Systematic Analysis of Public Transit Data Availability in Canada

Kevin Dick
Systems & Computer Engineering
Carleton University
Ottawa, Canada
kevin.dick@carleton.ca

Azizul Hasan†
Systems & Computer Engineering
Carleton University
Ottawa, Canada
†equal contribution

Jamil Dergham†
Systems & Computer Engineering
Carleton University
Ottawa, Canada
†equal contribution

A. J. Clarke†
Systems & Computer Engineering
Carleton University
Ottawa, Canada
†equal contribution

Hoda Khali
Systems & Computer Engineering
Carleton University
Ottawa, Canada
hodakhalil@cunet.carleton.ca

Gabriel Wainer
Systems & Computer Engineering
Carleton University
Ottawa, Canada
gwainer@sce.carleton.ca

***Abstract*—Regional authorities will publish public transit route and timetable service offerings through the General Transit Feed Specification (GTFS) as a standard format. Systematically collected GTFS data can be used to study the structure, organization, and availability of public transit services nationally. In this work, we provide a systematic approach to collection of public transit data from various sources, preparation of a high-quality GTFS data inventory and analysis of the public transit offerings lending to numerous insights about transit availability at various geographic scales. Using Canada as a case study, we collected GTFS for the 213 candidate census subdivisions ([CSDs] representing cities, towns, municipalities, *etc.*) with populations greater than 20,000. These data were then cleaned and leveraged along with CSD-specific census statistics to comprehensively compare public transit offerings between provinces/territories and across CSD types. We determined that, despite a systematic collection process, the majority of CSDs lack official GTFS data, certain provinces are under- or over-represented in our analysis. We further proposed using the median of transit stop spatial density as a national baseline revealing provinces such as Québec are severely lacking in public transit offerings. GTFS data analysis is insightful for understanding the current state and progress in urban public transportation, which is highly relevant to the United Nation's Sustainability Development Goals. Our aggregated dataset and open-sourced codebase are publicly available at: github.com/chazingtheinfinite/canada-transit-study.**

***Index Terms*—GTFS, transit data, data availability, systematic analysis**

I. Introduction

Public transit systems are typically fixed transport routes with spatially separated embarkation/disembarkation stops organized with a timetable of temporally separated trips for use by the general public [1]. These systems help to provide efficient and sustainable mobility in cities globally [2]. Regional authorities will publish service offerings through the General Transit Feed Specification (GTFS) as a standard format [3], [4]. Certain online repositories including TransitLand and TransitFeeds seek to aggregate GTFS data, however there are numerous challenges to the study of public transit offerings.

As outlined by Kujala *et al.* in [3], there exist three major challenges in the systematic collection of GTFS data from independent transit authorities:

- **Fragmentation:** public transit feeds for a given city may be provided by multiple, independent transit authorities requiring the intelligent merger of multiple GTFS feeds for a single region of interest.
- **Redundancy:** directly in opposition to the first challenge, at times the GTFS feeds for a given sub-region actually represent coverage for multiple sub-regions resulting in duplication of service and redundancy.
- **Invalid Feeds:** GTFS data may contain errors and when passed through a validation procedure they are flagged as having erroneous data.

To the best of our knowledge, there does not exist a standard data description of detailed fixed-route ridership throughout Canada, a finding seen in other countries such as the United States [5] and throughout Europe [3]. This research proposes a systematic GTFS data discovery, collection, and analysis pipeline to study public transit offerings within Canada.

A systematic study of public transit offerings within and/or between regions is crucial for public transportation management since planning adaptations can be implemented to increase supply in high demand areas or adapt the inefficient deployment of transit services in low transit demand areas [6]. Additionally, the nation-wide collection of these data may then be used to evaluate a country's achievement of the United Nation's sustainability goals. Of particular relevance to this work is the sustainability development goal indicator 11.2.1, measuring population convenient access to public transit services [7], with "convenient" understood as "living within a 0.5 km (or 500m) walkable distance of the nearest stop" [7].

A recent study sought to leverage GTFS data for a network-wide assessment of locations where reactive treatments or proactive infrastructural changes may effectively improve tran-

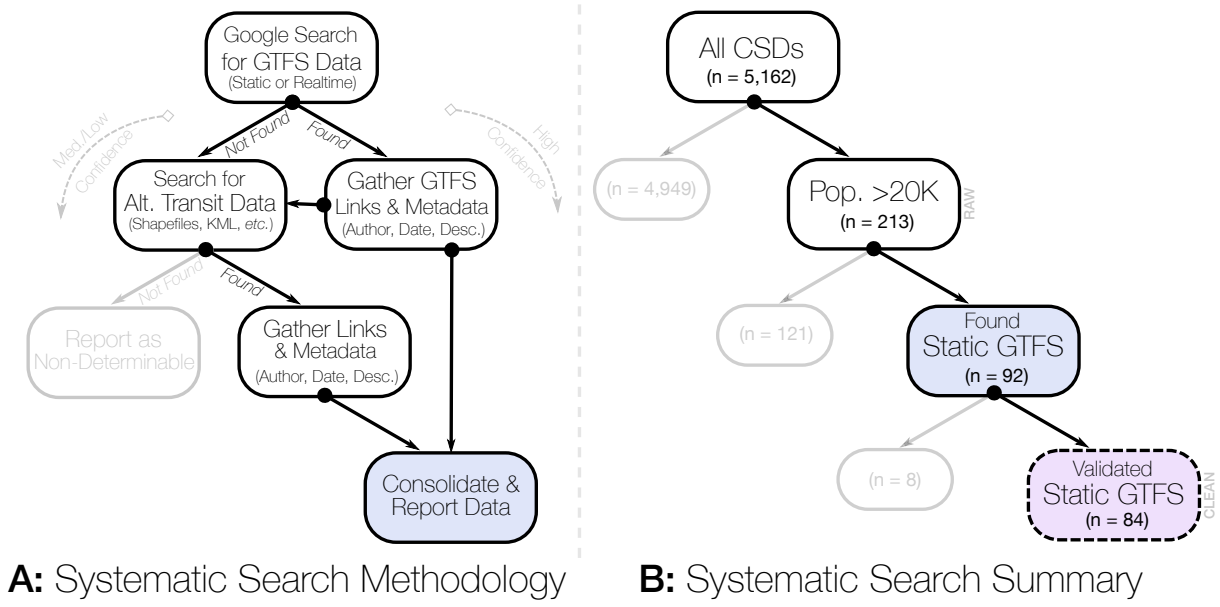

Fig. 1. Overview of the Canada-Wide GTFS Systematic Search Methodology and Summary of Results.

sit efficiency and reliability [8]. In a Canada-specific example, the GTFS data of Calgary, Alberta was used as a case study for the development of visualization tools displaying transit movement and for the measurement of transit system operation geographically and statistically [9].

Despite their broad utility in studying transit offerings at intra- and inter-regional levels, to the best of our knowledge, there does not exist a centralized repository collecting comprehensive and high-recency GTFS data for Canada. While crowd-sourced repositories such as TransitFeeds aggregates snapshots of these data, neither comprehensiveness nor recency are guaranteed. Consequently, in this work, we make the following three contributions:

1) provide a systematic approach to collect public transit data from various sources within a predefined jurisdiction
2) prepare a high-quality inventory of aggregated GTFS data (made publicly available)
3) analyze the public transit offerings lending to numerous insights between regions and cities

Using Canada as a case study, we collected GTFS for the 213 candidate census subdivisions ([CSDs] representing cities, towns, municipalities, *etc.*) with populations greater than 20,000 [10]. These data, when systematically collected throughout a country can be used to study the structure, organization, and availability of services nationally and between sub-regions. The codebase and all of the collected data are made freely available in a GitHub repository: github.com/chazingtheinfinite/canada-transit-study.

II. Data Collection & Methodology

Our work required the initial collection and filtration of all CSD population data [10] to only consider those regions with a population of at least 20,000 residents (n=213). We then systematically searched for publicly available links to GTFS and/or alternative public transit representations (*e.g.* Shapefiles, Keyhole Markup Language [KML] files, JPEG/PDF maps, *etc.*).

A. Systematic GTFS Data Collection Procedure

Six data collectors were tasked over two months (Feb. & Mar. 2022) to comprehensively search for relevant GTFS and alternative transit data for each of the $n = 213$ CSDs included within this study. Given the unstructured nature of hosted GTFS data, the collectors were asked to exercise good judgement while searching for all possible sources of data on public transit. This process included searching municipal Open Data portals, the serving public transit authorities websites, and crowd-sourced repositories. While the URL links to hosted GTFS data were our primary sought -after data, the collectors additionally acquired information on all additional relevant (meta)data, such as transit authority, authors, contact information, recency, *etc.*). The data that was collected came in a variety of formats such as GTFS, Real-Time GTFS (GTFS-RT), Comma-Separated Value files (CSV), KML files, Shapefiles (SHP), GeoJson files, and PDF maps. All public transit data were consolidated in a standard reporting document subdivided by CSD.

We illustrate our systematic data collection procedure and resultant dataset in Fig. 1. Our subsequent analysis considered only those CSDs for which the following criteria were met:

1) Official census subdivision data were provided by Statistics Canada (see [10]).
2) A link to *Static* GTFS .zip archive was found online through a *reasonable* search procedure (see Fig. 1A).
3) The discovered Static GTFS data were validated as matching the GTFS standard (see Fig. 1B).

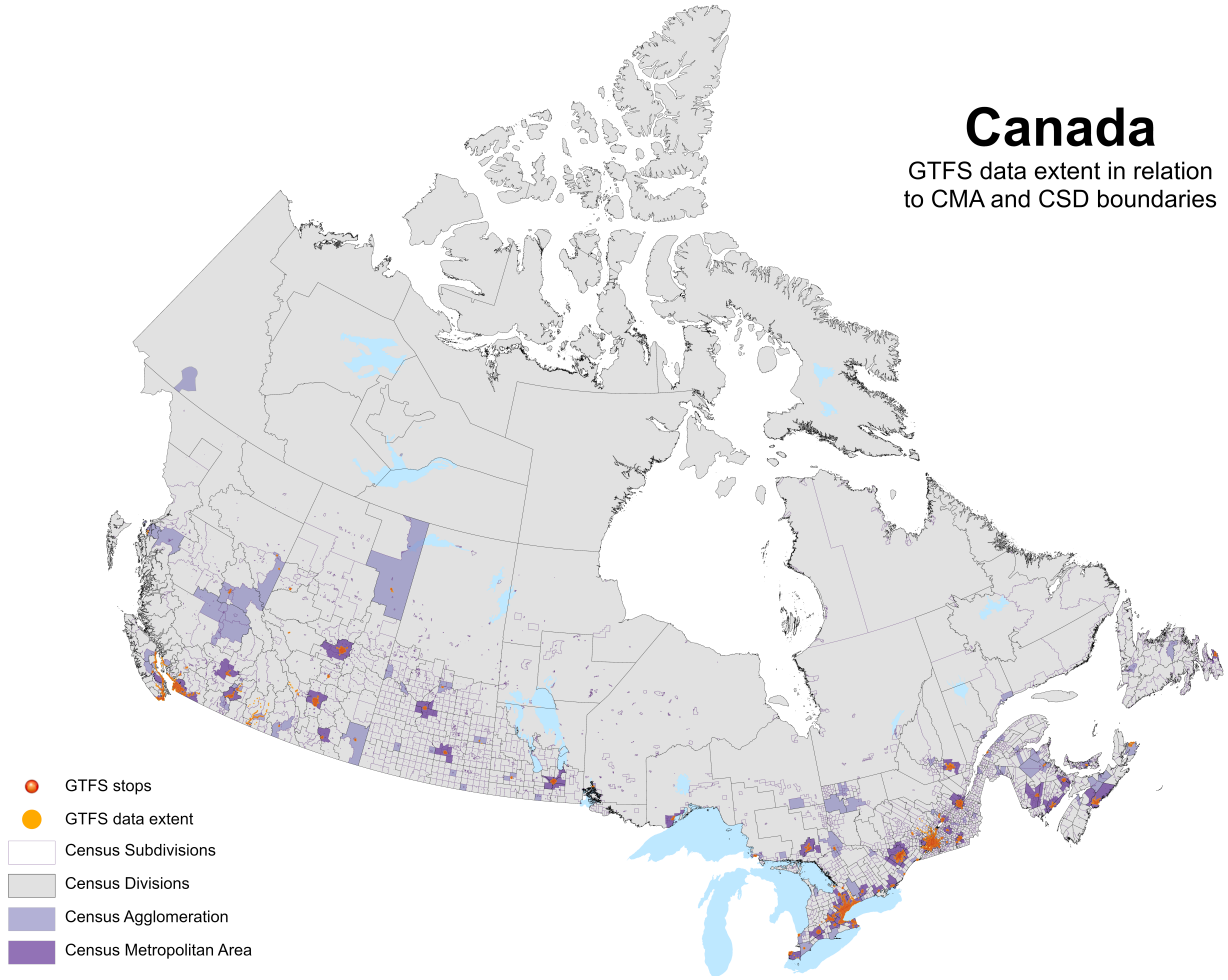

Fig. 2. Geospatial Representation of Public Transit Services throughout Canada and throughout Census (Sub)divisions.

We note that larger municipalities are typically served by multiple transit operators. For instance, Montréal is served by Société de transport de Montréal, Montreal Metro, and Exo. For the instances where multiple transit authorities provided independent static GTFS data, we appropriately merged the multiple feed statistics into a single comprehensive representation of the area.

B. Automatic GTFS Data Downloading & Preprocessing

After collecting GTFS metadata and the URLs, we sought to collect all of the discovered Static GTFS data archives using an automatic download script which would ensure that our work is both transparent and could easily be reproduced. Ethical data acquision practises were used as defined in [11]. Additionally, this script ensured that these data could be systematically formatted and transformed into a SQLight database which is the standard for interfacing with GTFS data using specific analysis libraries. Once collected these data could be geospatially plotted and visualized as in Fig. 2.

C. Integration of Census Population and GTFS Transit Data

Once all $n = 92$ GTFS archives were downloaded, through the SQLight processing step, each archive was passed through a GTFS validator that examined the contents of the archives to determine whether errors or violations of the standard were present. In $n = 8$ CSDs, the GTFS data were rejected resulting in a fully validated and final dataset comprising $n = 84$ Static GTFS archives (39.4% of all candidate CSDs). Figure 1B depicts the results from this systematic search summary and the stage-wise exclusions.

Finally, using the PyGTFS library, summary statistics were extracted for each CSD and merged with the official census subdivision data published by Statistics Canada [10]. These GTFS-derived statistics include measures such as the number of routes, the number of stops, the number of service days, as well as the coverage area (two-dimensional geographical height and width). From these metrics, we could also derive intuitive measures to express GTFS availability as a function of the service area and the number of stops therein.

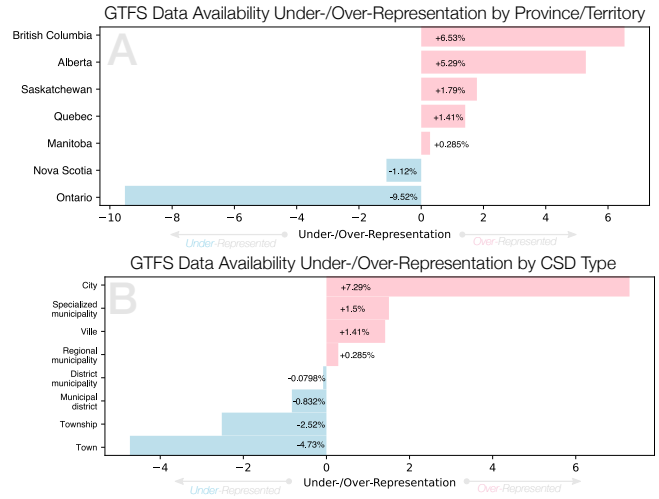

Fig. 3. GTFS Data Availability Under-/Over-Representation Analysis by Province/Territory and by CSD Type.

D. GTFS-Derived Measure of Stop Spatial Density

The metrics derived within this work to express the spatial availability of public transit services for the targeted census population are gross generalizations. That is, despite the specified latitude and longitudinal location data available from a given GTFS feed, the census data available within this study does not precisely localize the population within the service region. Ideally, high-resolution census data would be leveraged for all census-sampled resident to refine the population coverage area; such a study is left to future work with high-resolution census data. Consequently, to systematically compare public transit offerings between regions, generalized assumptions were incorporated into the metric calculations. For example, the *Stop Spatial Density* was computed in a way analagous to the simple weighted average proposed be Lamporte *et al.* [12] which assumes a uniform distribution of population over the serviced land area:

$$\text{Stop Spatial Density} = \frac{\text{number of stops}}{\text{GTFS Area}} \tag{1}$$

While relatively simple to implement, the resulting metric (as expected) tends to be inaccurate for sparse rural and/or non-uniformly populated areas. An analogous temporal metric of Static GTFS service offerings over the reported service window is possible, however this would require the temporal alignment of the available GTFS data (data recency is not necessarily guaranteed) as well as the extraction of route-specific trip frequencies to appropriately measure the temporal coverage over a particular service period. Such a study would be complimentary, however, it is left to future work.

E. Census Population Metrics Correlated with Static GTFS Metrics

Having integrated both census population data (for census years 2011, 2016, and 2021) with GTFS-derived public transit metrics, we generated a comprehensive correlation table (Table I) with both Pearson and Spearman correlations listed row-wise in pairs. Cell colouring in the table varies according to the magnitude of the correlation with magenta representing moderate-to-high positive correlation, grey representing neutral/no correlation, and blue representing moderate-to-high negative correlation. Statistically significant correlation coefficients are marked with asterisks. In addition, correlation scatterplots were created to visualize a subset of paired variables (Fig. 4).

F. Measuring Under- and Over-Representation of GTFS Data Availability

An important question to this work is whether or not the collected GTFS data is equally available across the geographical regions (*i.e.* provinces/territories) and CSD types (*i.e.* "city", "town", "ville", "municipal district", *etc.*) under study. To quantify whether the available Static GTFS data is under- or over-represented in a given region or CSD type, we determined the difference in the proportion of represented CSDs *with GTFS data* ($n = 84$) from the proportion of total candidate CSDs ($n = 213$). Thus, when grouping all CSDs by province/territory, an equal GTFS availability in a region would result in the proportion of municipalities with discoverable GTFS being equal to the proportion of candidate CSDs located in that region. This analysis is extensible to CSD-type (a classification determined within the census data). The provinces/territories and CSD-types for which no Static GTFS data were discoverable were left out of this analysis given that their under-representation is undefined (tends to $-\infty$). The findings are summarized in Fig. 3.

G. Evaluating Stop Spatial Density by Province/Territory & CSD-Type

The stop spatial density metric defined in this work is a heuristic measure of the availability of public transit availability in a given CSD controlling for the service area. By comparing how individual CSDs relate to the National Baseline (defined as the median stop spatial density of all CSDs), general and specific trends may emerge. To that end, we plot the rank-order distribution of CSDs (coloured by province/territory) in Fig. 5A, their grouped province-specific distributions as boxplots (in decreasing median stop spatial density) in Fig. 5B, and finally, their grouped CSD-type distributions as boxplots (also in decreasing median stop spatial density) in Fig. 5C.

H. Globally Situating Average Canadian Commute Times

Finally, to broaden the scope of our analysis and to situate metrics from Canada's public transit system among other countries globally, we extracted average commute times (measured in minutes) from 147 cities in 28 countries from the Moovit Public Transit Index [13]. Grouping cities by country and plotting each distribution in decreasing order by median commute time, we generate a comparison of average commute times between each country's cities and in relation to the global median (defined as the median commute times for all cities considered within this analysis).

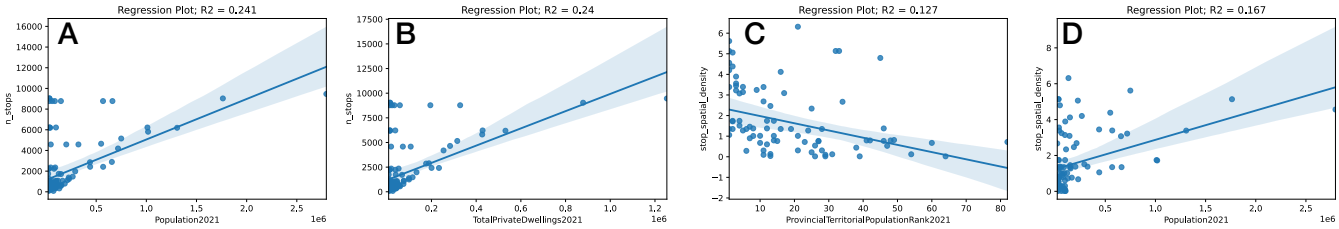

Fig. 4. Regression Plots between Population-Based Regional Census Measures and Transit Offerings.

III. Results & Discussion

This work sought to systematically collect all public transit data available within Canada to evaluate the availability of online Static GTFS data nationally. To the best of our knowledge, there does not exist a standard protocol for the dissemination of transit data online; public transit providers and local governments organize these data according to their own data management principles resulting in an unstructured and gap-filled representation at the national scale. This work represents a large-scaled effort to simultaneously collect Canada-wide public transit data and assess the challenges and limitations of navigating the existing Static GTFS publishing system. The data collected herein enables the critical assessment of Static GTFS data availability throughout Canada as well as assist in other analyses pertaining to public transit availability and network density, including relevant to achieving the UN SDGs. It is the hope of this work that better practises in the online publication of public transit data at the municipal, provincial/territorial, and federal levels will emerge. With progressively improved public transit data publishing additional insights into improving public transit availability and improved overall system efficiency can be discovered.

A. Current Static GTFS Offerings Correlate Moderately well with the Served Population

An anticipated finding of this work, when correlating population census metrics with Static GTFS service measurements, is a general moderate-to-high outcome (Table I). The correlation table highlights a number of expected trends such as the number of stops correlating significantly with CSD population for each of 2011, 2016, and 2021 ($p \leq 0.001$) and with a generally increasing Pearson coefficient ($2021 > 2016 > 2011$) given that current Static GTFS services are reflective of the needs of the current population (Fig. 4A). We note that population density per square kilometer (2016) is the highest positively correlated value with the number of stops ($0.5751; p \leq 0.001$), the number of trips ($0.5639; p \leq 0.001$), and the number of service days ($0.6468; p \leq 0.001$); the 2021 values are consistent with these findings in both magnitude and significance. Also, as expected, are the moderate-to-high negative correlations with national and provincial population rank (Fig. 4C). Since a low rank is, by definition, inversely proportional to population, negative correlations are expected. Interestingly, the national and provincial ranks have the largest Spearman coefficients among all census metrics with respect to the stop spatial density ($p \leq 0.001$). We considered both Pearson and Spearman correlation coefficients as a means of validating the trends observed throughout this analysis.

B. Discoverable Online Static GTFS Data Availability is Not Equal

A critical finding of this work reveals that the discoverable online Static GTFS data indeed under- or over-represents certain provinces/territories and CSD-types (Fig. 3). By comparing the set of CSDs for which Static GTFS data is available to the set of CSDs with a population over 20K, we can take the difference of these proportional representations to determine whether we observe a strong under- or over-representation once grouped. An equal distribution would find the same proportions indicating that GTFS availability is likely independent of province/territory or of CSD-type. As illustrated in Fig. 3, there are substantial differences when grouping by province/territory. At one extreme, British Columbia is determined to be greatly over-represented within this study by a factor of +6/53% (Fig. 3A) whereas, at the other extreme, Ontario is even more severely under-represented by a factor -9.52% (Fig. 3A).

A major contributing factor to British Columbia's over-representation is the ease of access to Static GTFS data through the BC Transit provider's online repository. By making easily available these standard data for all regions served by this provider, CSDs both highly populated and population-sparse are captured in our analysis. Conversely, Ontario representing a considerably large portion of the CSDs within our study does not have a standardized means for sharing Static GTFS data resulting in a sever under-representation within this study.

Similarly, when replicating this analysis for CSD types, we find that cities are greatly over-represented by a factor of +7.29% (Fig. 3B) whereas towns, at the other extreme, are under-represented by a factor -4.73% (Fig. 3B). These results are intuitive given that CSDs classified as a "City" will typically have a greater served population as well as a larger number of resources with which to organize public transit data. Much smaller CSDs, classified as towns, townships, or municipal districts will generally lack the necessary resources to make broadly available their public transit services in Static GTFS format.

TABLE I
CORRELATION TABLE BETWEEN ALL CENSUS SUBDIVISION POPULATION MEASUREMENTS AND GTFS TRANSIT OFFERINGS.

Census Value	Correlation Measure	Number of Routes	Number of Stops	Number of Trips	Number of Days	Stop Spatial Density
Population, 2021	*Pearson*	0.4028 ***	0.4912 ***	0.4037 ***	0.3687 ***	0.4084 ***
	Spearman	0.3326 **	0.3655 ***	0.3127 **	0.3833 ***	0.4487 ***
Population, 2016	*Pearson*	0.3977 ***	0.4877 ***	0.4057 ***	0.369 ***	0.4087 ***
	Spearman	0.3314 **	0.359 ***	0.3038 **	0.3774 ***	0.4545 ***
Population, 2011	*Pearson*	0.3914 ***	0.486 ***	0.411 ***	0.3741 ***	0.4075 ***
	Spearman	0.3324 **	0.3604 ***	0.3055 **	0.3767 ***	0.4439 ***
Population 2016-2021, % Change	*Pearson*	-0.1337	-0.1254	-0.0583	-0.0908	-0.0582
	Spearman	-0.0748	-0.0434	0.0191	-0.0596	-0.159
Total Private Dwellings, 2021	*Pearson*	0.4057 ***	0.4898 ***	0.4318 ***	0.3938 ***	0.4046 ***
	Spearman	0.3222 **	0.3518 **	0.2977 **	0.3794 ***	0.4381 ***
Total Private Dwellings, 2016	*Pearson*	0.4018 ***	0.488 ***	0.4347 ***	0.397 ***	0.4061 ***
	Spearman	0.3135 **	0.341 **	0.2856 **	0.3736 ***	0.4404 ***
Total Private Dwellings, % Change	*Pearson*	-0.0109	-0.0366	0.0028	-0.0674	-0.0981
	Spearman	0.1324	0.0864	0.1406	-0.0238	-0.1216
Private Dwellings Occupied by Usual Residents, 2021	*Pearson*	0.4063 ***	0.4902 ***	0.4291 ***	0.392 ***	0.4059 ***
	Spearman	0.3197 **	0.3505 **	0.2977 **	0.377 ***	0.4331 ***
Private Dwellings Occupied by Usual Residents, 2016	*Pearson*	0.4006 ***	0.4868 ***	0.4306 ***	0.3925 ***	0.4072 ***
	Spearman	0.3177 **	0.3455 **	0.289 **	0.3707 ***	0.4336 ***
Private Dwellings Occupied by Usual Residents, % Change	*Pearson*	-0.0502	-0.0688	-0.0266	-0.073	-0.1236
	Spearman	0.0491	0.0122	0.0647	-0.0639	-0.1576
Land Area In Square Kilometres, 2021	*Pearson*	-0.0222	-0.0693	-0.0611	-0.0592	-0.0573
	Spearman	0.2373 *	0.2174 *	0.146	0.1628	0.2759 *
Land Area In Square Kilometres, 2016	*Pearson*	-0.0233	-0.0703	-0.0615	-0.0597	-0.0573
	Spearman	0.1974	0.1788	0.1122	0.1265	0.2712 *
Population Density Per Square Kilometre, 2021	*Pearson*	0.3562 **	0.5104 ***	0.5497 ***	0.6002 ***	0.3294 **
	Spearman	0.3086 **	0.3925 ***	0.3564 **	0.4091 ***	0.3013 **
Population Density Per Square Kilometre, 2016	*Pearson*	0.417 ***	0.5751 ***	0.5639 ***	0.6468 ***	0.3022 **
	Spearman	0.3472 **	0.4292 ***	0.3904 ***	0.4451 ***	0.308 **
National Population Rank, 2021	*Pearson*	-0.2393 *	-0.2436 *	-0.08	-0.1199	-0.36 ***
	Spearman	-0.3326 **	-0.3655 ***	-0.3127 **	-0.3833 ***	-0.4487 ***
National Population Rank, 2016	*Pearson*	-0.2358 *	-0.2362 *	-0.0703	-0.1073	-0.3633 ***
	Spearman	-0.3314 **	-0.359 ***	-0.3038 **	-0.3774 ***	-0.4545 ***
Provincial Territorial Population Rank, 2021	*Pearson*	-0.3016 **	-0.2642 *	-0.0639	-0.1557	-0.3568 ***
	Spearman	-0.3391 **	-0.317 **	-0.2245 *	-0.4537 ***	-0.506 ***
Provincial Territorial Population Rank, 2016	*Pearson*	-0.3011 **	-0.2634 *	-0.0523	-0.1457	-0.3613 ***
	Spearman	-0.3451 **	-0.3234 **	-0.2307 *	-0.4593 ***	-0.5124 ***

In summary, a number of lessons may be learned from these findings. Most notably, the broad availability of Static GTFS data for those CSDs served by BC Transit suggests that with the definition of a standard data publication protocol in a centralized repository by an organization with the means to oversee that publication, these data can be made easily accessible for public consumption and/or enable subsequent downstream application/study. Conversely, without specific oversight at the regional or provincial level, the availability and management of these critical infrastructure data are left to highly localized transit authorities to dictate resulting in a piecemeal representation of these data. From Fig. 3B, there is a need to support CSDs classified as ”towns” or ”townships” to provide the requisite expertise to make their public transit services easily accessible online and enable, for example, subsequent studies into the efficiencies of their public transit offerings.

C. Stop Spatial Density Analysis May Reveal Potentially Lacking Public Transit Infrastructure

Leveraging the stop spatial density metric derived in this work for the analysis of Canadian public transit system services, we examined the general distribution of these values with respect to the provinces/territories and their CSD type. We define an arbitrary national baseline measure as the median stop spatial density among all CSDs considered in this work. In plotting the rank order distribution of this metric in Fig. 5A, we get an initial sense for which CSDs of particular provinces/territories rank in the top-half versus those ranked among the bottom-half, relative to the national baseline.

Extending this analysis, when grouping the CSDs by province/territory (Fig. 5B), we observe highly varied distributions. We note that, due to lack of discoverable data, provinces

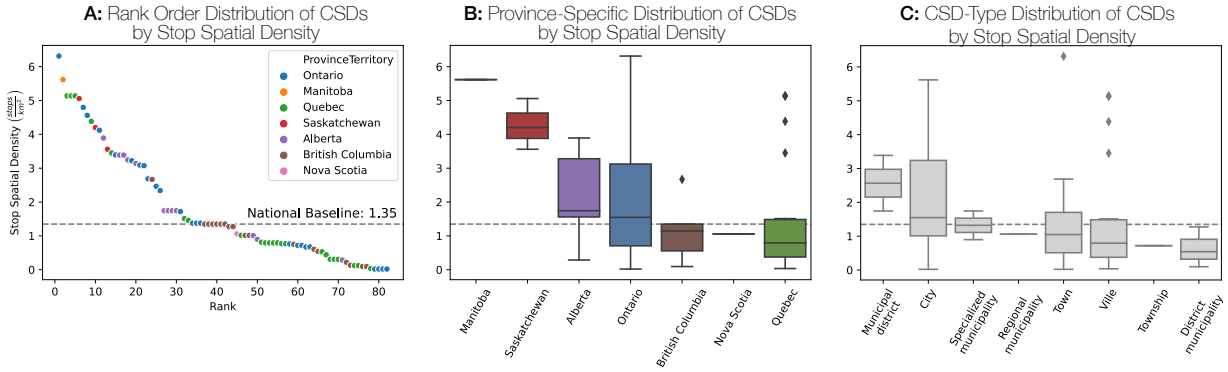

Fig. 5. Distribution Plots Depicting Stop Spatial Densities of CSDs for which Static GTFS Data are Available.

such as Manitoba and Nova Scotia have severely limited data from which any reasonable conclusion can be drawn. In the case of Manitoba, only Winnipeg is represented and with an ordering according to median stop spatial density, it evidently ranks first among these provinces. In reality, the distribution of stop spatial densities is more likely to resemble that of the the other Prairie provinces. Interestingly, even though British Columbia has been discovered to be over-represented within this study, it ranks relatively low compared to the other provinces and the national baseline. This may be due to the fact that BC Transit published considerably more Static GTFS data for smaller population CSDs in British Columbia effectively pulling its stop spatial density distribution towards lower values, which may effectively reflect a truer representation of general stop spatial densities within and across provinces. At the lowest end of the spectrum, Québec depicts the lowest median stop spatial densities with the majority of its represented CSDs falling below the national baseline. While the province was found to be somewhat over-represented in the dataset, this finding may point to a more fundamental lack of public transit infrastructure. The three outlier CSDs represent Montréal, Laval, and Québec City.

Finally, we extended this analysis to consider the stop spatial distributions when grouped by CSD type and obtained a much more nuanced outcome. Both ”municipal district” and ”city” are generally distributed above the national baseline, the remaining classifications generally distribute around the median except for ”district municipalities” which are consistently below the baseline.

While we define the national baseline as 1.35 stops per square kilometer (based upon the median value of all stop spatial densities in this analysis), at the spatial resolution considered within this study we do not have a measure for an *ideal* target stop spatial density.

Further analyses related to proximity and convenient access to means of transportation, including international comparisons, could reveal how Canada and different municipality types compare on this metric globally. A study on the proportion of the population living in relative proximity to a public transit stop can present an opportunity to determine candidate locations for new transit infrastructure as well evaluate regional alignment with a derived national target baseline.

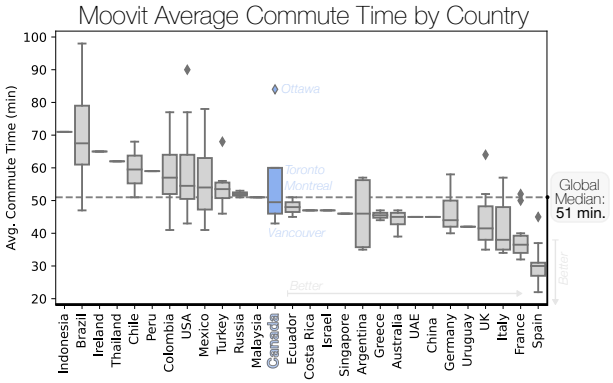

Fig. 6. Moovit’s Average Commute Time by Country. Data available under a CC BY 4.0 licence from the Moovit Public Transit Index [13].

D. Canada is Around the Global Median by Average Commute Time

Expanding the scope of our analysis to the global level, we sought to situate Canada among other countries based on average commute time as a metric. To this end, the Moovit Public Transit Index information was plotted as a boxplot in Fig. 6 where the average commute time of each city was grouped by country and sorted according to their median value. The global median is highlighted to depict how the distributions of countries vary around this value. Canada, with only four major cities represented, is situated approximately in the middle and, interestingly, outranks the two other North American countries, Mexico and the United States of America. At the lowest end, we note several European countries well below the global median (Fig. 6). Ottawa, as an outlier, appears to have a considerably longer average commute time, faring worse than the majority of cities worldwide. These data suggest a number of countries from which which public transit infrastructure systems might be modelled to improve, on average, the daily commute time.

IV. Conclusion & Future Directions

Public transit is a critical infrastructure system of great importance to the general public. In this work, we provide a systematic approach to collect public transit data from various sources, prepare a high-quality inventory of aggregated GTFS data, and analyze the public transit offerings lending to numerous insights and at various geographic scales. We sought to collect GTFS data for the 213 candidate Canadian municipalities (CSDs) and derived a number of findings on public transit provision nationally and on potential approaches towards making GTFS data broadly accessible for downstream applications. This work serves as a model for the large-scale collection, aggregation, and analysis of public transit data. The Static GTFS data collected in this work can contribute to a variety of future projects, including the calculation of Canada's achievement of the United Nation's sustainability development goals related to proximity and convenience access to means of transportation. Our aggregated dataset and open-sourced codebase are available at: github.com/chazingtheinfinite/canada-transit-study.

Acknowledgment

The authors would like to thank Dr. Alexandra Bozheva and Yann Pelcat of Statistics Canada for their guidance throughout this work. The authors would also like to thank Michael Nwokobia, Mingqi Feng, and Blessing Okeme for their assistance in collecting portions of the data used in this work.

References

[1] G. Desaulniers and M. D. Hickman, "Public transit," *Handbooks in operations research and management science*, vol. 14, pp. 69–127, 2007.

[2] D. Banister, "The sustainable mobility paradigm," *Transport policy*, vol. 15, no. 2, pp. 73–80, 2008.

[3] R. Kujala, C. Weckström, R. K. Darst, M. N. Mladenović, and J. Saramäki, "A collection of public transport network data sets for 25 cities," *Scientific data*, vol. 5, no. 1, pp. 1–14, 2018.

[4] S. K. Fayyaz S, X. C. Liu, and G. Zhang, "An efficient general transit feed specification (gtfs) enabled algorithm for dynamic transit accessibility analysis," *PloS one*, vol. 12, no. 10, p. e0185333, 2017.

[5] P. Carleton, S. Hoover, B. Fields, M. Barnes, and J. D. Porter, "Gtfs-ride: Unifying standard for fixed-route ridership data," *Transportation Research Record*, vol. 2673, no. 12, pp. 173–181, 2019.

[6] K. Kaeoruean, S. Phithakkitnukoon, M. G. Demissie, L. Kattan, and C. Ratti, "Analysis of demand–supply gaps in public transit systems based on census and gtfs data: a case study of calgary, canada," *Public Transport*, vol. 12, no. 3, pp. 483–516, 2020.

[7] R. Global Urban Observatory Unit and C. D. Branch, "Metadata on SDGs Indicator 11.2.1 Indicator category: Tier II," *United Nations Habitat*, 2018. [Online]. Available: https://unhabitat.org/sites/default/files/2020/06/metadata_on_sdg_indicator_11.2.1.pdf

[8] Z. Aemmer, A. Ranjbari, and D. MacKenzie, "Measurement and classification of transit delays using gtfs-rt data," *Public Transport*, pp. 1–23, 2022.

[9] P. Prommaharaj, S. Phithakkitnukoon, M. G. Demissie, L. Kattan, and C. Ratti, "Visualizing public transit system operation with gtfs data: A case study of calgary, canada," *Heliyon*, vol. 6, no. 4, p. e03729, 2020.

[10] Statistics Canada, "Population and Dwelling Counts: Canada and Census Subdivisions (Municipalities)," 2022, data retrieved from Statistics Canada https://doi.org/10.25318/9810000201-eng.

[11] A. Luscombe, K. Dick, and K. Walby, "Algorithmic thinking in the public interest: navigating technical, legal, and ethical hurdles to web scraping in the social sciences," *Quality & Quantity*, vol. 56, no. 3, pp. 1023–1044, 2022.

[12] G. Laporte, J. A. Mesa, and F. A. Ortega, "Locating stations on rapid transit lines," *Computers & Operations Research*, vol. 29, no. 6, pp. 741–759, 2002.

[13] Moovit, "Compare average commute times — Moovit Public Transit Index." [Online]. Available: https://moovitapp.com/insights/en/Moovit_Insights_Public_Transit_Index-commute-time