might have also spent a much larger portion of their time indoors in winter. When spring comes and the temperature starts increasing, this provides a large momentum for people to go outdoors and therefore changes especially the casual usage patterns for bike share most significantly. This interpretation is also supported by the high usage of lake-side stations starting in spring.

Based on the results from the bike share station comparisons, cluster 1 denotes the average stations in the system. Cluster 2 represents the residential and employment zone stations.

Institutional zone stations form cluster 3. Cluster 4 indicates the commercial zone stations. The transitional zone stations are captured by cluster 5. Cluster 6 is the outlier station. Cluster 7 denotes the recreational zone stations.

According to the findings above, there is little spatial pattern difference between the member and non-member trips in autumn and winter. This might indicate that riders, although being of different user types, used the bike share services in similar ways like for commuting purposes. It could be a good idea to introduce reduced seasonal casual trip pricing for autumn and winter, which can not only boost usage of the bike share system in these under-demanded seasons but also promotes equitable transportation services for commuters. Also, reduced trip pricing or membership discounts can be offered to riders based on geographic locations. For example, offers can be distributed through mobile apps to riders who constantly have their trips start or end around or at St James Town, as stations in this region serve relatively more vulnerable communities. In the meantime, the stations that are collocated with recreational sites, although being located at the edges of the whole system, have relatively higher usages in spring and summer. It might be worth trying to introduce more recreational areas in the transition areas to attract more riders to use these stations.

There are some limitations in the analysis arising from the data challenge. The data used in the analysis are generally not collected in the same year. The outlier station which forms cluster 6 results purely due to the mismatch between 2021 trip data and 2022 station data. If there are any significant yearly variations in the socio-demographic, built environment variables, and station characteristics like locations or capacities, the reliability of taxonomy is likely to be undermined.

Also, it is beyond the scope of this study to conclude any correlation or causation between any of the variables and trip counts at station levels. Future studies are needed to address these problems.

7. Conclusion

In this case study, I provide insights into the Toronto bike share system at station levels. These findings can function as the first step toward the understanding of the usage of bike share services in Toronto. I explored the seasonal trip pattern differences at station levels grouped by rider types, as well as similarities and differences between stations at different locations.

I found that there is little difference between trip origin and destination's spatial patterns throughout the year. The trip spatial patterns are similar between spring and summer, as well as between autumn and winter. There is seasonal high demand for stations located along Lake Ontario in spring and summer. In the meantime, the downtown core stations enjoy stable high demands in all seasons. The differences between member and non-member riders are most significant during spring and summer while being negligible in autumn and winter. The stations located within the transition zone between the center and the edges of Toronto have the lowest usage among the whole bike share system throughout the year. Meanwhile, the stations situated at the edges of the system mostly enjoy having recreational sites or waterbodies around them and have relatively higher casual riders in spring and summer.

These results provide novel understandings of the performance of the Toronto bike share system and rider preferences for using the system, which can be used to support or formulate further developments or government policies.

References

Allen, J., Palm, M., Tiznado-Aitken, I., & Farber, S. (2022). Inequalities of extreme commuting across Canada. *Travel Behaviour and Society*, *29*, 42-52. https://doi.org/10.1016/j.tbs.2022.05.005

Cheng, L., Wang, K., De Vos, J., Huang, J., & Witlox, F. (2022). Exploring non-linear built environment effects on the integration of free-floating bike-share and urban rail transport: A quantile regression approach. *Transportation Research Part A: Policy and Practice*, *162*, 175-187. https://doi.org/10.1016/j.tra.2022.05.022

De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. L. (2000). The mahalanobis distance. *Chemometrics and intelligent laboratory systems*, *50*(1), 1-18. https://doi.org/10.1016/S0169-7439(99)00047-7

DMTI Spatial Inc. (2014). *Land Use (LUR)* [digital resource: vector]. Scholars GeoPortal. http://geo.scholarsportal.info/#r/details/_uri@=2785150059

Economic Development & Culture. (2022). *Places of Interest and Toronto Attractions* [digital resource: vector]. City of Toronto Open Data. https://open.toronto.ca/dataset/places-of-interest-and-toronto-attractions/

El-Assi, W., Salah Mahmoud, M., & Nurul Habib, K. (2017). Effects of built environment and weather on bike sharing demand: a station level analysis of commercial bike sharing in Toronto. *Transportation*, *44*(3), 589-613. https://doi.org/10.1007/s11116-015-9669-z

Hosford, K., & Winters, M. (2018). Who are public bicycle share programs serving? An

evaluation of the equity of spatial access to bicycle share service areas in Canadian cities. *Transportation research record*, *2672*(36), 42-50. https://doi.org/10.1177/0361198118783107

Information & Technology. (2020). *Intersection File - City of Toronto* [digital resource: vector]. City of Toronto Open Data. https://open.toronto.ca/dataset/intersection-file-city-of-toronto/

Liu, Q., Deng, M., Shi, Y., & Wang, J. (2012). A density-based spatial clustering algorithm considering both spatial proximity and attribute similarity. *Computers & Geosciences*, *46*, 296-309. https://doi.org/10.1016/j.cageo.2011.12.017

McKenzie, G. (2019). Spatiotemporal comparative analysis of scooter-share and bike-share usage patterns in Washington, DC. *Journal of transport geography*, *78*, 19-28. https://doi.org/10.1016/j.jtrangeo.2019.05.007

Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, *40*(2), 99-121. https://doi.org/10.1023/A:1026543900054

Statistics Canada. (n.d.). *2016 Census Tracts Cartographic Boundary File* [digital resource: vector]. https://www12.statcan.gc.ca/census-recensement/2011/geo/bound-limit/files-fichiers/2016/lct_000b16a_e.zip

Statistics Canada. (2017). *2016 Census Profiles Files / Profile of Census Tracts* [Data set]. Computing in the Humanities and Social Sciences (CHASS) at the University of Toronto. http://dc.chass.utoronto.ca.myaccess.library.utoronto.ca/cgi-

bin/census/2016/displayCensus.cgi?year=2016&geo=ct#vars

Toronto Parking Authority. (2021). *Bike Share Toronto Ridership Data* [Data set]. City of Toronto Open Data. https://open.toronto.ca/dataset/bike-share-toronto-ridership-data/

Toronto Parking Authority. (2022). *Bike Share Toronto* [GBFS]. City of Toronto Open Data. https://open.toronto.ca/dataset/bike-share-toronto/

Toronto Police Services. (2022). *School Locations - All Types* [digital resource: vector]. City of Toronto Open Data. https://open.toronto.ca/dataset/school-locations-all-types/

Toronto Transportation Services. (2022). *Bikeways* [digital resource: vector]. City of Toronto Open Data. https://open.toronto.ca/dataset/bikeways/

Tzeng, S., & Hsu, H. Y. (2022). *The R Package HCV for Hierarchical Clustering from Vertex-links*. arXiv. https://doi.org/10.48550/arXiv.2201.08302

Urbanek, S., & Rubner, Y. (2022). *emdist: Earth Mover's Distance*. The Comprehensive R Archive Network. https://CRAN.R-project.org/package=emdist