Transit Network Complexity in the Context of Transit Itinerary Inference with Travel Survey Data and GTFS

Marshall V. Davey

A Thesis

In

The Department

Of

Geography, Planning, and Environment

Presented in Partial Fulfillment of the Requirements for the Degree of Master of Science (Geography, Urban, and Environmental Studies) at Concordia University Montreal, Quebec, Canada

March 2019

CONCORDIA UNIVERSITY
School of Graduate Studies

This is to certify that the thesis prepared

By: Marshall Vincent Davey

Entitled: Transit Network Complexity in the Context of Transit Itinerary Inference with Travel Survey Data and GTFS

and submitted in partial fulfillment of the requirements for the degree of

Master of Science (Geography, Urban and Environmental Studies)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

______________________________________ Chair
Pierre Gauthier

______________________________________ Internal Examiner
Sebastien Caquard

______________________________________ External Examiner
Martin Trepanier

______________________________________ Supervisor
Zachary Patterson

Approved by ______________________________________
Chair of Department or Graduate Program Director

____
Dean of Faculty

Date ____________________

Abstract

Transit Network Complexity in the Context of Transit Itinerary Inference
with Travel Survey Data and GTFS

Marshall Davey

Researchers and planners have taken great interest in the rich data-resource that smartphone and GPS travel surveys can now produce. The interpretation of this data has become a popular topic with methods such as transit itinerary inference (TII) from travel survey data and GTFS emerging as useful tools in the field of travel behavior analysis. This exploratory research develops metrics to quantify a characteristic of GTFS data that complicates the overlay processing of travel survey GPS points and bus route geometries in TII: the spatiotemporal overlap of bus routes in the GTFS record. Accurate route inference is difficult in regions where rider data coincides with overlapping routes and various TII approaches have been tested to address this challenge. In this research, detecting overlap, and quantifying the degree of overlap on road links is achieved in 5 study regions through the application of two proposed measures: The Overlapping Routes on Links (OROL) index, and the Probability of Passage (POP) score. The latter's output is seen as one way to improve route matching rates in TII. These measures build off the traditional Line Overlapping Index (LOI) and improve upon it by providing previously unobtainable road-link level detail; the OROL index, in fact, represents a spatially precise decomposition of the LOI. To ensure accurate analysis between networks, an additional novel procedure is developed that converts GTFS data into a simplified stand-in road network representation, thus providing a base layer for disaggregate network measures, and replacing the need for additional road network sources entirely.

Acknowledgements

This thesis is dedicated to my family and loved ones for their continuing support throughout the journey of my higher education. This thesis was made possible thanks to the tutelage and patience of Zachary Patterson, my colleagues in the trip lab, Ali and Kyle, as well as the incredible community of Concordia's G.P.E. department.

Table of Contents

List of Tables and Figures

List of Acronyms

AVL – Automatic Vehicle Location data

BRT – Bus Rapid Transit

CSV– Comma-Separated Values file format

CT – Calgary Transit

ETS – Edmonton Transit Service

GeoJSON – Geographical JavaScript Object Notation

GIS – Geographic Information System

GPS – Global Positioning System

GTFS – General Transit Feed Specification

GUI – Graphical User Interface

LOI – Line-Overlapping index

OD - Origin Destination

OROL – Overlapping routes on links

POP – Probability of Passage

SQL – Structured Query Language

STM – Société de Transport de Montréal

TTC – Toronto Transit Commission

1 Introduction

Observing how riders move about in a transit system is a crucial first step for making any informed decisions about expanding or changing service. Given the great expense of transit operations and the long-term scope of their operations, planners need to implement the best possible solutions for the specific needs of the citizens on a limited budget. Maximizing the benefit of new routing and infrastructure requires not only knowledge of how the existing network is being used, but also projections of how alterations or extensions to it will impact the usage of the system and the community at large. The planning of public transit systems in particular has the power to empower or hinder communities with regards to access to employment, education, food and health services, and social and cultural events.

Once transit systems are in place, methods of observing and measuring their effectiveness are of the utmost importance for tailoring the service and correcting any oversights. Historically, network usage statistics are gathered via rider counting and phone interviews which are then used to create Origin-Destination (OD) cost matrices. OD cost matrices are used to calculate the least cost path between points and are an indispensable tool for planning the layout of transit routes. Unfortunately, the high cost of conducting rider surveys and processing the data leaves transit agencies with little choice than to conduct the studies at intervals several years apart. Luckily, advances in GPS smartphone technologies and the ubiquity of smartphones in North America is opening new avenues in the study of transportation systems.

Transportation planning is a field that has benefited greatly from advances in mobile, and GPS technologies. In particular, the collection and processing of data from smartphone travel surveys is a topic that is gaining traction in both private and public sectors (Nitsche, Widhalm, Breuss, & Maurer, 2012; Shen & Stopher, 2014; Zahabi, Ajzachi, & Patterson, 2017). When combined with GIS technologies, this type of survey data can provide detailed descriptions of how users move in a transit network, and even allow for the inference of respondent transit itineraries (Zahabi et al., 2017).

In addition to new sources of transportation demand side data, supply side data provided by public transportation agencies is also facilitating research efforts. The now popular General Transit Feed Specification (GTFS) is a data format transit agencies use to publish their routing and scheduling information and has become a *de facto* standard as more and more agencies choose to make the data public via open-data portals online (Hadas, 2013). Combining this freely available GTFS data with rider GPS data provides researchers and planners with the data they need to not only observe respondent transit itineraries, but to do so with a previously impossible level of precision.

Thanks to these developing technologies and data sources, new ways of interpreting and inferring information from the collected data are also being developed. The observation and inference of several trip characteristics from GPS data is being explored and increasingly reliable methods for inferring details such as origin and destination, travel mode, vehicle type, direction, and even trip purpose are emerging in the literature (*6*, *21–25*, *28*). Recent work in this area has begun to examine methods for inferring the transit itineraries of travelers. As such, "transit itinerary inference" (TII) aims to completely describe each segment of a rider's trip using spatiotemporal data collected from a variety of sources (Thiagarajan et al., 2010; Zahabi et al., 2017).

Transit itinerary inference has begun to be explored relatively recently, primarily with the aid of vehicle location technologies and with transit fare-card (smartcard) data (Gordon, Koutsopoulos, Wilson, & Attanucci, 2013; Nassir, Khani, Lee, Noh, & Hickman, 2011). More recently, methods to infer transit itinerary combining smartphone travel survey data and GTFS have emerged (1). One such approach is the TII algorithm developed by Zahabi et al. in which an iterative process examines GPS points and their coincidence with GTFS records, direction of travel, and even instances where GPS signals are not present in order to infer a rider's full itinerary (Zahabi et al., 2017). The inference of this information is accomplished over several steps: first, transit trips are extracted from the smartphone travel survey data by filtering GPS data to reveal patterns of mobility vs. immobility. Next, once a series of GPS points are identified as a trip and grouped together, these trips are overlaid with GTFS scheduling and routing data to determine a collection of candidate transit routes whose path in the road network matches that of the rider. A trip breaker algorithm then breaks the trips into sub-segments according to where the collections of candidate routes change. Afterwards, each sub-segment is compared to each other to determine which routes are common to all sub-segments. Finally, it is only when a route belonging to all the matches also coincides with a route belonging to the boarding stop that the algorithm assigns that route to that trip.

This GIS-based, algorithmic approach was able to reliably infer transit route details for 87% of the distance travelled by transit in the pilot study region of Montreal. The remaining 13% of transit distance occurred in areas with route overlap and represents "route ambiguity" within the network (Zahabi et al., 2017) .

Figure 1 provides an example of transit route ambiguity. The figure shows a schematic representation of an intersection (avenue du Mont-Royal/avenue du Parc in Montreal, Canada), along with three transit lines (shown in color), and the GPS points that represent travel survey data. All three of the transit lines travel northward along Parc. North of Parc, two of the transit lines continue north, while the other (route 129) deviates to travel along another street (Cote Ste. Catherine). The transit network is ambiguous for all the GPS points in this example. Transit Itinerary Inference work amounts to reducing network ambiguity by controlling for time of day as well as by following the location data of a user across time (Thiagarajan et al., 2010; Zahabi et al., 2017). Although transit itinerary is inherently a function of transit route ambiguity, there are no adequate indicators able to measure this characteristic for transit networks. This thesis proposes two methods that build on the traditional Line Overlapping Index (described below) to better describe the degree to which networks are characterized by overlap, or ambiguity.

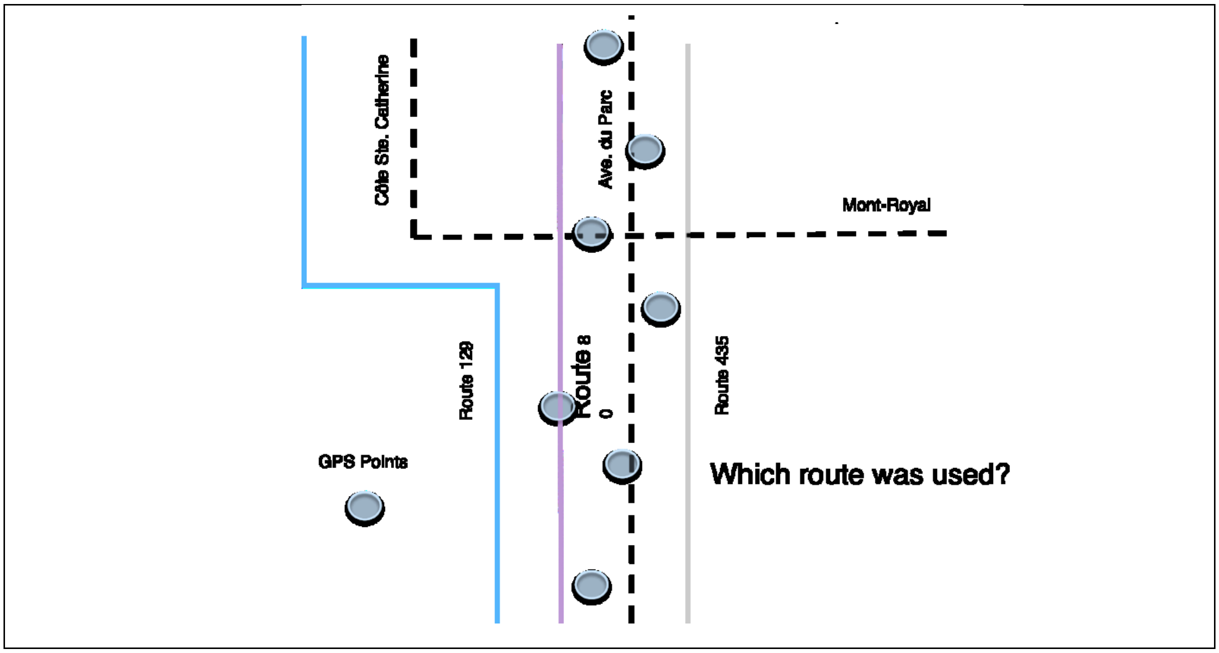


Figure 1: Transit network ambiguity

A different route inference method, developed by Carrel et al., begins with a similar GIS approach to identify candidate routes, but then employs an undirected tree, graph-theory approach to match an entire trip vector to one common route (Carrel, Lau, Mishalani, Sengupta, & Walker, 2015). This procedure forgoes the segmenting approach described in the previous method that could be said to contribute to the issue of route ambiguity (since 'ambiguity' results when a route number match cannot be found amongst all trip sub-segments) and yet the authors acknowledge that route overlap as a hindering factor in their procedure. This method examines GPS rider data gathered from survey devices, the GTFS record, and Active Vehicle Location (AVL) data to construct a 3-dimensional (latitude, longitude, time) search box around the rider's trip points. From this, all transit routes that pass through the box are recorded. Next, Dynamic Time Warping is used to calculate similarity between the GPS trace and route locations, and finally a distance threshold is applied to select candidate routes. Via these methods the team was able to correctly infer 93% of the 103 sample transit trips recorded in their survey of San Francisco (Carrel et al., 2015).

While the authors do not explore the cause of the remaining 7% of undescribed trips, they mention the occurrence of multiple route matches hindering the route inference process. In this study the authors match rider trajectories to AVL data and only use GTFS to verify if busses were running on time. I believe the fact that this procedure examines AVL data (which provides the actual location of vehicles during service runs), and yet still runs into problems of overlapping routes underscores that this challenge is inherent to transit itinerary inference processes regardless of whether the static or real-time GTFS records are used. Even while relying on Active Vehicle Location data bus routes will still converge for transfers as well as at termini at given times in their schedules, effectively increasing the number of candidate route matches.

The motivation for this research is to address the issue of route overlap by developing measures that can locate and quantify the occurrence of route overlap in GTFS datasets that meet the minimum condition of having a geographically faithful shapes.txt file. The networks examined in this study belong to the following Canadian cities, presented in descending order of metropolitan population: Toronto, Montreal, Vancouver, Calgary, and Edmonton.

While the TII developed by Zahabi et al. performed relatively well in the pilot study region of Montreal, it is an open question as to whether it would perform as well in other cities. In the absence of validated trip data from the other study regions to test the relationship between route attribution accuracy and network shape, this research will develop a ranking of the study regions according to how well a TII route matching procedure is expected to work in each network.

In order to develop the ranking of cities, two existing metrics will be employed alongside the measures developed in this research. Of the two existing metrics, only one of them is calculated using spatial data, while the other relies solely on timetable information. The two extant measures employed in this study are: 1) the Active Routes count: tabulates and reports the number of transit routes active for each hour of the day; 2) the Line-Overlapping Index, which provides one dimensionless score for the whole network representing the degree to which routes converge onto common paths. And finally, the two novel metrics proposed in this research are the Overlapping Routes on Links measure (OROL) that examines each road link with active routes and produces a total route overlap count for each link, and the Probability of Passage score (POP) that expands upon the OROL measure by also tabulating the departures of each route present on a link and calculating a departure ratio for each route (the POP score). The spatial layers that result from the OROL and POP methodologies are then used to measure network statistics such as total length of overlapping routes, the ratio of overlapping routes to total network distance (OROL %), as well as portion lengths of overlap once the road links are filtered using POP scores (more on this in the methodology section).

While the OROL methodology was developed with a different goal in mind than the LOI, it was interesting to discover during this research that the OROL methodology offers a new pathway to producing LOI scores, doing so in a spatially disaggregate manner thanks to the availability of spatially disaggregate data and GIS. In effect, the OROL calculation presents a methodology to *decompose* the single LOI value measured for a network into degree of overlap categories. Each overlap category can then be located and measured in the network to produce a more detailed accounting of overlap.

In order to calculate OROL and POP scores, a road-network shapefile is required so that each road link can be examined for the correspondence of bus routes. Initial tests of this process raised concerns over the accuracy of length measures derived from publicly available road files. After encountering files from different sources with varying topographical rules, or worse yet, files from OpenStreetMap that have sidewalks and bike paths coded as roads, the idea of generating road-networks directly from the GTFS itself was proposed. The use of the GTFS data to create a “road-free” network will be described in the methodology section of this thesis as it was a crucial development in the design of these measures. The GTFS-to-roads section further below also includes more detailed examples of network encoding configurations that lead to the mis-attribution of routes and erroneous length calculations (see Figure 8 page 36).

Thanks to this novel GTFS-to-roads approach, a quasi-road network file with consistent topological rules can be generated quickly for each city in the study group. In addition to consistent topology, these resultant layers also make GIS processing more efficient by eliminating any road links from the analysis that do not have active routes for a given test period.

As such, this paper contributes to the current literature by developing two spatially fine (link by link) measures of transit overlap that are comparable across transit networks and can be derived exclusively with the use of GTFS data. The link by link nature of these measures and the availability of the resultant length values are improvements over the classic Line Overlapping index that provides a single dimensionless score for the entire network (Derrible & Kennedy, 2011; Musso & Vuchic, 1988). Likewise, the spatially disaggregate GIS layers that are produced throughout the procedures are useful for encoding any link-level statistic encountered while analyzing different types of networks and represent a contribution this research brings to the field of network analysis.

What follows in the literature review is a re-cap of the current state of overlap analysis in the context of network complexity as well as a collection of classical transit indicators that can be calculated using only the GTFS dataset of a transit system.

2 Literature Review

With the increased accuracy of Transit Itinerary Inference procedures established as the motivating factor behind this study, the literature review will now turn to the development of transit measures.

Developing measures that quantify and locate the overlap of transit routes in static GTFS records is the primary goal of this research, and as such, experiments testing route matching rates in TII processes fall outside the scope of this research. With this in mind, the research focuses primarily on the development of measures, the development of a GIS network spatial disaggregation methodology, and the testing of the accuracy of each developed measure's outputs. To situate this work in the growing body of literature, the following review focuses on the measure of transit networks, the development of these measures, and how modern data-sets facilitate and even motivate such efforts.

The research presented in this paper contributes to a growing body of literature that examines useful ways in which GTFS data can be leveraged in the domain of transit network analysis. What follows below is a brief history of the development of transit indicators punctuated by the arrival of GTFS in 2005. From there, the review will cover how GTFS has been used in conjunction with these older metrics, as well as how it permits for the relatively easy calculation of Network and Graph theory approaches to network measures. Finally, the review concludes with current measures of network overlap that pertain to transit itinerary inference with a special focus on the way each expresses route ambiguity.

2.1 Classic transit indicators

Transit Level of Service (LOS) indicators have long been a topic of interest to city and transit planners alike. Given that many transit agencies are publicly funded, the ability to accurately measure and report the functioning of a transit system is key in order to ensure that public funds are used efficiently. Historically, the development of such indicators has typically been driven by governmental or industry bodies with some of the first indicators ever proposed resulting from studies commissioned by governmental bodies like the Pennsylvania Department of Transport in 1973, and even earlier than that, the Public Administration Service in 1958, based in Chicago (Allen & DiCesare, 1976).

There exist different classes of indicators; some are designed to be used by administrators making budgetary decisions, others by transportation engineers, and others still that are intended to inform riders about the agencies offerings (Fielding, Glauthier, & Lave, 1978). There are efficiency indicators, and effectiveness indicators, as well as Level of Service indicators and even Transit Hygiene indicators – this final type is perhaps the most conceptual of the group as it pertains to the level of satisfaction of using the service (Alter, 1976). Since a user's level of satisfaction with the service is a subjective matter, the factors impacting Transit Hygiene are typically measured quantitatively and don't lend themselves to being tabulated from tables or maps. Such indicators are important to transit planners for understanding why ridership may be low when the LOS and financial indicators all report that the system is running efficiently and affordably (Alter, 1976).

Some of the most popular and pertinent transit indicators stem from a seminal work titled "Evaluation of Public Transit Services: The Level-of-Service Concept" authored by Colin H Alter et al. in 1976. The paper proposed a set of basic indicators designed to give governmental administrators the information they needed to properly manage public transit services (Alter, 1976). Table 1, shows some of the basic indicators proposed in their paper:

Introduced by / Source	**Examines**	**Indicator**	**Calculation**
Transit Service Evaluation: Preliminary Identification of Variables Characterizing Levels of Service, William G Allen, et al. 1976	Routes	Route Density	route-km / square km
		Route Distribution	vehicle km/ service area pop
		Route Coverage (area)	route km * 0.4km / square km
		Route Coverage (population)	route km * 0.4km / population
		vehicle use (distance)	daily vehicle km / scheduled # of vehicles
		vehicle use (time)	daily vehicle hours / scheduled # of vehicles
	Frequency	Headway	average time between busses
	Capacity	Vehicle seat capacity	population / total seats
		route capacity	max # of passengers / hour
	Non-user related	Route Congestion	# of busses on any street segment / hour

Table 1: Early transit indicators

It is interesting to note that 7 out of the 10 listed measures can be calculated with GTFS even though these measures were proposed 29 years before the GTFS format was created. This helps demonstrate how the network information required to provide accurate schedule planning and location services is the same type of data required to calculate various network statistics. We cannot say that the data stored in GTFS was created as such to facilitate the calculation of these indices, but rather the data required by both researchers and itinerary systems are one and the same.

2.2 GTFS

The General Transit Feed Specification (formerly known as the Google Transit Feed Specification) is a standardized data-format that transit agencies employ to disseminate their time tables and routing information to scheduling and itinerary platforms such as Google Maps (Antrim et al., 2013). The datasets are published in open data portals hosted by transit agencies and the classic GTFS format is referred to as a *static* record. This is in contrast to GTFS-r which is a real-time reporting service that can track vehicle locations through online Application Programming Interfaces (web API). Contained in a collection of .csv files, a static GTFS dataset contains the location of all the agency's stops and routes, stored as latitude and longitude points, the complete timetables of all lines, and additional information for trips such as wheelchair accessibility or special service schedule exclusions.

Prior to digital GTFS datasets the primary means of disseminating this information was via printed maps and schedules, rendering compiling and analyzing this information an arduous task. GTFS presents increased accessibility to this data as the new format encompasses all of the required data in the common .csv format. In addition to saving the time of manually compiling schedules, the task of calculating the spatial characteristics and relationships of the network is also easily accomplished thanks to Geographic Information Systems (GIS) software.

Historically, the calculation of even the simplest network characteristics such as route lengths, coverage areas, and stop densities presented a labor-intensive process. The field of transit network analysis, and the creation of indices to conduct the analysis, has been opened wide to researchers thanks to this new concise source of network information.

Much of the current research applying GTFS to network analysis has examined how to recreate classic indicators used since the beginning of transit analysis, and others still rely on outside sources of information such as GIS road network files and census level demographic data (Antrim et al., 2013).

It is expected that different indicators require different data, some of which is more readily available than others; therefore, it follows that the data available to researchers is often a major consideration in choosing an indicator for the purpose of a study. It is in this regard that

GTFS data represents a great source of information for researchers. GTFS is especially handy to researchers since the data comes packaged in a common scale and resolution, and was compiled under one common context (Hadas, 2013). This is in contrast to data gathered via traditional sources that often vary in many characteristics. The fact that most transit agencies using GTFS make the data publicly available also contributes greatly to its utility for researchers.

The GTFS data format originated in 2005 due to a joint effort between the TriMet transit agency of Portland, OR. and Google to bring transit planning software to the citizens via Google Maps platform (J. C. Wong, 2013) ; as such, the data contained in the files is only intended to drive transit itinerary software. This limits the possible types of analysis to service level evaluation and topographical analysis of the network (Catala, Downing, & Hayward, 2011). Since the data does not reflect any rider or performance levels of the system, efficiency and hygiene indicators are not possible. Even though GTFS was not specifically intended to be a data source for researchers and planners, the pursuit of GTFS transit indices has been a popular topic. As with the early transit indicators, governmental and industry bodies remain a strong driving force in the development of GTFS indicators. What follows is a list of measures possible with GTFS datasets proposed by the National Center for Transit Research, FL.:

Source	Type	metric
National Center for Transit Research, University of South Florida for the Florida Department of Transportation	Service Evaluation metrics	service area
		service coverage
		time and distance calculations
		route and service directness
		stop location and spacing optimization
		service frequency
		span of service

Table 2: Classic metrics that can be calculated with GTFS

The Transportation Research Board (a branch of the National Academy of Science, Engineering, and Medicine) has outlined several useful metrics in their Transit Capacity and Quality of Service Manual. The manual is produced with the support of the U.S. Department of Transportation and is intended as a guideline of best practices for planners. What follows is a brief description of some of the indicators proposed in the manual authored by the Transportation Research Board that only need GTFS data to be calculated (Table 3 adjacent) (Group, 2013).

While the manual makes no specific mention of GTFS in regards to these metrics, they can all be calculated using the information stored in a GTFS dataset.

Source	Metric	Measure	Note
Transit Capacity and Quality of Service Manual	Daily Average Headway	measures the time between arrivals at each stop for each line	
	Route Length and Stop Density	looks at the relationship between average route length and number of stops	especially helpful in identifying routes with unique attributes such as exceptionally long stretches representative of commuter lines, or exceptionally dense lines typical of urban cores
	Hours of Service	looks at scheduling and routing to provide availability measures at the route segment and corridor resolution	

Table 3: Capacity and Quality of Service Indicators form the Transit Capacity and Quality of Service Manual

2.3 Network and Graph Approaches to Network Quantification

While Network and Graph theory approaches have lent their utility to transportation network analysis since the 1950s (Allen & DiCesare, 1976), the relatively recent advent of the GTFS data format and GIS software has done much to advance new methods in the domain of transit network design and analysis (J. Wong, 2013).

Network and Graph theory have been applied to a host of indicators in the context of quantifying and describing transit networks. These approaches are best suited to travel time and accessibility indicators. Previous work by Yuval Hadas in 2013 applied graph methods to GTFS data in a paper in which they measured four network indices. Their paper outlined connectivity, coverage, accessibility indicators as well as a route overlap indicator, all of which are listed in the table below (Hadas, 2013).

One of the most important contributions to Graph theory applications to transit network analysis comes from Vuchic and Musso who, in their 1988 and 1991 publications, outlined a collection of transit metrics that are suited to graph analysis in the context of transit networks. Divided into network *size and form* as well as network *topology* indicators, some of these measures have been carried forward into recent work that uses GTFS data (Derrible & Kennedy, 2011; Musso & Vuchic, 1988). The indicators that can be computed using GTFS data have been included in the table below.

Recent work from Polytechnique Montréal, Quebec also took a graph-theoretical approach to developing transit metrics that use GTFS data. In their work, Fortin et al. calculate indicators at the stop and route level of analysis (listed below). The authors decided to omit network level analysis from their study since in the graph theory approach the network level results would be a summary of the route and stop level results (Fortin, Morency, & Trépanier, 2016).

Source	Metric	Measure	Note
Innovative GTFS Data Application for Transit Network Analysis Using a Graph-Oriented Method (Fortin, P. 2016)	Active pairs of stops	Connectivity	Stop level analysis
	Extent of stop service	Connectivity and frequency	Stop level analysis
	Service speed	Average speed on each road link	Route level analysis, helpful for revealing changes in service at different times of the day due to hypothetical disturbances to the road network
Vuchic and Musso (1988,1991,2005)	Number of stations on line		
	Number of inter-station spacings on line		
	Length of line		
	Number of transfer stations		Used to express connectivity

Vuchic and Musso Continued (1988,1991,2005) Summarized in Applications of Graph Theory and Network Science to Transit Network Design (Derrible, S. et al. 2011)	Number of lines in network	Network size and form	
	Number of stations in network		
	Number of inter-station spacings		
	Route length of network		
	Number of Circles		Also known as cyclicity
	Number of station to station travel paths	Network size and form	
	Average inter-station spacing	Network topology	
	Line overlapping index		Provides one score for entire network
	Network complexity		
	Network connectivity		
	Directness of Service		
Assessing Public transport systems connectivity based on Google Transit data (Yuval Hadas, 2013)	Coverage level indicator		
	Network speed indicator		
	Intersection coverage indicator		Same as coverage level indicator but emphasizes flow at the intersection level allowing for node analysis
	Stop transfer potential		
	Route Overlap Indicator	Measures the overlapping extent of pairs of routes	Used to measure efficiency of transfers

Table 4: Graph and Network Theory Indicators Possible with GTFS

2.4 Measuring Network Overlap

It is important to note that three principal "overlap" indicators exist in the context of network and transit analysis. The Line Overlapping index (Derrible & Kennedy, 2011; Musso & Vuchic, 1988), a "transit overlap" measure (actually a sub-component of a Robustness indicator) (Liao & van Wee, 2017), and the Route Overlap indicator (Hadas, 2013). What follows below is a brief description of each measure and their applicability to the problem of route overlap in the GTFS record.

Of the three established indices, the Line Overlapping Index (LOI) introduced by V. Vuchic and R. Musso provides an accurate depiction of route overlap in the GTFS record, however as will be described below, the single numerical value the calculation produces sheds little light on the varying degrees and locations of overlap in the network. First published in their book *Characteristics of Metro Networks and Methodology for their Evaluation* in 1988, the LOI measure expresses the degree to which rail vehicles share common tracks (*15*). The measure was originally intended to communicate the degree of complexity in planning and scheduling as networks with many routes that share paths (rail or road) are more susceptible to disturbance since only one line sharing the rails or roadway needs to fail to disrupt all other vehicles on that shared portion. A network with few routes, but a high LOI will be more susceptible to network wide disruptions than another network with the same number of routes but a lower LOI.

Although the measure was designed for rail networks, the concept of overlap and the nature of its spatial calculations lend themselves to bus network analysis for the purpose of this study. Similar to the goals of this study, the LOI helps answer the question "how much of this network has overlap?" However, as we will see in the Results and Discussion section further below, the numerical result does not always present an intuitive impression of the network.

The LOI score works by calculating the ratio of the sum of all route lengths over the geometric union length of the network. That is to say, the geometric union only counts the lengths of overlapping regions *once*, while the sum all of route lengths includes those lengths for each instance of overlap. This produces a dimensionless indicator with a minimum value of 1.0 that is comparable from network to network regardless of size or configuration (full formulation provided in methodology section) (Musso & Vuchic, 1988). While this value is useful for comparing the amount of overlap experienced over a time series, or from one network to another, the dimensionless, spatially aggregate result is of no particular use to TII practitioners examining a given network. For example, an LOI increase from 1.46 to 1.56 from one time-window to the next represents an increase in overlapping length within in the network, but this increase cannot account for *where* the changes happened or if the changes will have any particular impact on TII procedures. Links with 3 or more overlapping routes may have experienced a net decrease in length, while links with two or more may have experienced an increase in overall length; these sorts of changes in the network cannot be expressed using the LOI index alone.

The second existing measure of overlap, developed in network measurement research conducted by Liao and van Wee explored the notion of "transit overlap" as the availability of route options between origin-destination pairs. In their work, transit overlap is just one factor in the formulation of their network robustness measure and thus is not included in the collection of indices listed in the tables above. Their research describes the robustness of a transit system as its ability to recover from disturbances such as infrastructure failure and motor vehicle break down. In order for the system to recover quickly travelers must be presented with multiple options to complete a trip (Liao & van Wee, 2017). A network that contains many options to get from origin to destination (high transit overlap) is said to be robust, and the number of route options between origin and destination pairs represent the amount of overlap within that network. It is interesting to note how this measure is almost the inverse phenomenon of that examined by the LOI. I.e.: The conceptualization of the LOI is concerned with disruptions on the network hindering routes (a disruption affects many lines due to overlap), while the transit overlap calculation is concerned with the ability of the network to circumvent disruptions (the disruption has little affect because of overlap).

Since the concept of overlap presented by Liao and van Wee focuses on something altogether different than what hinders TII route matching, it will not be considered for the study.

Similar to the Line Overlapping Index, the Route Overlap indicator comes very close to providing a measure of route overlap on roadways in the GTFS record, but its pair-wise route comparison method results in simplified overlap reporting that does not quantify the degree of overlap on roadways.

Developed by Yuval Hadas and published in 2013, the Route Overlap indicator is intended to measure the efficiency of transfers within the network by revealing pairs of routes that have overlapping sections. The analysis compares two routes at a time to verify if they have overlapping portions and is mainly used to determine how easy it is for riders to transfer between lines. For example, a transfer that requires the riders to cross a street is said to not overlap, whereas routes that use a common stop facilitate transfers. A small degree of overlap will make transfers more efficient, while a large degree of overlap might indicate inefficient planning and could mean the service is not efficiently covering the road network. The output of this analysis is a percentage of overlap between each pair of connecting routes from which a matrix can be constructed revealing all connecting routes in the network. While the handling of the spatial data is similar to what is needed to quantify overlap on road links, the methodology and results do not provide network level insight into the nature of overlap occurring in the network.

For my research, the Line Overlapping Index from Vuchic and Musso will serve as a reference point against which my Overlapping Routes on Links metric will be compared. The research contained in this thesis diverges from previous work in that it examines a way of measuring the network that relies solely on GTFS data. Additionally, by handling the data in a spatially disaggregate manner, overlap can be identified and tabulated on a link-by-link basis in the network, thus providing a fine scale examination of network overlap previously unachieved.

3 Methodology

This section describes in detail the programing environment constructed for the analysis as well as the step-by-step procedures for the calculation of each measure. The section closes with a procedural outline of the GTFS-to-Roads procedure that creates a stand-in road layer from the geometries contained in the GTFS shapes.txt file. This final contribution is crucial in carrying out the Overlapping Routes on Links and Probability of Passage calculations.

The indices proposed in this section are intended to provide: 1) a precise depiction of the spatiotemporal overlap of bus routes in a static GTFS dataset, 2) an idea of how well a TII algorithm will perform in a given network if functioning by overlapping GTFS route shapes with a rider's GPS record, and 3) a GIS toolkit for detecting overlap and pre-processing GTFS data to serve as new input layers in TII processes.

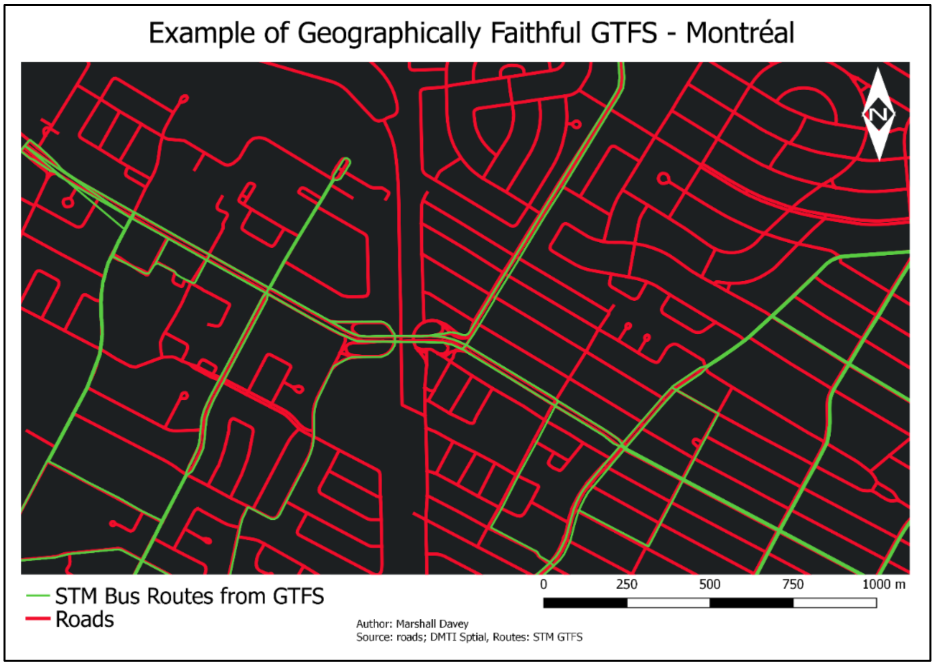
Throughout the GIS procedures scripted to conduct this research many spatial layers are generated containing various network statistics; the geometric union for each sample period, route length and road-link length details, link centroid locations, complete route lists stored in arrays for each link, departure counts for each route on links, as well as global statistics such as total network length, convex hull area, and coverage areas calculated by buffering route shapes at different buffer sizes. Even though the final Overlapping Routes on Links (OROL) result presented in this research is a single numerical value similar to the Line Overlapping index, the methodology that brings us this result provides more information about each network than the previous comparable measures of overlap. Just as the final OROL network layers serve as the input for the POP methodology, it is proposed that the POP output layers can replace the use of GTFS routing data in TII procedures. Each of these layers, which present a more nuanced 'snapshot' view of the network than GTFS can provide, contains the necessary information to conduct travel survey route matching procedures. It is even possible to conduct a Route Overlap analysis as proposed by Liao and van Wee., but thanks to the OROL and POP output layers, it can be conducted in one network-wide sweep rather than on a pair-wise basis.

3.1 Required data and programming environment considerations

GIS analysis was performed using the PostGIS and PostGIS_topology libraries operating in a PostgreSQL relational database. Python scripts control the overall procedure via the Psycopg2 Python library; queries are fed to the database, and data and messages are retrieved from Postgres. QGIS was used for visualization and 2D map making via a connection to the database server, while 3D maps were created in ArcScene, an extension of ArcMap Desktop. The following Python libraries are referenced by the scripts: gdal, psycopg2, os, shapely, time, and sys.

The only data required to conduct this analysis is the static GTFS dataset of a transit agency, real-time GTFS records are not suitable for this analysis in its current form. Once the GTFS zip files of a study region have been collected, the dataset must be geographically validated. In short, a GTFS dataset is *geographically faithful* if the route shapes follow the shape of the road network. The *non-geographically faithful* datasets on the other hand have bus routes represented as the shortest Euclidean distance between stops. This is a crucial aspect of the data requirements of this research as the final expression of the metrics is a percentage of distance measures.

In order to confirm the shape of the routing data provided, the "shapes.txt" GTFS file must be converted into a linestring layer by connecting each consecutive lat/lon point while grouping them by shape id number. This is accomplished with a Python script that loads the data into a database and then executes a PostGIS function to create the new linestring geometry. Once this new spatial layer is created it can then be displayed in QGIS and compared to a street network file obtained through the built-in OpenStreetMap tool. If the routes are confirmed to follow the shape of the underlying roads, the GTFS dataset is suitable for the methods developed in this research (see figure 2 below). In the absence of OpenStreetMap data, the suitability of the shapes file can still be determined by observing each resultant route linestring for the presence of curves, as well as the coincidence of multiple routes that follow the same curves. Conversely if the linestring layer displays abrupt changes in heading between points, and there are no curved edges in the network, it was likely generated from a non-geographically faithful GTFS record.



Author: Marshall Davey
Source: roads; DMTI Sptial, Routes: STM GTFS

Figure 2: Geographically Faithful GTFS bus routes in Montreal

To isolate the shapes of active routes at different times of the day for the OROL and POP calculations, a route selection table is generated for every 15-minute window in the weekday service schedule. The reader may notice further below that windows can run *past* the 24-hour clock with departures and overlap values being recorded for the 25th, 26th, and even the 29th "hour" of a day. These timestamp values are used to depict when a given day's service schedule extends past midnight into the next morning. For example, a route operating every half-hour late on a Saturday night may continue with 30-minute headways until 2:00 or 3:00 am, even if, technically speaking, the Sunday service schedule headways are 60 minutes. Thus, "27" hours represents 3 hours past midnight and a departure at 28:00 is understood by the rider as departing at 4:00am. The values above 24 are not visible to riders as they are only intended to help sort GTFS records. For the OROL and POP calculations, the total range of hour values is retrieved from the stop_times table and then routes are selected for 15-minute subdivisions in each hour.

For the study region of Montreal this resulted in the generation of 104 route selection tables. The 15-minute subdivision was selected to improve the accuracy of the process since larger time windows could lead to the misrepresentation of the network. For example, if two routes leave the same first stop at the top of the hour, and one departs every 15 minutes, and the other only departs once per hour, using time windows of one hour will simply report that two routes are present on that first link. 15-minute increments on the other hand will report 2 routes for the first window, and then only one route for the remaining windows of that hour. 15 minutes was also the time increment employed by the TII experiment applied to the Montreal survey data (Zahabi et al., 2017). In the context of matching travel survey GPS traces to GTFS data, it was assumed for this study that GPS data recorded on bus lines running behind schedule would not run more than 15 minutes late. In other words, if a rider alights a bus at 16 minutes past the hour, and the route has departures every 15 minutes, it is assumed the rider has alighted the bus which departed at 15 minutes past the hour, and not the 1st departure that happens to be running extremely late. Applying this methodology to route matching on trips running exceptionally behind schedule can lead to the miss attribution of routes, however this type of error is expected to be less common than the mis-counting of active routes that would occur as described above if the sampling window were prolonged.

3.2 Study Region

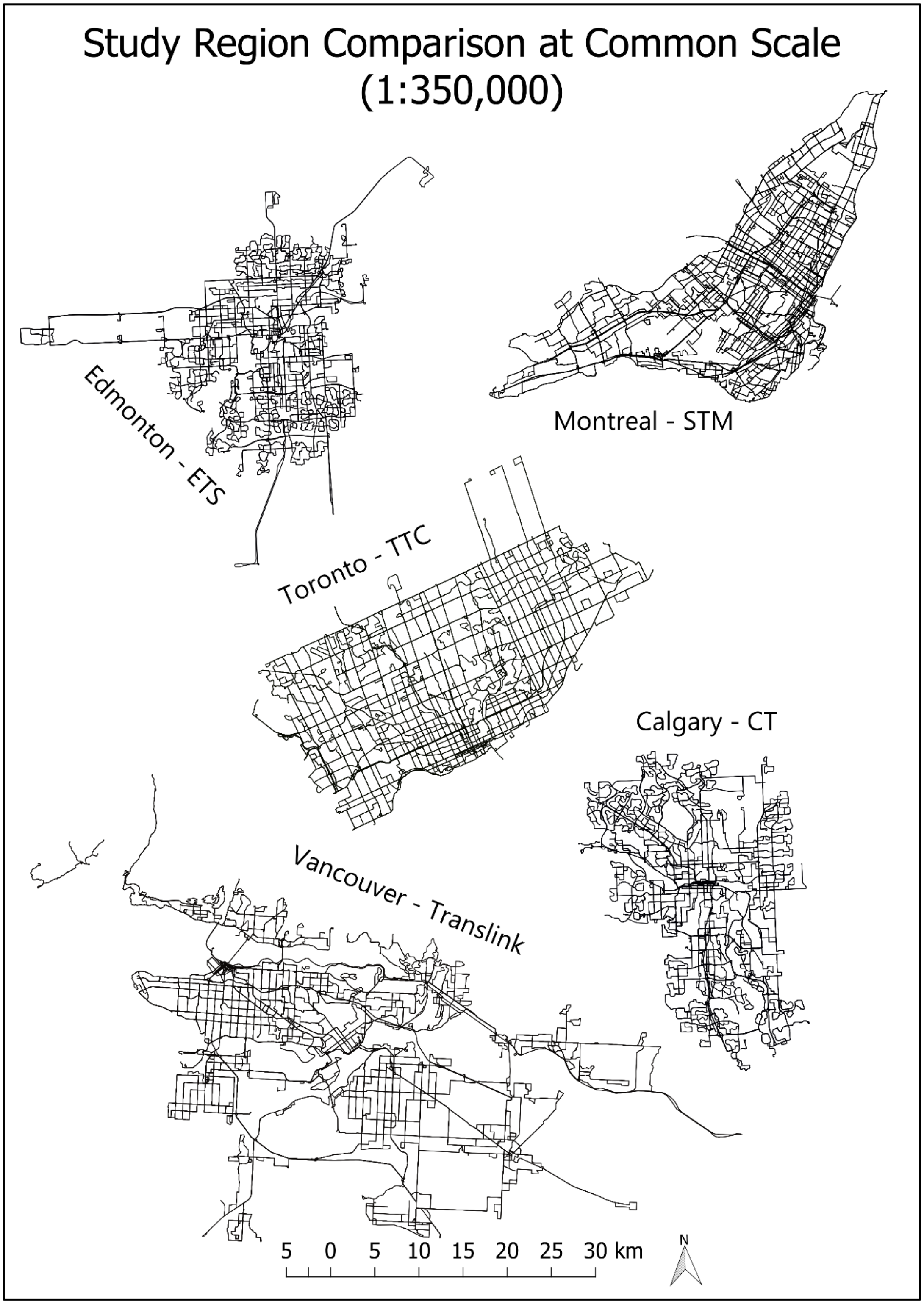


Figure 3: Side by side comparison of study region networks at 1:350,000 scale

This research examines 5 distinct study regions consisting of Canada's most populous census Metropolitan Areas: Toronto, Montreal, Vancouver, Calgary, and Edmonton (see figure 3 above). While the research initially set out to examine overlap in several more transit networks, the requirement that all GTFS datasets be geographically faithful reduced the sample set to Canada's top 5 most populous cities. Other, smaller transit agencies in Canada were discovered to meet the GTFS requirement but were omitted from the study due to having population and modal transit share values much lower than those of Montreal that serves as the study's reference point.

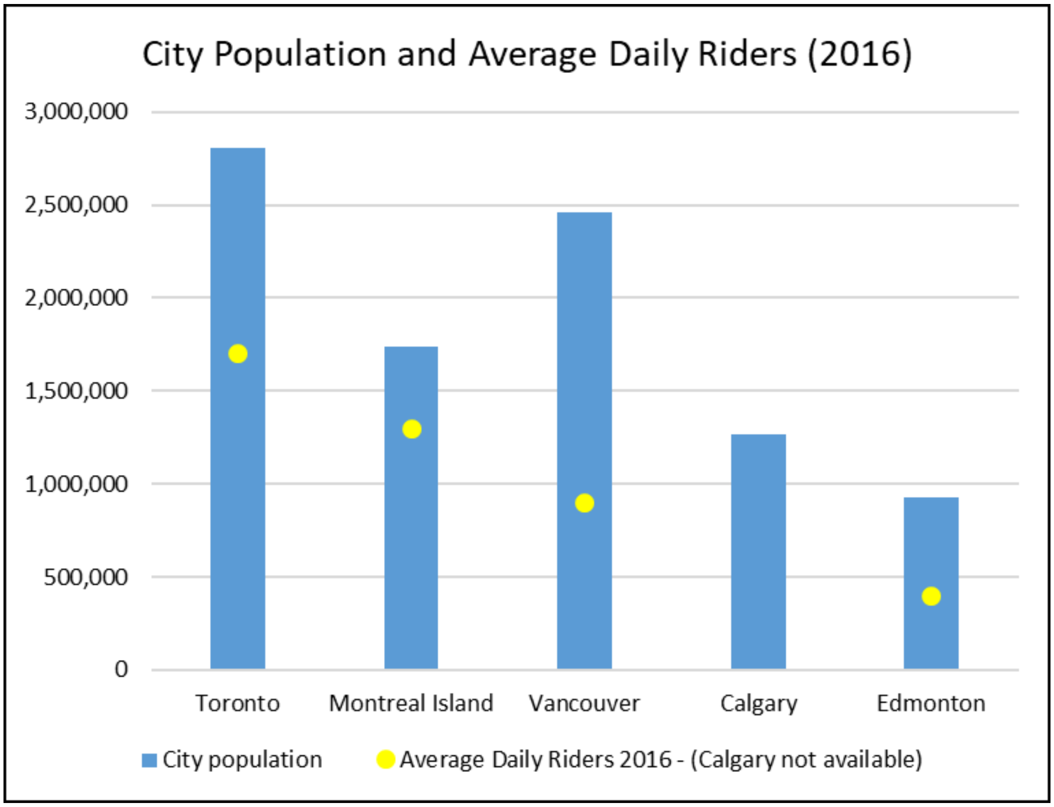


Figure 4: City Population and Average Daily Riders 2016

Figure 4 provides a comparison of city population vs average number of daily riders: Edmonton and Toronto have a daily ridership that is roughly 1/2 of the city population, Montreal's ridership is roughly 3/4 of the city population value, and Vancouver's is roughly 1/3. It should be noted that daily rider count values do not explicitly represent the portion of the local population using public transit system as many people living outside of the population count zones travel into the city each day and contribute to these numbers. The term 'city population' is used flexibly in this study to denote the population statistics that most closely coincide with the areas serviced by the transit network. For example, the population of Montreal's census metropolitan area (CMA) is roughly double that of the Island of Montreal, yet the largest transit agency primarily services the island of Montreal only. For this reason, only the STM network was chosen to represent Montreal and the population and area values are representative of its corresponding coverage, not a particular jurisdictional area. Each of the study regions are handled in a similar manner to provide an accurate view of each *network* rather than each city. Similarly, for Toronto, the coverage of the TTC's network coincides closely with the city limits and not the CMA's boundaries. The only study region whose network closely matches the CMA area is Vancouver whose Translink network services almost the entire CMA. Even though figures 5 and 6 below report Vancouver's city area as almost 3 times larger than the coverage area derived from buffers placed around route shapes, the single agency successfully services the entire region. It turns out the extremely high CMA area value results from the particular geography of the area which include a large bay, inlets, and rivers.

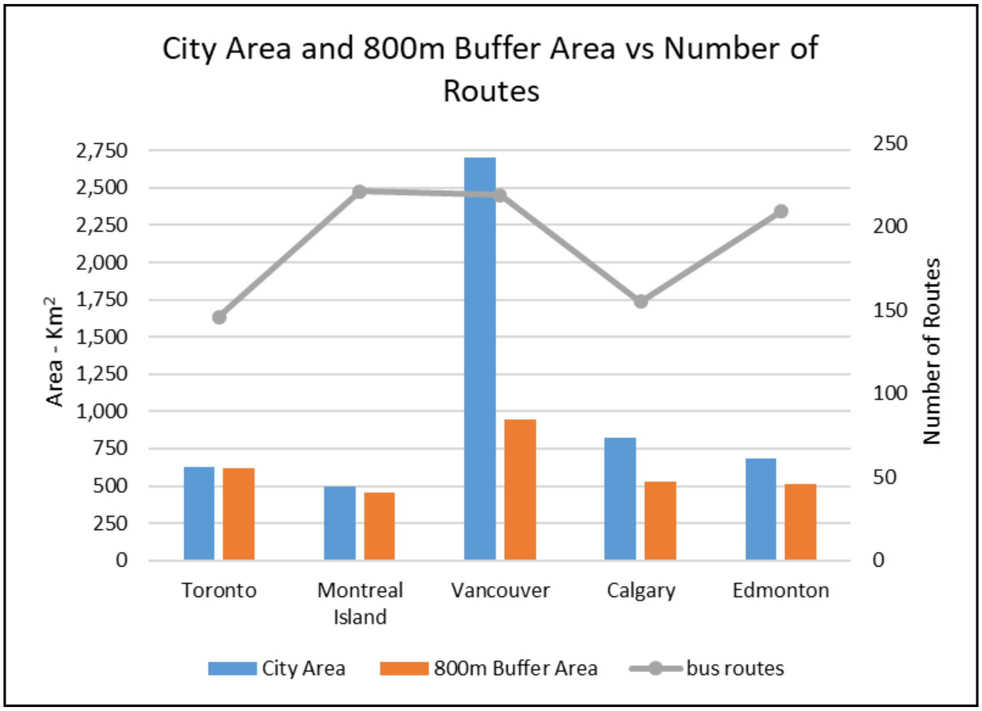


Figure 5: City area vs coverage area of network expanded with 800 meter buffers and number of bus routes

To quantify the coverage area of each network, four different area values were compiled. Buffers were created around the route line shapes with both 400-meter and 800-meter radii. The choice to build buffers around lines instead of stops was made to reflect the nature of this study which focuses on line features and not the point features found in the GTFS record. Next the convex hull area of each network is calculated by forming a polygon that encloses all edges of the network, much like stretching a rubber band around all the map features. Finally, city area values were determined form various combinations of census border files. Figure 6 below provides the surface area values for each study region and compares the area of the city based on its jurisdictional boundaries vs. the buffer areas, and final the convex hull area of its bus network.

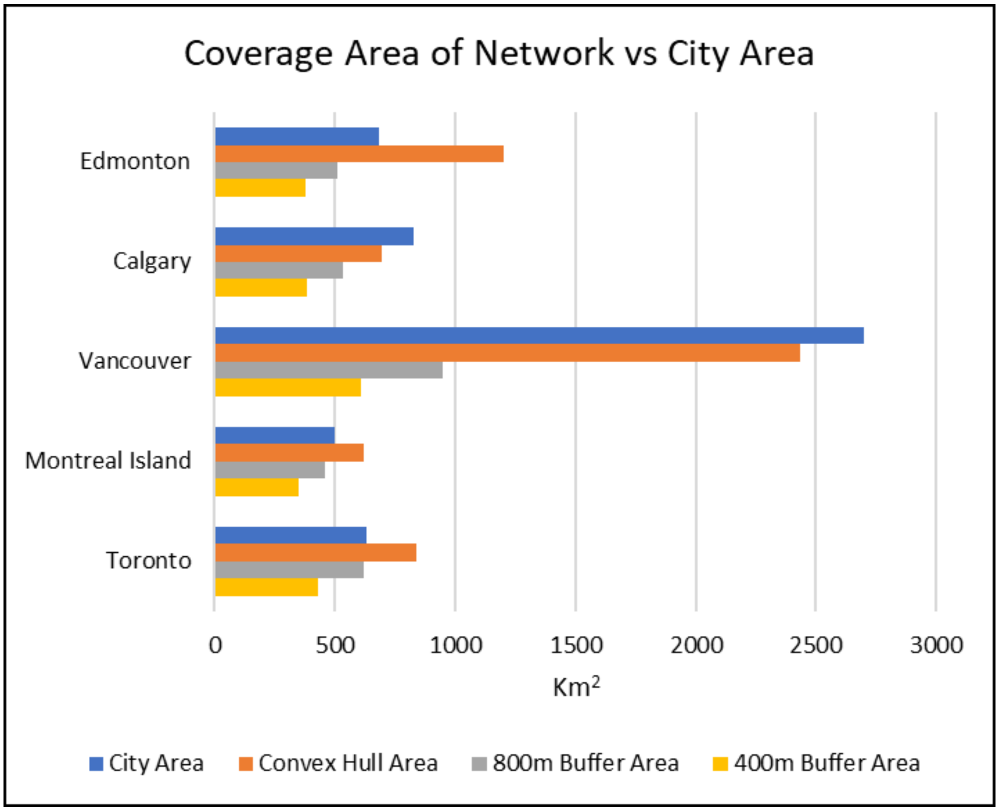


Figure 6: City area, Convex hull area, 800meter coverage area, 400meter coverage area for the 5 study regions

These values are provided to help introduce the reader to each network and give context to the results of each metric. Since route overlap is similar in concept to density, coverage area information is useful for forming hypotheses and assumptions. For example, the assumption can be made that a network with many routes but small coverage area should theoretically experience more overlap than a network that covers a very large area with fewer routes. The methods developed in this research will provide quantitative measures that can address exactly this type of assumption.

It is interesting to note in figure 5 that Vancouver has a comparable number of routes as Montreal, yet according to figure 4 it has a much lower daily rider count. One might surmise that this represents low efficiency in route planning or simply low ridership for Vancouver. Figure 6 then reveals that according to each coverage area calculation, the service area of Vancouver's Translink network is significantly larger than Montreal's, essentially stating that Vancouver covers more ground with its routes than Montreal. It is also known that Montreal has many "out of city" daily riders that enter the network, thus influencing the daily rider count. In short, route counts and coverage areas alone do not tell the whole story and cannot be used by themselves for making accurate predictions regarding overlap in networks.

This underscores the importance of the choice of coverage area statistic as CMA values and census border files do not provide important context regarding the geography of each city. In addition to the inclusion of water bodies that transit can't service, it was discovered that the Vancouver CMA extends Northward into the mountains where the population is sparse and altogether un-serviced by Translink. Vancouver's North-East CMA boundary was observed to be up to 50Km away from the closest Translink route. For these reasons I believe it is important to test different area calculation methods to better understand how much ground is actually covered transit agencies in transit studies.

These graphs and observations are also provided to help illustrate the difficulty encountered when determining the jurisdictional or geographic extent used when comparing transit networks in general. If Montreal's full CMA population was provided in figure 4, for example, we would discover that the ratio of CMA population to daily ridership is more similar to that of Vancouver.

For these above reasons, and to ensure reliable outputs from the SQL queries developed for this research, the decision was taken to examine *one* GTFS dataset per city and to not undergo the complex task of merging GTFS datasets within one study region - a task that often involves strenuous data management and conversions.

3.3 Calculation of Indices

The calculation methodology for each index is provided below. Where applicable, the mathematical formulation of a measure is provided. The section closes with a detailed description of the GTFS-to-roads procedure which generates stand-in road network layers from the GTFS route shapes. The generation of these road layers is a crucial step in this methodology; any mention of links, link lengths, or centroids refers to information derived from these layers.

3.3.1 Active Routes

The Active Routes count is tabulated for each hour of the weekday service schedule and presents the total count of unique routes with at least one departure during each sample period. It tabulates departures through the use of a python loop that iterates over time values. The loop first creates an SQL window of the trips.txt table to isolate trips that occur during the weekday service schedule, these trips are then grouped by shape_id to prevent duplicate counting. With the shapes of weekday service trips isolated, the query then examines the stop_times.txt for routes with departures that match trip ids and that fall within the given sample window. If a route has a departure from its first stop within the time window it is deemed "active." While the results of this calculation do not directly relate to the ability of a TII process to reliably infer route information, it illustrates the variation in service levels throughout the day. As the resultant graph reveals further below, each city experiences a predictable increase in service during peak hours, but what is interesting to observe from the results of this calculation is how the difference in regular vs. peak hour service varies from city to city.

3.3.2 Line Overlapping Index

The Line Overlapping Index works by summing the total length of all the bus routes and then dividing this by total length of roadways covered by these routes. Alternatively stated, the denominator of the ratio is the total length of all routes, *excluding* lengths that overlap existing routes (the geometric union), while the numerator represents the total length of all routes as if stacked end-to-end. This produces a dimensionless score with a minimum value of 1.0, representing to what extent the coverage is duplicated by different routes. A hypothetical city that has only one bus route per road would have a value of 1.0. In terms of gauging the potential for ambiguous routing data, the hypothetical value of 1.0 would represent no route ambiguity possible during the inference process. In such cases rider itinerary could easily be inferred. Conversely, a city that has two routes on each road would yield a value of 2.0 and any GPS data collected in a survey would be considered ambiguous. Since a network with a score of 1.0 is largely hypothetical, the results of this calculation for Montreal will serve as the baseline score for this study. The mathematical formulation for the Line Overlapping Index is as follows:

$$\lambda = \frac{\sum_{i} R_i}{R} = 1 + \frac{\sum_{k=2}^{k_{\max}} R_m^k}{R} \quad (1)$$

R = Length of Network

R_i = Length of route i

$R^k{}_m$ = Length of all overlapping segments,

where k = number of overlapping lines, m is a road link identifier

Equation 1: Line Overlapping Index

Using the active route selection methodology described above, the collection of routes for a given time window are manipulated in one query to produce the LOI ratio for each time frame. The length of the geometric union of the network is produced using the PostGIS functions ST_Length(ST_Union(ST_SnapToGrid("geometry", "snap value"))) where geometry is the collection of route shapes, and ST_SnapToGrid is used to align component edges to a common grid to avoid "non-noded intersection" topology errors. This error was encountered in many GTFS datasets and result from LineString features that cross each other, but with no node recorded at the point of intersection to denote the overlap of the two shapes (a common topological error). Different "snap values" were employed depending on the needs of the network, this value represents the resolution of the new grid the points are transcribed to. This geometric union length is then divided by the more easily obtained SUM(ST_Length(geometry)) which sums the total length of all routes.

3.3.3 Overlapping Routes on Links (OROL)

The Overlapping Routes on Links index (OROL) is the most calculation intensive index mentioned thus far, also different about this procedure is that it involves the comparison of two spatial layers. It references both the GTFS data sets as well as a shape file that represents the road network. As will be described further below, this stand-in road layer is actually constructed from the GTFS data itself and therefore only the single data source is required.

Using GIS intersection and buffering techniques, the OROL procedure examines each individual road link for the presence of overlap in the coincident GTFS record, calculates the sum length of overlapping routes, and writes this sum in a new table that contains the link id, link geometry, and total overlapping route count.

Care had to be taken while handling the GTFS data to ensure that each active route *shape* is present in the analysis, but then to group all route counts by the *route id* of each shape. This aspect of the analysis was discovered to be of the utmost importance in accurately counting the routes present on a link since it was possible for one bus route to have two different paths (shapes) on the road network within the same sample window. Sometimes this was due to both directions of a route being listed with the same route_id, which is not always the case from agency to agency, or other times the same direction might have two different shapes during the 15-minute sample period simply due to small differences in how the vehicles enter and exit bus termini. Failure to account for this discrepancy in GTFS records can lead to the over-estimation of overlap in the network. For example, the results may report a route count of 2 on a link, and then closer inspection reveals that both routes have the same headsign and direction id, meaning they are effectively the same route.

The OROL scores are calculated for 15-minute periods over a full day of the weekday service schedule following the methodology of the TII research conducted in Montreal. This is to say that for every 15-minute period throughout the service schedule, an individual spatial layer displaying all of the routes with departures during that period is generated in the database. For the study region of Montreal this resulted in the creation of 104 “routes info” tables, each named with its time frame and transit agency abbreviation.

Once the count of overlapping routes is tabulated for each link, we can then calculate the portion of the network that has overlapping routes (route count >= 2). The resultant percentage represents a theoretical maximum amount of ambiguous distance a TII may encounter during one sample period. Using the validated trip data and route detection rate from the TII study in Montreal we can estimate the degree of success a TII will have in another network simply by ranking each network according to its OROL percentage, a table containing these rankings is provided in the results section (Table 6: Ranked results, page 72).

The calculation is performed as follows: first working on the stand-in road network layer, buffers 40cm in diameter are generated around the centroid of each road link and assigned the same ID number as its parent road-link. This relatively small buffer size of 40cm is capable of capturing all coincident links due to the nature of the stand-in road network that follows almost exactly the path of the bus routes. This small buffer size is also helpful for avoiding the overlap of buffers in complex areas such as merges and multilane intersections that can lead to the over estimation of overlap.

Next, an intersection is performed between each active route layer and the buffers. This step records which bus routes pass through each of these buffers and writes the following information to a new table: buffer/link ID as the primary key, shape_id, route_id, direction_id, trip_id, and route_headsign. At this stage instances of overlap are represented by the repetition of link_ids in the first column. i.e.: a link id will appear in the table as many times as there are routes on it.

Next, this table is analyzed and compiled into a new spatial table that contains one row per link_id. Unique route ids (or headsigns depending on the structure of the GTFS) are counted on each link and this information is written to the new table as a route count value along with each link's geometry.

The percentage of the network consisting of links with 2 or more routes can now be calculated by summing the lengths of all links that meet the filter criteria, divided by the total network length. This percentage is calculated for each 15-minute examination period and written to a CSV file which is used to produce the graph of results. The mathematical formulation of the OROL score is provided below:

$$\boldsymbol{OROL}\,\% = \frac{\boldsymbol{\Sigma_{i=1}^{n} R_i}\ \boldsymbol{where}\ \boldsymbol{c_i > 1}}{\boldsymbol{\Sigma_{i=1}^{n} R_i}} * 100 \tag{2}$$

i = road link

R_i = Length of link i

C_i = route count on link

Equation 2: Overlapping Routes on Links Percentage

To further illustrate the difference in LOI and OROL calculations as well as the variability of the factors in each equation several network diagrams are present below in Figure 7. The three factors presented: N, GU and OVR are all length values derived from the sample networks with N being the total network length, GU, the length of the geometric union and finally OVR being the length of all links with a route count above or equal to two.

In each sample network, the geometric-union of the network results in the same shape with the same length (700 units). For these examples it should be assumed that bus routes run in the same lane; they are depicted side-by-side for illustration only. The route count classes in the OROL output layers are color coded to depict how each road link is assigned a class of overlap. Comparing the first row of diagrams to the second row, we can see that increasing overlap length (OVR) increases both LOI and OROL. Comparing the second row to the third we see that increasing network length and having a higher maximum degree of overlap does not automatically change the OVR value. This is a good example of how the LOI communicates more about *degree* of overlap, whereas OROL is more concerned with *amount* of overlap. After examining several sample networks it was determined that each of the factors involved in these calculations, namely: N, GU, and OVR can all increase or decrease independently of each other depending on the particular network configuration.

The comparison of sample networks and discussion of variables is included to help impress upon the reader the difference in how each measure communicates information about the network. If one can imagine the complex nature of the geometric relationships needed to describe transit networks at different times of the day, it is clear that the difference be LOI and OROL results, and the higher values for LOI themselves, do not lend themselves to intuitive conclusions.

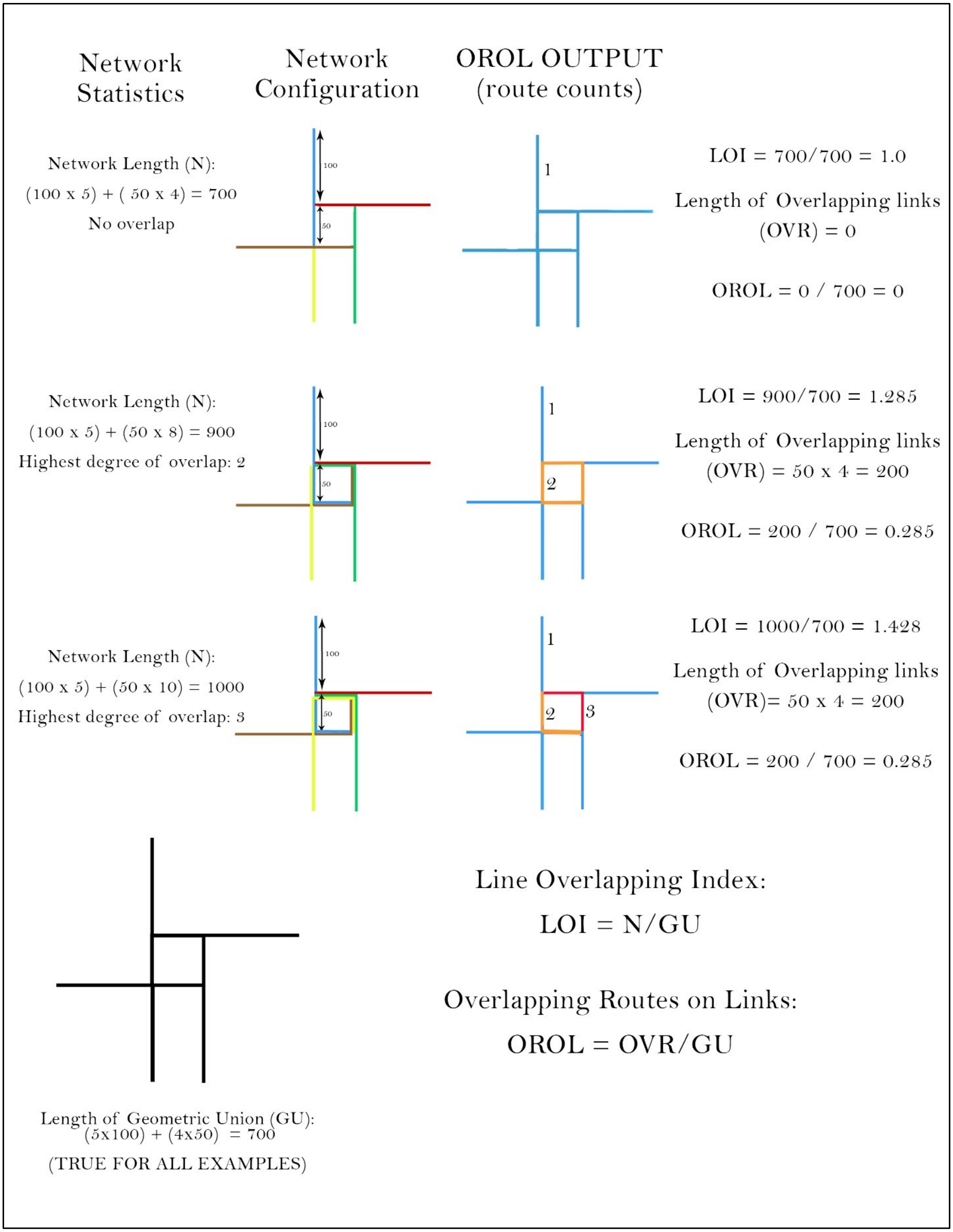

Network Statistics
Network Configuration
OROL OUTPUT (route counts)
Network Length (N):
(100 x 5) + (50 x 4) = 700
No overlap
100
50
1
LOI = 700/700 = 1.0
Length of Overlapping links (OVR) = 0
OROL = 0 / 700 = 0
Network Length (N):
(100 x 5) + (50 x 8) = 900
Highest degree of overlap: 2
100
50
1
2
LOI = 900/700 = 1.285
Length of Overlapping links (OVR) = 50 x 4 = 200
OROL = 200 / 700 = 0.285
Network Length (N):
(100 x 5) + (50 x 10) = 1000
Highest degree of overlap: 3
100
50
1
2
3
LOI = 1000/700 = 1.428
Length of Overlapping links (OVR)= 50 x 4 = 200
OROL = 200 / 700 = 0.285
Line Overlapping Index:
LOI = N/GU
Overlapping Routes on Links:
OROL = OVR/GU
Length of Geometric Union (GU):
(5x100) + (4x50) = 700
(TRUE FOR ALL EXAMPLES)

Figure 7: Sample network diagrams depicting LOI and OROL calculations

It has been noted that when the maximum overlapping route count in the network is 2, the LOI and OROL calculations will return comparable values (i.e.: 1.46 and 0.46 respectively), an example of this is provided in the second row of networks in figure 7. However, as the maximum degree of overlap in the network increases above 2, the OROL percentage will remain fixed and the LOI value will increase (see 2nd and 3rd networks in figure 7).

The difference in reported scores between LOI and OROL, when the maximum degree of overlap surpasses 2, results from the definition of overlap as coded into the OROL equation. In its intended and standard form, the OROL percentage defines any link with 2 or more routes as "ambiguous" since these are the areas which may obfuscate route matching processes during TII.

If the OROL qualification for overlap is instead incrementally increased (route count >=2, route count >= 3… route count >= q_{max}), the sum of OROL percentages for each category of overlap will produce the same value as the LOI minus 1. In one sense, the OROL score of a network is a special case of LOI where the maximum degree of overlap never surpasses two. The relationship between LOI and OROL is expressed in equation 3 where q takes on the value of route count categories detected within the network. When the qualification for overlap is parameterized like this, the OROL methodology effectively presents a method for decomposing the LOI score into meaningful categories of overlap.

$$\boldsymbol{LOI - 1 = \sum_{q=2}^{q_{max}} OROL_q}$$

$$\boldsymbol{where\ OROL_q = \frac{\sum_{i=1}^{n} R_i \ where\ c_i \geq q}{\sum_{i=1}^{n} R_i}} \tag{3}$$

q = overlap qualifier

q_{max} = maximum overlapping route count value found in network

i = road link

R_i = Length of link i

C_i = route count on link

Equation 3: OROL as the Decomposition of LOI

3.3.4 Probability of Passage Score (POP)

The Probability of Passage score provides more information regarding the nature of each instance of overlap by examining how many departures each route has during the sample period. Expanding upon the layers generated during the OROL calculations, this procedure produces a POP value for each route present on a link which then allows for the calculation of the portion of network consisting of ambiguous links once the links have been filtered for POP values above a user determined threshold.

The POP score of a route is the ratio of the sum of that route's departures divided by the total number of departures on that link during a given sample period (Equation 4). For example, if a given road link has two routes present during the examination period, and route 1 has one departure, while route 2 has two departures, we can say the POP score for route 1 is 1/3 (0.33), and the POP score for route 2 is 2/3 (0.66). On this link the OROL score will simply report a route count of 2 which then gets classified as overlap. Here the POP score provides a more nuanced description of the link by allowing the user to assume that route 2 is more likely to have been used during this time period.

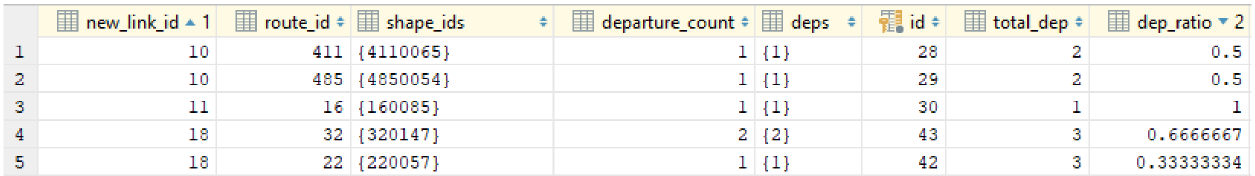
$$\boldsymbol{POP_{r,i} = \frac{\Sigma\, departures_{r,i}}{\Sigma_{r=1}^{n} departures_{i}}} \tag{4}$$

r = transit route

i = road link

Equation 4: POP score

Table 5 below shows the POP analysis results for 3 links in a sample network. The first link (new_link_id 10) has two routes present, each with one departure during this time window. Each route therefore has a POP score of 0.5. This link is considered ambiguous since there is no clear "winner" between the routes. The next link in the table (new_link_id 11) has one route with one departure and a POP score of 1.0, this link is unambiguous. The final link has an OROL count of 2 with route 32 having 2 departures, and route 22 having 1 departure. This sample link provides the necessary conditions for the selection of the route that is more likely to pass during this time window.



	new_link_id ▲ 1	route_id	shape_ids	departure_count	deps	id	total_dep	dep_ratio ▼ 2
1	10	411	{4110065}	1	{1}	28	2	0.5
2	10	485	{4850054}	1	{1}	29	2	0.5
3	11	16	{160085}	1	{1}	30	1	1
4	18	32	{320147}	2	{2}	43	3	0.6666667
5	18	22	{220057}	1	{1}	42	3	0.33333334

Table 5: Sample of POP table showing 3 links with dep_ratio as POP score.

Similar to how the OROL percentage represents the overlapping portion of the network, the POP % also reports overlap but only after links have been filtered according to a threshold POP limit: α (see equation 5)

$$POP\ \% = \frac{\sum_{i=1}^{n} R_i \ where\ POP_{r,i} < \ \alpha}{\sum_{i=1}^{n} R_i} \tag{5}$$

r = transit route

i = road link

α = probability of passage threshold

R_i = Length of link i

Equation 5: POP percentage

The advantage of the POP score is that thanks to the probability of passage scores the researcher can effectively reduce the amount of ambiguous links, and thus, the overall ambiguity of the network. Where the OROL scoring system reports ambiguous links as any link having two or more routes, the POP score allows the researcher to choose a threshold of probability with which to filter the network. For the above example, if the researcher decides that a link containing a route with a POP of 0.6 or higher should not be considered ambiguous, and instead we treat the route with the highest POP score to be the unambiguous "winner" of the link, we have effectively removed an ambiguous length of road from the TII's input data.

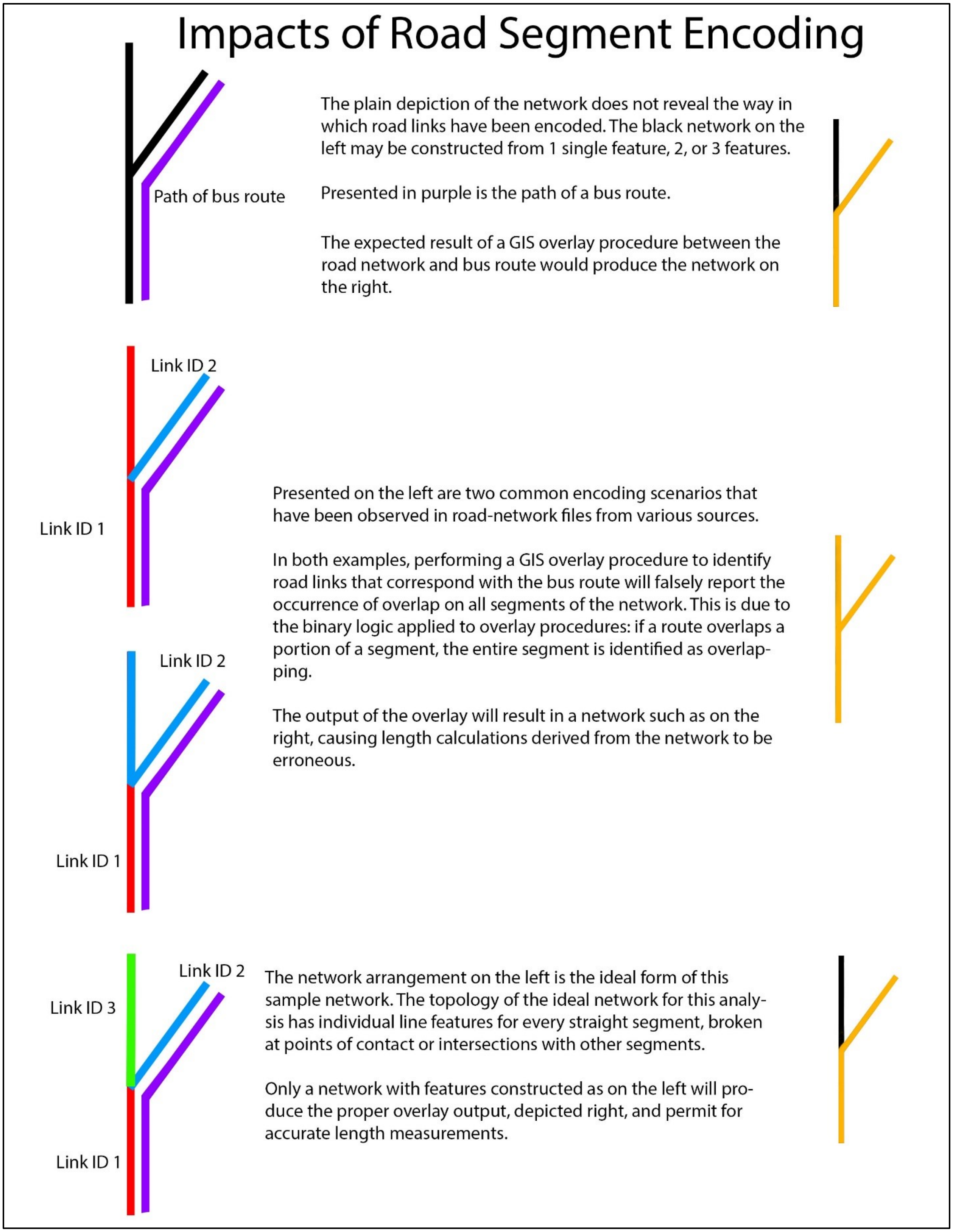
The end result of the POP procedure is a spatial layer that contains multiple entrees for links with overlapping routes, each with its own geometry as well as the POP score. This allows the researcher to set a POP threshold (α) determining which level of POP is sufficient for their needs, and then export multiple spatial layers representing different degrees of overlap in the network.

3.3.5 Generation of the Meta-Road-Layer

As mentioned above, finding reliable road network files proved problematic in early tests of these processes. Having consistent topological rules from one road network file to another is key in producing comparable results between different cities. For example, a bus route may turn off of a roadway while the road layer's linestring feature continues straight, a GIS intersect procedure will flag this link as active along its entire length, thus exaggerating the length of the path actually travelled by the bus. To avoid such errors, a novel process was developed in which the routing information contained in the GTFS files will be used to generate a "quasi-road network file". Since the OROL and POP indices examine each link in a road layer for the presence of a bus route, it follows that these procedures need only examine links that actually have bus routes on them. Initial versions of the OROL procedures involved scanning each link in a road layer that spans the entire city. This approach was computationally intensive and as a result inefficient for studying multiple regions (initial tests would take close to 20 hours to complete analysis of a network). Next, the pre-generation of a road layer via the intersection of bus routes and road features was attempted. It became apparent via initial trials that such GIS processes were also computationally demanding and obtaining road network files which matched the coverage of the transit network was also difficult. Road layer files that match the extent of the transit network were difficult to come by and early attempts involved stitching together different road files which sometimes had different attributes. Layers obtained from the OpenStreetMap project were discovered to have different topological rules from one city to another, and worse still, pedestrian and cycling paths were sometimes registered as roads which led the attribution of bus routes where no roads even existed.

Thus, the idea to use the geometries of the GTFS routes themselves to serve in the place of a road network file was proposed. By creating a stand-in road network file that contains *only* the links worthy of study the processing demands were greatly reduced, the topology of each study is guaranteed to be consistent as the topological rules are determined by the script that generates the layers, and furthermore, the research now only requires a single source of data.

Presented below in Figure 8 are road network examples depicting how inconsistent topology can contribute to erroneous length measurements:



Impacts of Road Segment Encoding

The plain depiction of the network does not reveal the way in which road links have been encoded. The black network on the left may be constructed from 1 single feature, 2, or 3 features.

Presented in purple is the path of a bus route.

The expected result of a GIS overlay procedure between the road network and bus route would produce the network on the right.

Presented on the left are two common encoding scenarios that have been observed in road-network files from various sources.

In both examples, performing a GIS overlay procedure to identify road links that correspond with the bus route will falsely report the occurrence of overlap on all segments of the network. This is due to the binary logic applied to overlay procedures: if a route overlaps a portion of a segment, the entire segment is identified as overlapping.

The output of the overlay will result in a network such as on the right, causing length calculations derived from the network to be erroneous.

The network arrangement on the left is the ideal form of this sample network. The topology of the ideal network for this analysis has individual line features for every straight segment, broken at points of contact or intersections with other segments.

Only a network with features constructed as on the left will produce the proper overlay output, depicted right, and permit for accurate length measurements.

Figure 8: Impacts of Road Segment Encoding

The new “road network” file is produced as follows (a flow chart of this process is provided in Figure 10 on page 41): first, generate a geometric union of all the bus routes as described in the routes.txt and shapes.txt files. Next, the unioned geometry is converted to a GeoJSON format to facilitate reading the coordinates of each segment of the unioned feature. From here a custom script employs a loop that unpacks sequential pairs of coordinates and constructs individual linestrings from each pairing. These coordinate pairs represent the beginning and end points of straight segments of the original feature, as well as straight lines between points of inflection. With these pairs now isolated in a list, they can be joined together into linestrings according to their sequence. What results is a new linestring file that follows route shapes but with individual features between every intersection and/or point of inflection.

For the sample region of Montreal, the above described process converted 624 bus route shapes into one multi-linestring feature, and then decomposes it into 188,564 individual linestring features.

At this stage, however, there still exists many overlapping linestrings that will lead to incorrect route attribution, and worse still, incorrect length calculations when determining the percentage of overlap. In order to isolate the minimum amount of linestrings required to accurately describe the road network, a series of additional GIS and data management procedures are employed as described below.

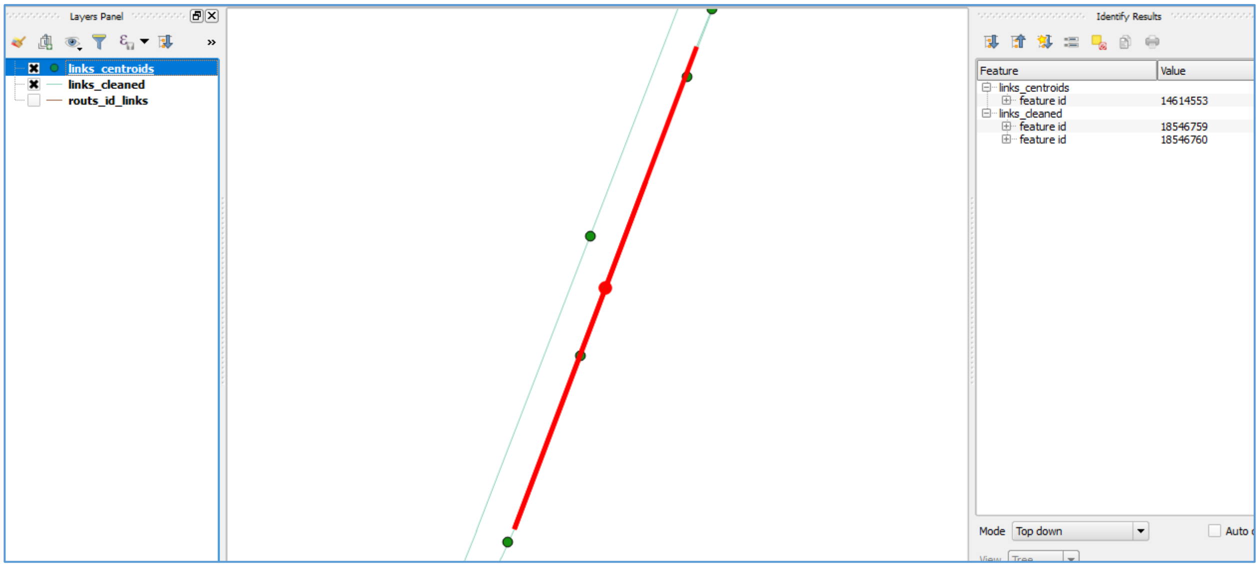
First, perfectly overlapping links are filtered and all but one are deleted from the record. This reduces the link count from 188,564 down to 75,366. Next, links that perfectly overlap but have opposite start and end points are identified and all but one are deleted from the record. This is accomplished by identifying lines with intersecting centroids. The assumption is made that if two lines intersect *and* share a common centroid then only one of the lines is required. This assumption is reliable due to the way the script unpacks the coordinates from the GeoJSON format; instances of a shared centroid from lines with different headings are not possible after the initial script essentially breaks lines at each crossing point. This step reduces the count of linestring features from 75,366 to 64,673 for Montreal’s STM network.

At this stage, the resulting linestrings provide the general shape of the road network occupied by routes, but there are still many overlapping or near-coincident lines where multiple routes share roadways. This is due to the topology of the route shapes as described in the shapes.txt coordinates. For example, two routes may travel down the same lane, yet the components of each respective linestring may have nodes and vertices in different locations along each line, thus failing the previous checks. Another complication is when the route linestrings are not encoded in a common centerline of a lane. In some areas of high overlap, near-coincident links were found to be as little as 8mm apart from each other on the ground, many of these were also non-parallel, leading to multiple links crisscrossing each other down the roadway. In order to present a clean depiction of the roadway the superfluous, crisscrossing links will have to be identified and removed from the record.

In practice, converting these conflicting lines into one simplified representation is not as straightforward as identifying links that just overlap or touch. GIS tools from QGIS, GRASS, ArcMap, PostGIS, and PostGIS_topology were all put to the test in an effort to clean these redundant linestrings and it was discovered that there is no existing tool that can handle the variety of geometric relationships existing between all of these lines in the network. Simply eliminating one of the overlapping lines in every instance of overlap or contact is not guaranteed to leave behind a complete depiction of the road way. Determining which of the 64,673 line segments must be retained to produce a simplified network proved to be the biggest challenge of developing this procedure.

The final class of superfluous links were discovered to occur most often in areas where shapes for multiple routes "zig-zag" back and forth over long distances down a single lane. These lines fail the filter criteria used to clean the network thus far. For example, a one lane roadway may have 2 routes represented by a single shape along most of its length, only to have the path break off into parallel lines for a short duration (each reporting only one route present), and then converge back into one line that once again reports the correct route count of 2. Inspection of these areas using google satellite imagery revealed no reason for the bus shapes to diverge in this manner while travelling down a roadway. Visual inspection of these lines in QGIS revealed that the near identical lines were at most 18mm away from each other on the ground.

The above described scenario of route count values suddenly changing on a long stretch of road was one of the main indicators that the overlap procedure was not producing correct results. Viewing the output layers in QGIS and applying contrasting colors to the route count categories allowed for easy inspection of an entire network to identify these zones. To further refine the link-deletion procedure sample geometries were extracted in these problematic regions. By focusing on these problematic geometries such as in figure 9, the testing of new GIS procedures was expedited allowing for multiple approaches to be tested in a timely manner.



☒ links_centroids
☒ links_cleaned
☐ routs_id_links

Feature	Value
links_centroids	
feature id	14614553
links_cleaned	
feature id	18546759
feature id	18546760

Mode Top down ☐ Auto

Figure 9: Near Duplicate Links in the GTFS record are difficult to remove and produce errors in length calculations.

Visible in Figure 9 is this last type of erroneous link; in this example the link highlighted in red coincides with two other, shorter links (approximately 8mm apart), whose centroids are visible in green. To address this last type of erroneous link, the most calculation intensive portion of the road generator process was developed with the aid of TRIP lab member Kyle Fitzsimmons. Referred to as the Route Flattening procedure, it begins by creating a table of all the nodes of all the links. Next, using the ST_DBScan function built into PostGIS, nodes within a threshold distance of links are all assigned the ID number of the closest link (Birant & Kut, 2007). The ST_DBScan algorithm functions by identifying groups of similar spatial elements using clustering techniques and was the only tool discovered that could address the varying geometric relationships of all these linestrings. The geometric relationships between these lines are essentially tossed aside by this technique as the ST_DBScan algorithm is only fed nodes, which are then categorized according to distance to the closest neighbouring linestring. This approach is fundamentally different than comparing and measuring linestrings against linestrings. After some testing, the search threshold distance of 30cm was chosen as the value least likely to negatively impact the output data.

Next, the script identifies all links that pass through, or are within 30cm of more than two end nodes. When this type of link is identified, it gets split by the middle node and the new section is assigned an identifier. If a link only touches two nodes it is a legitimate link and is classified as such. Once this step is completed, the network is then reconstructed to form the legitimate links and the first section of links that were subdivided at nodes. This final step brings the link count from 64,673 down to 63,453.

The resultant layer is now an accurate representation of the roadways used by the transit service and consists of one single line feature located approximately on the center line of the physical road ways. Three lane boulevards and highways are also represented as one single line per direction. This improves processing time as any GIS procedure involving the road network is now only fed the specific underlying links needed for analysis rather than iterating every road segment in the city. This also completely does away with the error of the misattribution of routes to nearby road-file segments during intersection and buffering procedures.

The full script of this procedure is available at https://github.com/TRIP-Lab/GTFS-to-Roads-converter

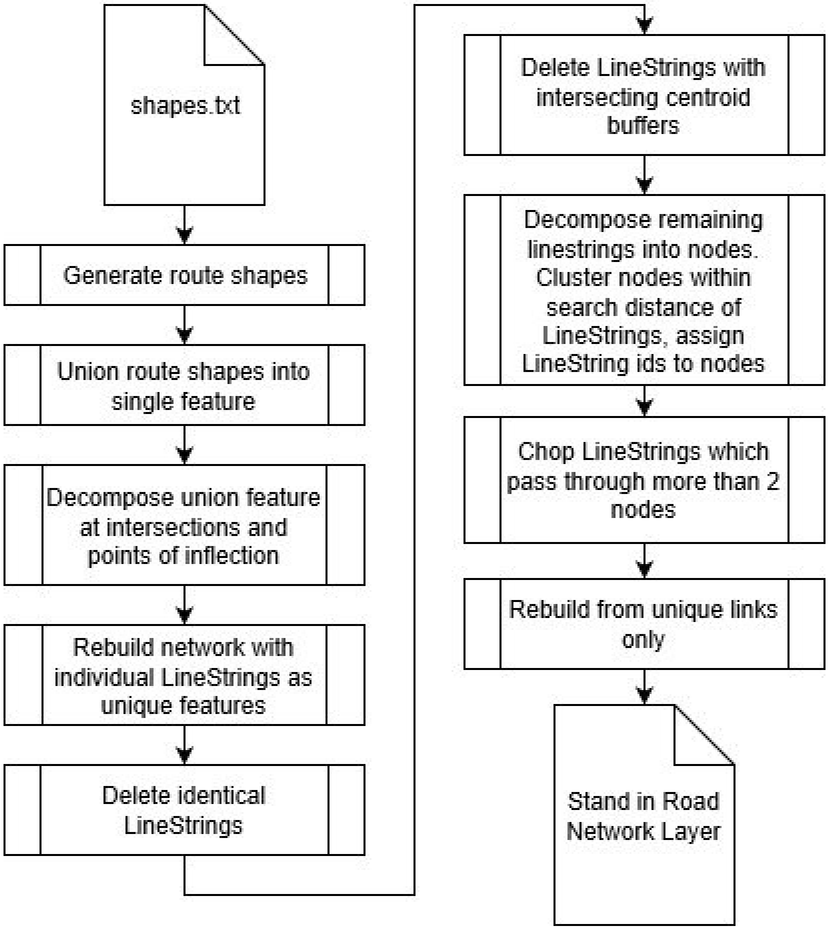

shapes.txt
Generate route shapes
Union route shapes into single feature
Decompose union feature at intersections and points of inflection
Rebuild network with individual LineStrings as unique features
Delete identical LineStrings
Delete LineStrings with intersecting centroid buffers
Decompose remaining linestrings into nodes. Cluster nodes within search distance of LineStrings, assign LineString ids to nodes
Chop LineStrings which pass through more than 2 nodes
Rebuild from unique links only
Stand in Road Network Layer

Figure 10: Procedural outline of the GTFS to meta-road-layer procedure.

.

To validate this new network representation, the GTFS-to-Roads output is loaded overtop of the original bus route shapes in QGIS and each layer is given contrasting colors. Next, by visually scanning over the entire network "holes" in the output layer can be identified where the GTFS-to-Roads procedure removed too many links on a conflicting region. Observing these regions allowed for the fine-tuning of the script parameters that identify near-coincident links. Namely, the ST_DWithin search distance which identifies links that fall 'within' a defined search radius of nodes, and the clustering search distance parameter in the ST_ClustrDBScan which decides which nodes should be 'clustered' together as a coincident group. If each search radius is too large, useful road links are deleted leaving 'holes' in the network where stretches of valid roadways have been deleted. Using too small a search radius conversely results in the retention of erroneous links. With the exception of Toronto's TTC, the 'within' search radius of 30cm and a cluster grouping of 1meter reliably produced a clean and complete depiction of each network: a relationship between networks believed to emerge due to the standardized topology of each resultant network. Toronto required a ‘within’ search distance of 40cm and the same cluster grouping distance of 1meter to specifically handle some conflicting lines where busses enter termini in the TTC network.

4 Results and Discussion

The previous section established how the mathematical formulation of each measure has been adapted to GIS procedures and how the stand-in road layer is generated to facilitate the calculation of the Overlapping Routes on Links and POP scores. What follows below is a detailed examination of the results for each metric as well as the presentation of maps for each region revealing links categorized by the OROL procedure. Where possible, the results of multiple measures are compared to each other to offer further insight into what each measure can or cannot express about each network.

4.1 Active Routes

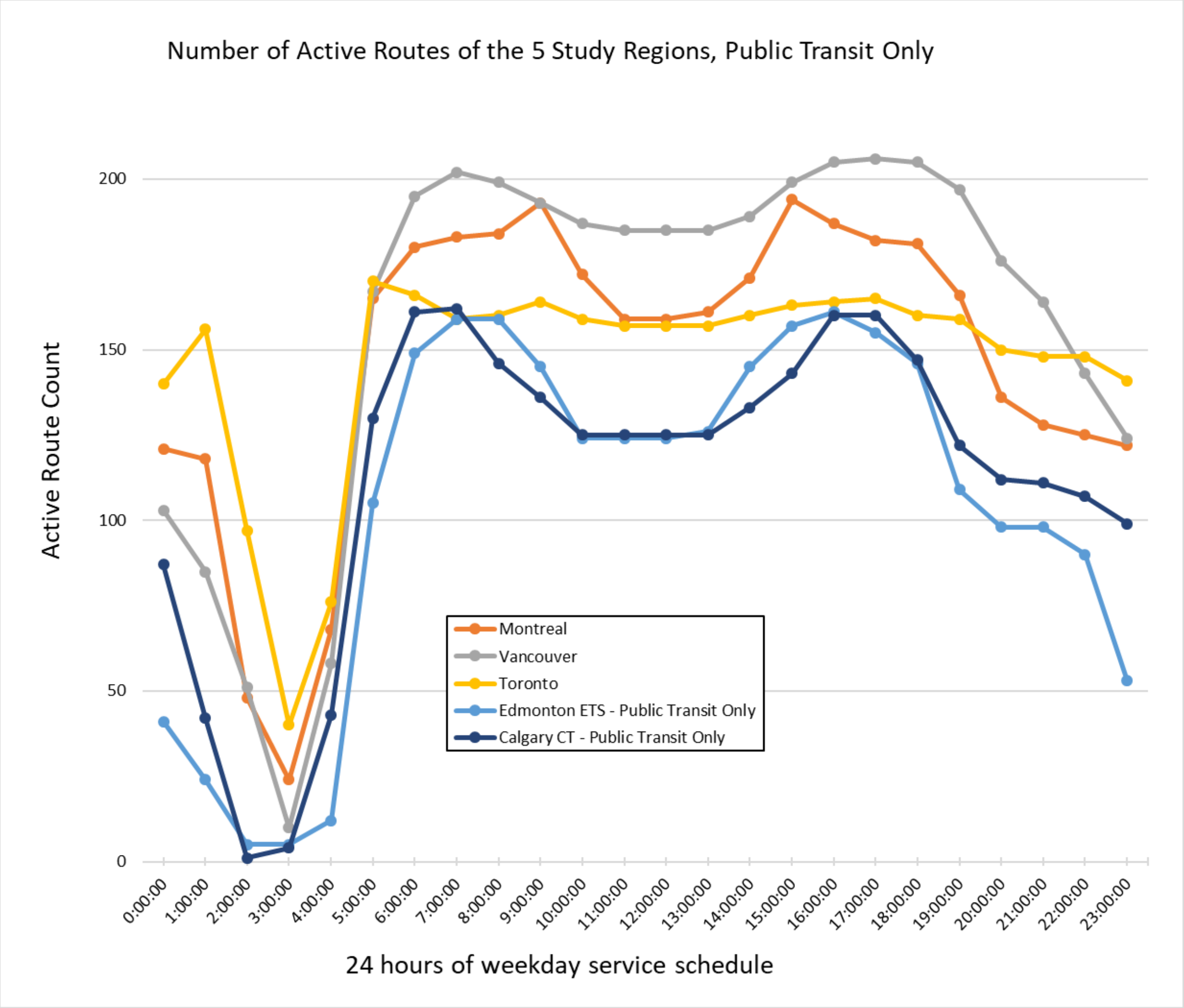


Figure 11: Number of active routes per hour of the day of the five study regions.

The change in the number of active routes over the 24-hour service schedule follows a predictable pattern in each of the study regions with the exception of Toronto; service is greatly reduced in the early morning hours and both rush-hour periods of the day exhibit the most active routes of the day. Toronto's results display the lowest amount of fluctuation over the daytime period. This phenomenon is also reflected in the OROL results graph further below. Upon closer inspection of Toronto's service and coverage levels it was determined that the city's broad surface area combined with a consistent rectilinear road network has led the network to consist of many routes that follow long, straight paths across this city. It is important to note that the count of active routes does not consider how many departures each route has for each hour. It is plausible that Toronto's service levels may in fact increase during peak hours due to more frequent departures, while not having to add additional routes to the network like other agencies in the study. If this is indeed the case for how the TTC addresses peak hour ridership, it would follow that the degree of overlap measured via the OROL index will fluctuate very little since the road links occupied by these routes will change very little over the time series. The OROL results graph further below will shed more light on this hypothesis.

Also of note in figure 11 is the labelling of Calgary and Edmonton that specifies these results are for "public transit only". Initial results of the active routes count revealed extreme peak hour values for both Calgary and Edmonton, sometimes more than doubling the number of active routes from one hour to the next. Upon examining the routes tables, the transit agency websites, and eventually following up with the agencies themselves, it was discovered that these two agencies also provide school bus service to the local school boards. Posted below in figures 12 and 13 are the active route count results for Calgary and Edmonton depicting the unfiltered data vs the count of public transit routes, excluding all known school routes. In the case of Calgary, the data was also filtered to include Bus Rapid Transit (BRT) routes since these vehicles run on roadways and can overlap with regular bus service routes.

Likewise, for each of the measures calculated during this research these cities displayed a major difference in results when calculating the unfiltered data-feed vs the public transit routes only. This underscores the importance of inspecting and understanding the GTFS data of any given agency; it became obvious early in this research that the contents and data structure of the GTFS datasets can vary widely from one agency to another. Edmonton, for example, has every bus route recorded as "route id = 1" across its entire fleet, drastically diverging from the typical use of this identifier. While GTFS is proposed as a standard, its main purpose is to facilitate itinerary calculations in scheduling applications such as Google Maps. Any research conducted comparing datasets should proceed with due diligence.

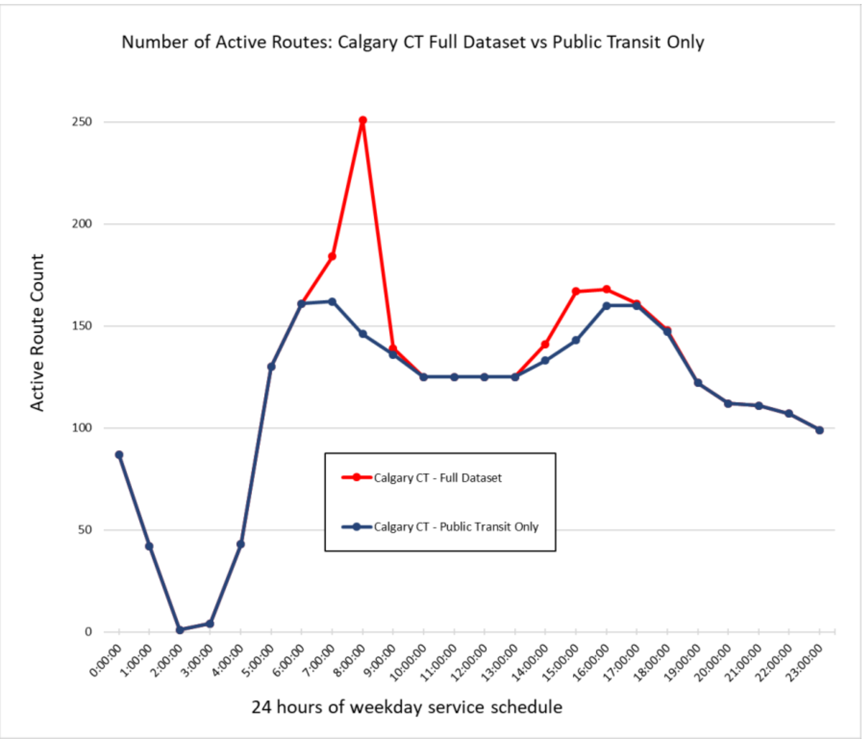


Figure 12: Number of Active Routes over 24 hours in Calgary: full GTFS record vs public transit lines only

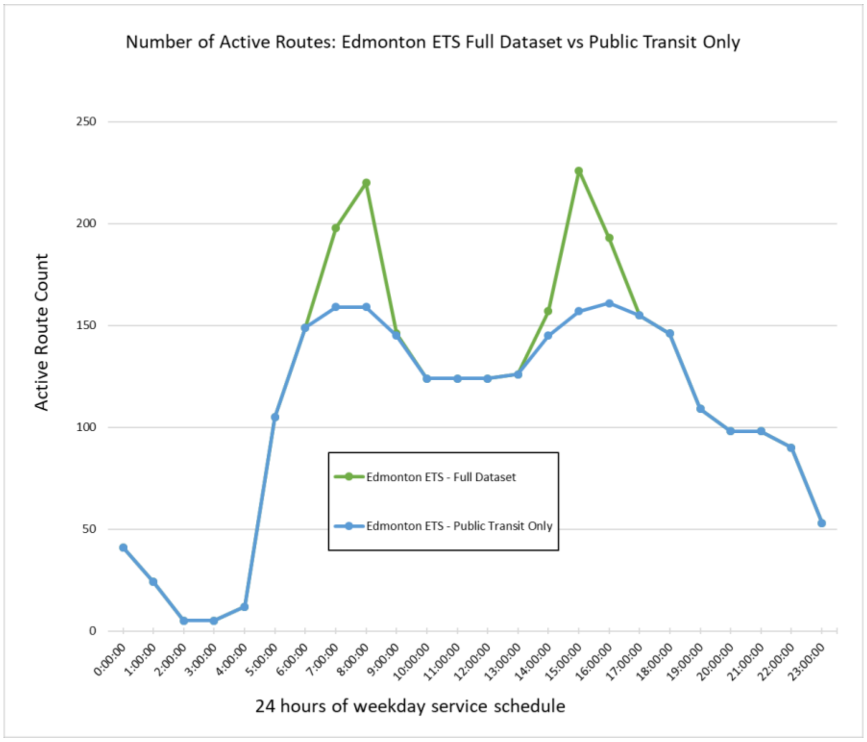


Figure 13: Number of Active Routes over 24 hours in Edmonton: full GTFS record vs Public transit lines only

The active route count, by itself, communicates very little regarding the change in degree of overlap throughout the day but when contrasted with the following measures it can provide valuable insight as to how the form of the network evolves through the day. It should also be noted that the active routes count does not take coverage area into consideration so an increase or decrease in active routes may not be correlated to changes in service levels or coverage.

4.2 Line Overlapping Index

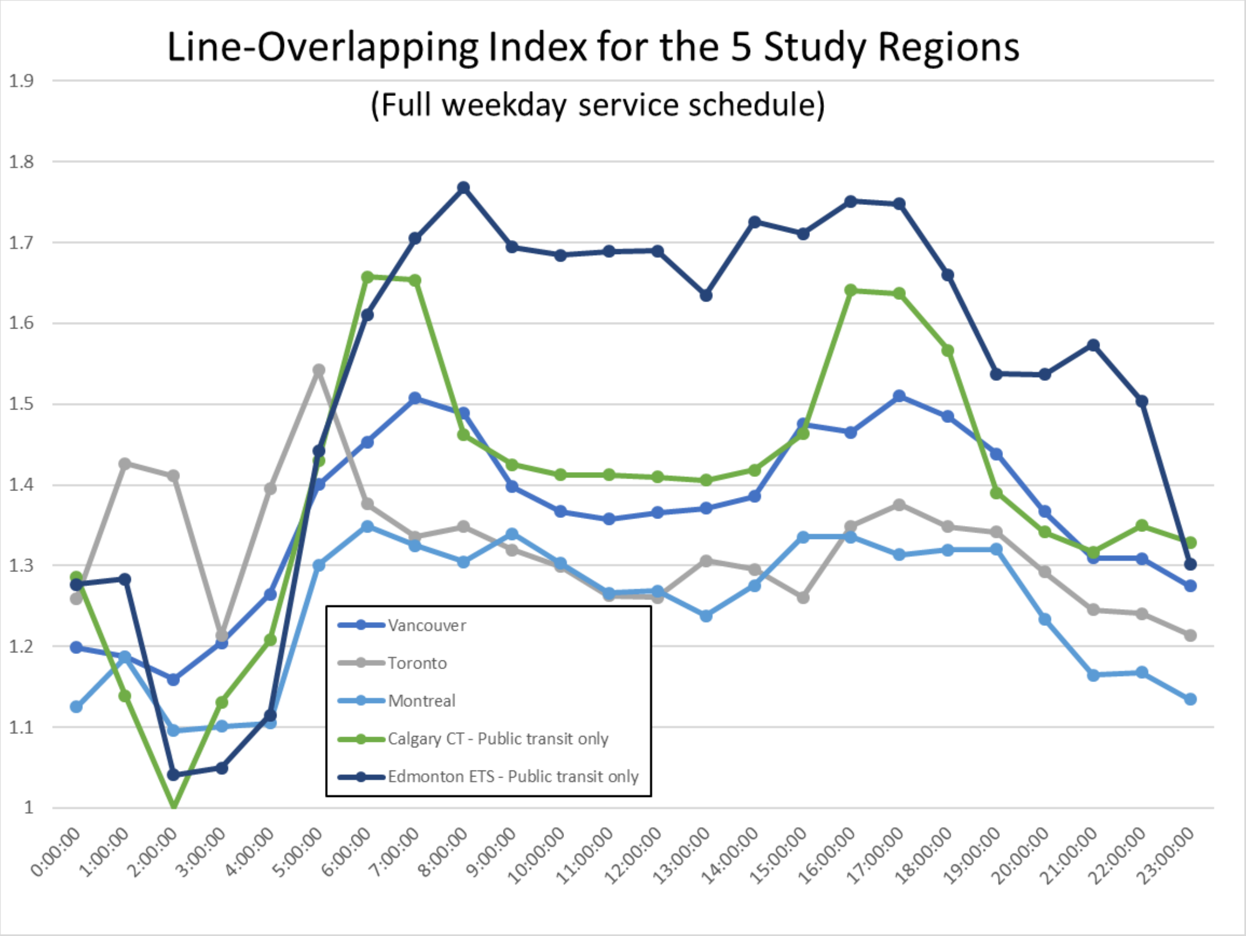


Figure 14: Line Overlapping Index scores of the 5 study regions over 24 hours

The resulting Line Overlapping Index graph (Figure 14) depicts the extent to which routes are covering common territory for each hour of the day, and therefore which hours of the day will be more likely to generate ambiguous trip data. Of note in these results is how only Calgary hits the base value of 1.0 at 2:00am, meaning all roads with routes were covered by only one route each, and none of those routes connect on the roadway for transfers (the recorded value was actually 0.9997… which is mathematically impossible for this equation and was therefore attributed to rounding errors encountered during the manipulation of coordinates). None of the other study regions hit the base value of 1.0 and the lowest reported values for each city occur in the early morning hours when the network is comprised mostly of night bus lines. Interestingly, Toronto's highest LOI values occur during night service hours due to the overall coverage length of the network drastically shrinking while the transfer portions of the network remain relatively fixed.

Edmonton displays the highest values of all networks during the mid-day period; examination of their network revealed many routes that converge on common links such as arterials, collector roads and bridges in the downtown region in particular, thus resulting in high degrees of overlap.

This index only examines the spatial correlation of the routes, meaning that similar to the active routes count, this measure does not consider the number of departures of each line. As such, this measure does not convey much information regarding the change of service levels or coverage area for the active routes. When compared to the active route count, when a time period experiences both an increase in active routes and increase in overlap index we can determine that portions of the new routes added to the network must be covering common territory. The GTFS shapes.txt was found to contain multiple linestring shapes for each transit route to represent the changes in their paths throughout the day. A route's path on the roadway can change at different times of the day to accommodate different transfer locations, terminus ingress/egress points, or even detours to service high schools at the end of the school day. In absence of the active route count, an increase in the line overlap index may in fact only be the result of the alteration of the shapes of the routes. In regards what these results reveal about route attribution rates; Toronto for example would have a decreased route attribution accuracy for trips recorded at 5:00am as compared to trips recorded at 11:00am, provided that trips were recorded on all portions of the network equally. In short, a high the degree of overlap found via this calculation translates to lower route attribution accuracy for TII processes that compare trips segments to GTFS data.

The presence of school bus routes in both the Edmonton and Calgary datasets impacted the LOI results with the largest difference occurring during morning peak hours. Amazingly, both networks achieve a LOI index above 2.0 meaning the total distance of all routes is more than double that of the actual roads covered by the routes. Calgary reaches a peak LOI of 2.13 and Edmonton reaches 2.10, both at 8:00am. The LOI values for 8:00am once filtered for public transit only are 1.46 and 1.76 respectively. The school routes account for over 900Km of network coverage divided between 112 school routes in Calgary, and 760Km of coverage in Edmonton, serviced by 74 routes.

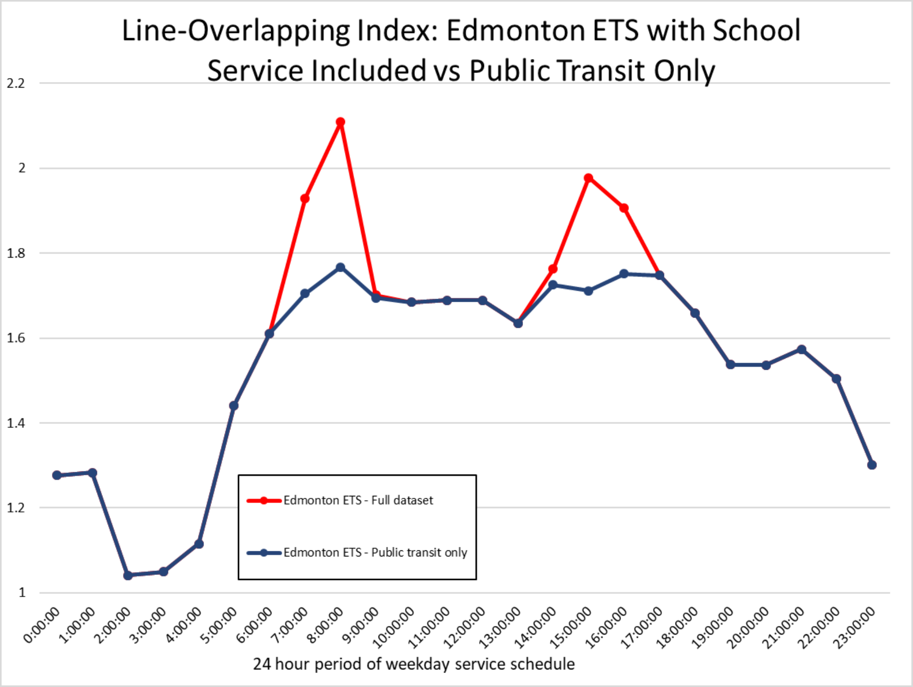


Figure 15: Line Overlapping Index: Edmonton ETS showing full GTFS dataset vs public transit only (school bus routes removed)

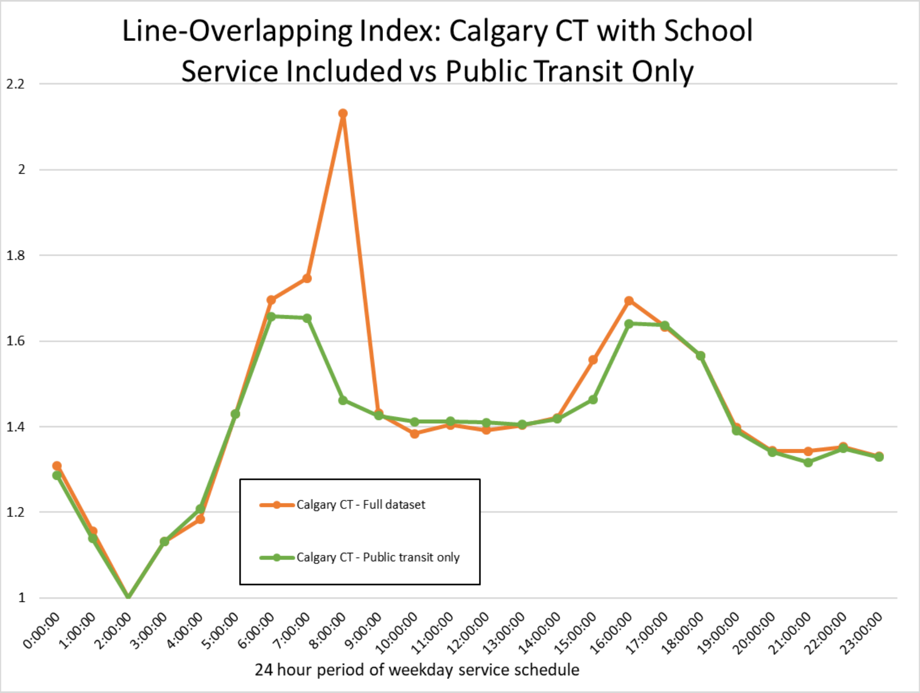


Figure 16: Line Overlapping Index: Calgary CT showing full GTFS dataset vs public transit only (school bus routes removed)

In the context of transit itinerary inference, higher LOI values translate to greater difficulty for TII procedures to match routes to GPS traces, but again, the spatially aggregate value only represents the theoretical maximum amount of overlap if trip data is recorded evenly throughout the entire network which may never occur in practice. These findings underscore the importance of filtering the GTFS feed for routes that pertain to a given procedure. The unfiltered datasets for both Calgary and Edmonton could result in reduced route attribution rates for travel survey trips recorded in each network. In fact, it was these extreme values when contrasted with the other regions that began the search that brought the school routes to our attention.

4.3 Overlapping Routes on Links (OROL)

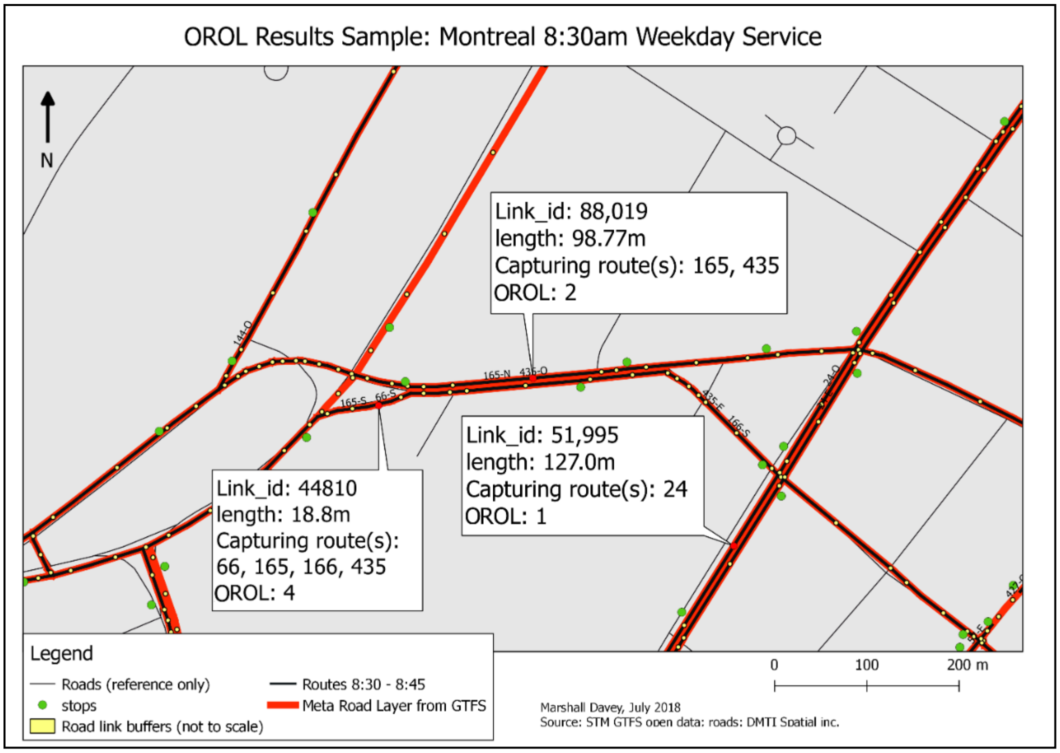


Figure 17: Sample of OROL results for Montreal showing the component layers of the OROL process

While the connection between LOI and OROL calculations is an interesting development in this research, the following discussion of OROL results will center on its standard form (overlap is any route count >=2) as this form directly addresses the problem of route overlap reducing route inference accuracy in TII.

The above map presents the OROL results as stored in a spatial database layer. Each text box depicts how a given link is recorded in the table. The spatial layer has link ids as primary key, an array containing all overlapping route numbers, the length value of the link, and finally the OROL count. Depicted in yellow, the buffers which capture route information during intersection procedures are formed around link centroids. The actual size of these buffers is so small that they would not be visible at this scale. Presented below in figure 18 is the complete results for all study regions. The Y axis percentage values represent the portion of the network consisting of links with overlap (OROL >= 2) as is consistent with the above described methodology.

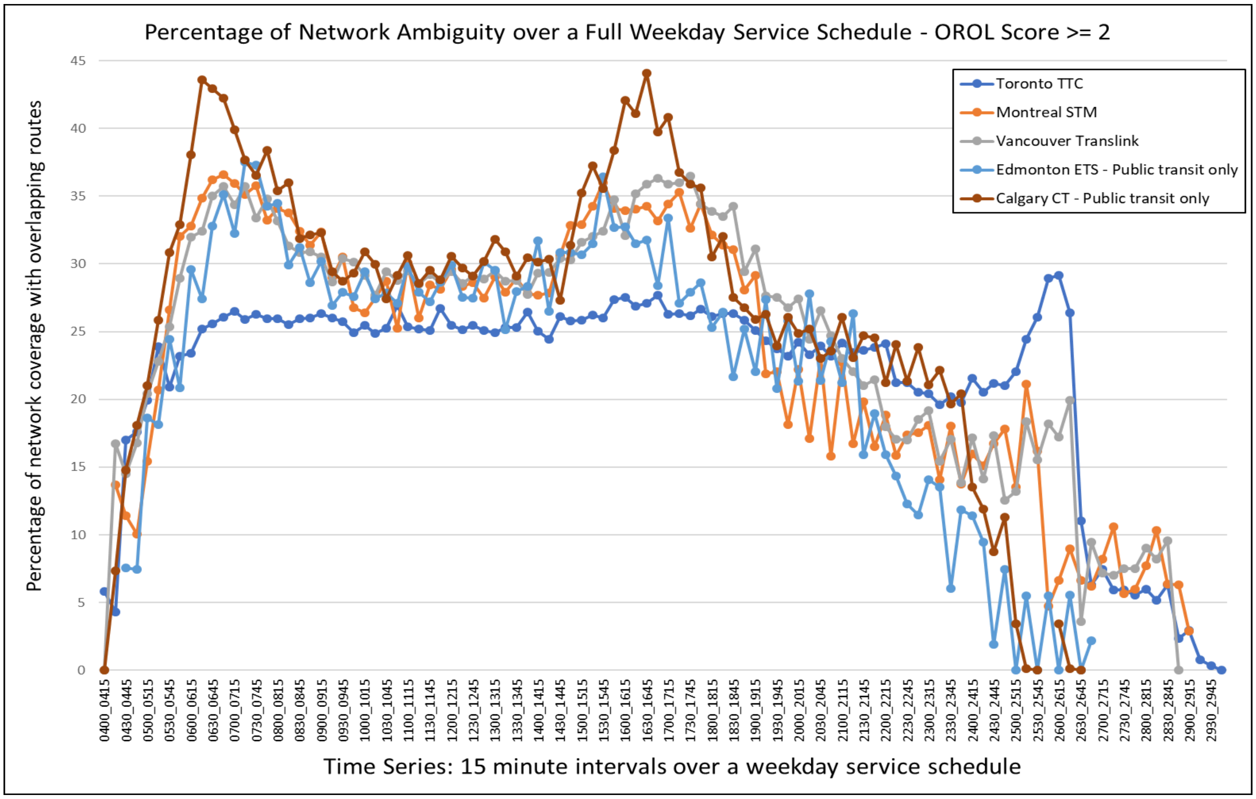


Figure 18: Overlapping-Routes-on-Links results: percentage of network length consisting of two or more overlapping routes

The Overlapping Routes on Links score expresses the percentage of network length that consists of two or more overlapping routes. This calculation is different than the Line-Overlapping index in three key ways: 1) it examines each individual road link and verifies the presence of routes, 2) it counts the amount of routes found on a link and records this value to a new column in the attribute record of the layer, and 3) it divides the entire coverage length into two categories – overlap: route count >=2, vs no overlap: route count = 1 (the route count limits can be set by the researcher). Each road link is categorized as either unambiguous, in the case of only *one* route present, or ambiguous, where *two* or more routes are present. The ranking of networks according to LOI score is different than the ranking when examined with the OROL methodology due to the fact that the components of the Line-Overlapping calculation inherently include the distance represented by both categories of links; both the denominator and numerator contain the lengths of links with only one route. The OROL procedure however, separates the two categories of links and sums the distance belonging to each.

Of note in these results are the relatively steady values for Toronto across the entire service schedule (Canada's largest city by metropolitan population, and 2nd largest by metropolitan surface area). This confirms the characteristic of the network that was first revealed with the Active Routes count and Line-Overlapping index; Toronto's service coverage area (620km^2) remains relatively fixed throughout the day in terms of number of routes and placement of the routes on the road network.

For each network, with the exclusion of Toronto, the portion of the network with overlap (y axis) experiences an up and down pattern from one time window to the next. This pattern results from more routes departing at the top of the hour and at 30 minutes into the hour, than the number of departures happening at 15 and 45 minutes past the hour – or in other words, the majority of the routes run every half an hour, while less routes run every 15 minutes. This up and down pattern reveals the implications the choice of examination period has on the analysis. If the OROL score was calculated every 30 minutes instead of every 15, the network would demonstrate a higher level of overlap than what is experienced on the network in real life. In the interest of transit itinerary inference, the GTFS data should be examined on as fine a spatiotemporal scale as possible.

Depicted below in figures 19 and 20, the OROL scores for Calgary and Edmonton were both impacted by the presence of school bus routes in the GTFS data. The time span of these graphs has been reduced to show only the differing portions. What these results reveal about the nature of the overlap is different than the LOI index in that when the LOI score increases above 2.0 it means the total network length is at least twice that of the geometric union, intuitively this may seem like an overlap rate of 100% or more (2 routes found on each road), yet the overlap percentage at the time of peak LOI is only 48.69%, indicating less than half the roads have overlap.

How is it possible for the total network length to be more than double the geometric union, while only experiencing overlap on 48.69% of the road length? Does this imply that the ~50% of the network with overlap actually has 3 routes on all links? By comparing the filtered vs non-filtered LOI and OROL results we can begin to understand how these school bus routes are added to the network. The LOI value dropped from 2.13 to 1.49 after removing school busses, while the OROL % went from ~46% to ~36%. From these values I can draw the following conclusions: 1) a 10% increase in overlap distance means some of the bus routes were added to roads which previously only had one transit route present, 2) since *adding* bus routes to the network can only produce a zero, or positive change in GU, and the LOI has increased, this implies N must have increased also, and by a larger degree than GU. Therefore, school routes added to the network *must* have greater lengths where the overlapping route count is 1. This is essentially a complex way of saying "new school bus routes were added to reach previously unserved regions, while portions of these routes will share roads with existing routes". This situation is similar to the 3rd sample network example on page 28 where the degree of overlap increases while OVR length remains fixed.

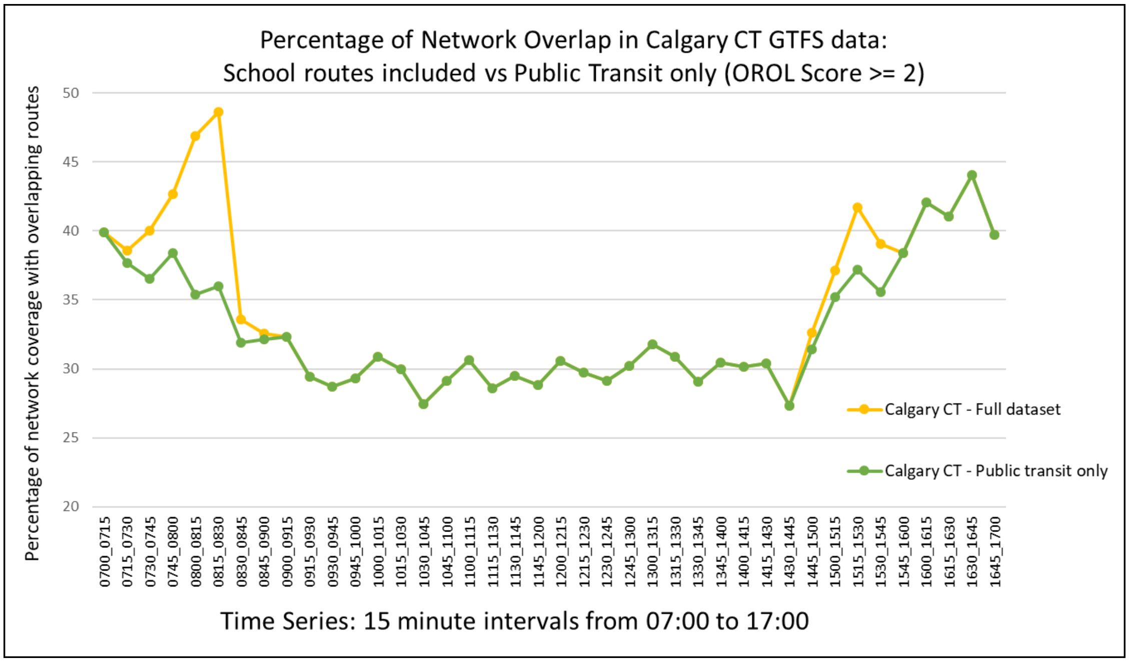


Figure 19: Overlapping-Routes-on-Links results: percentage of network length consisting of two or more overlapping routes for Calgary showing public transit service vs full service including school bus routes.

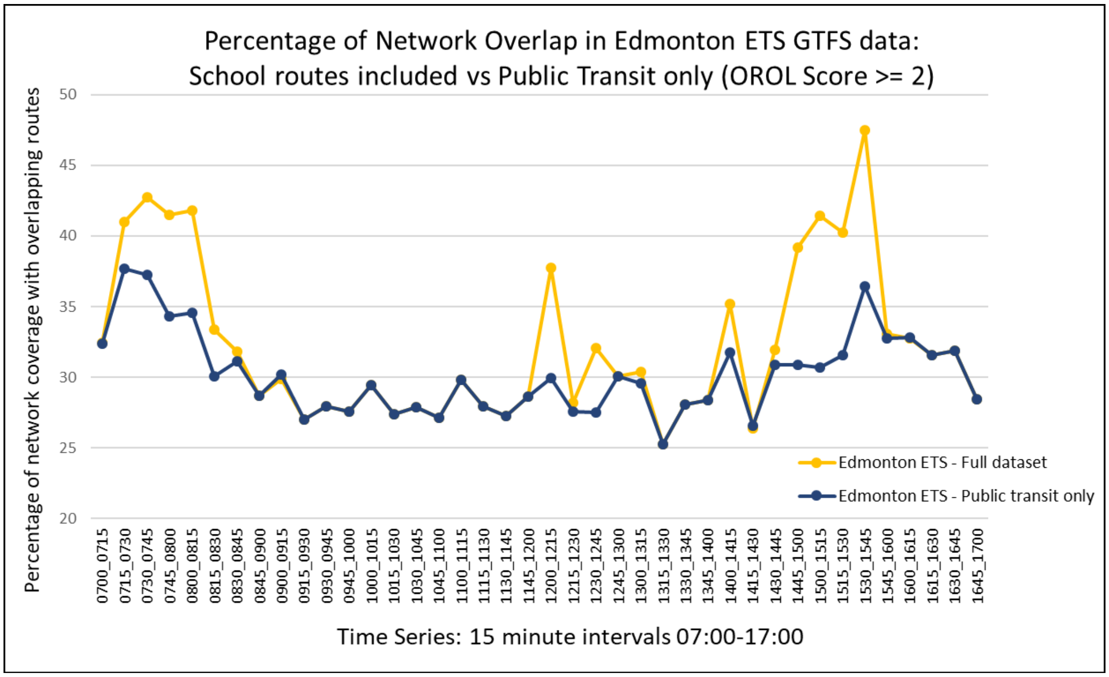


Figure 20: Overlapping-Routes-on-Links results: percentage of network length consisting of two or more overlapping routes for Edmonton showing public transit service vs full service including school bus routes.

The following four graphs present a summary of the information generated by OROL processing at the same time as demonstrating the analytical strength of the OROL approach to measuring overlap. The graphs are presented in pairs: Network length proportions of each category of overlap at 8:00am (Figure 21) are contrasted by the absolute length values (Figure 22) sampled at the same time, and then proportions vs actual lengths are again examined, but for the highest OROL% reported in each network, regardless of time period (Figures 23 & 24).

At 8:00am the network with the highest overlapping route value is Edmonton with 21 routes on one road link, however Figure 22 shows that this road link is only 1.5 meters long. While this single 1.5m road link with 21 overlapping route shapes seems like the most complex link in this group of results, it should be noted that Calgary's most complex link has 14 routes over a distance of 352 meters. This extended length of high overlap theoretically represents the collection of more rider GPS points in an ambiguous region of the network. Montreal has 16 routes on one link, Vancouver: 15, and Toronto and Calgary are tied with 14 routes on one link.

Returning to the earlier discussion of Calgary's high LOI values despite having the lowest active route counts; figure 21 shows how Calgary's percentage of length with only 1 route present is equal to that of Montreal's, however the absolute length of this category is actually much lower for Calgary than it is for Montreal. Comparing the absolute lengths of categories between Calgary and Montreal taken at the period of maximum overlap, we can see the categories of route count above or equal to two in Calgary have greater length than Montreal's, thus producing for the high LOI scores.

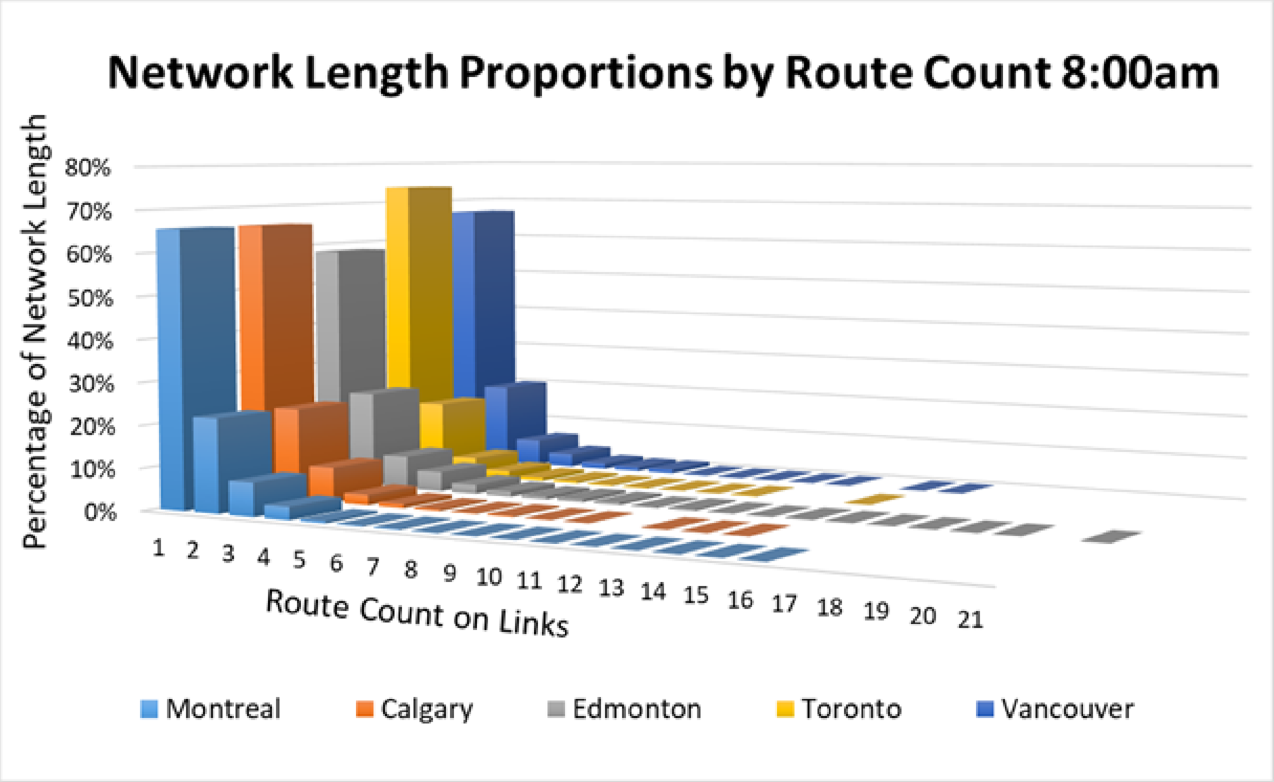


Figure 21: OROL Results: Network Length Proportions by Route Count 8:00am

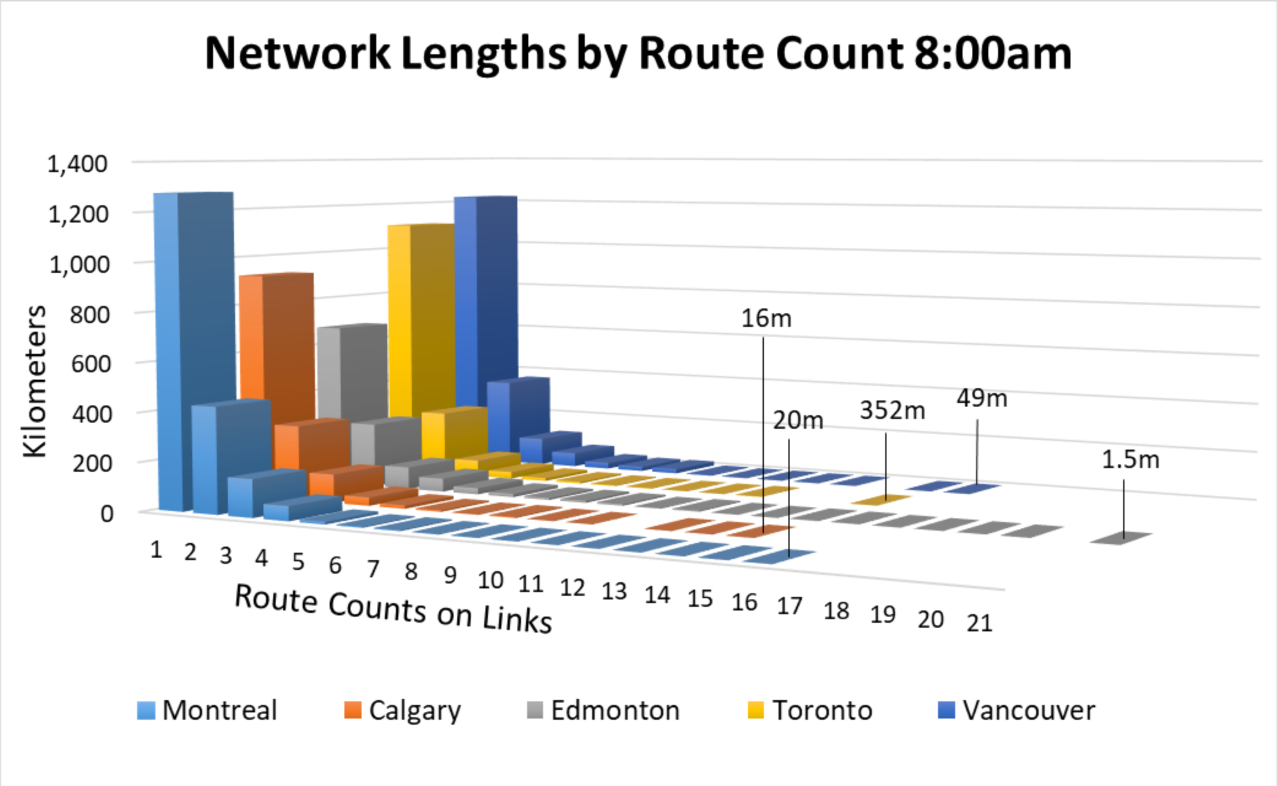


Figure 22: OROL Results, network lengths by route count: 8:00am

The graphs constructed from the maximum overlap values are presented to depict the theoretical "worst" arrangement of the network in regards to TII route matching. The time frame specified for each city represent the times at which a TII will have the most difficulty matching routes to GPS traces. Interestingly, Toronto's most complex network arrangement occurs at 2am, basically stating that trips recorded during peak service hours can actually have greater route attribution rates than trips recorded at 2am.

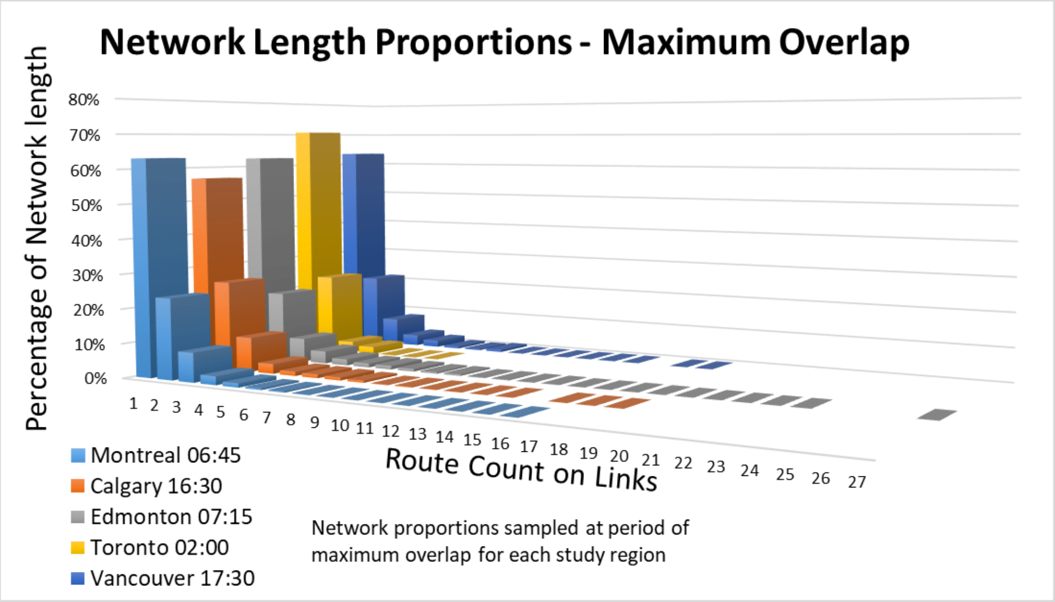


Figure 23: OROL Results, network length proportions at period of maximum overlap

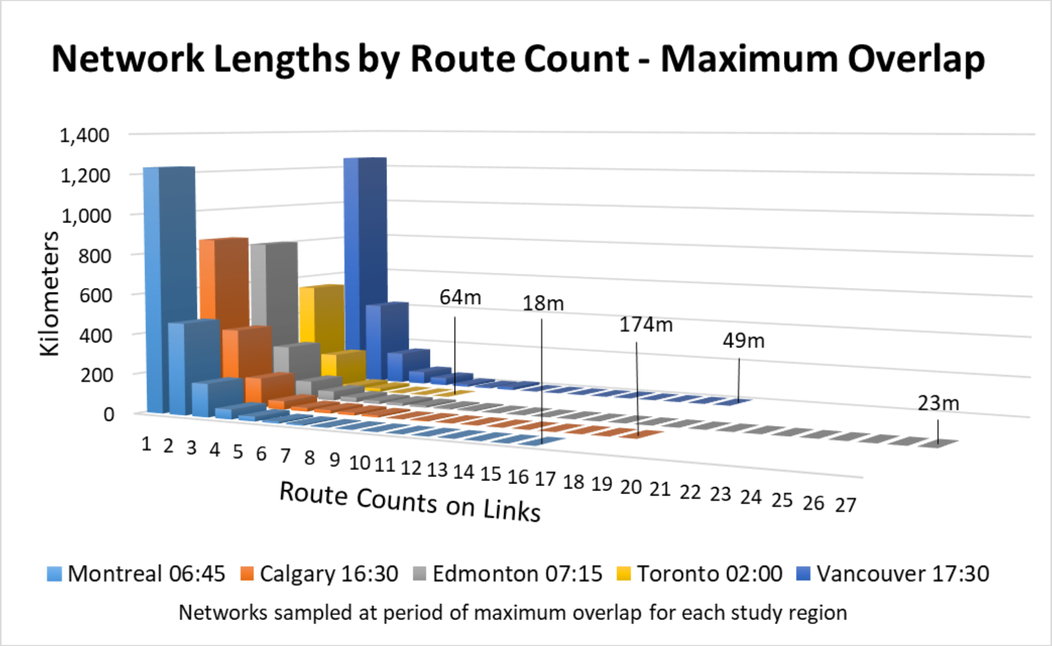


Figure 24: OROL Results: Network Lengths at period of maximum overlap

Presented below are the same Overlapping Routes on Links results as figure 18 on page 52, divided into three time segments: morning peak hour evolution (figure 25), midday and evening evolution (figure 26), and finally end of service day devolution (figure 27). The time segments have been divided for legibility of the graphs and according to the general evolution of each network. One would expect a clear pattern of low early morning service, peak service at morning and evening rush hours, and comparatively reduced service mid-day and late at night, yet all three metrics calculated thus far show that evening peak hour service level increases begin around 2pm.

Calgary's CT network reports the highest values, close to 45% in the morning and evening peak service periods. These results were closely inspected and reproduced to ensure the methodology was correctly expressing the overlap. Due to the varying structures of GTFS data from one agency to the next the SQL queries for several processes had to be edited to avoid the miss counting of trips and routes. After some inspection it was discovered the extreme OROL values reported for Calgary result from the Bus Rapid Transit lines that run exclusively during rush-hour and which follow longs paths along boulevards that coincide with regular service routes.

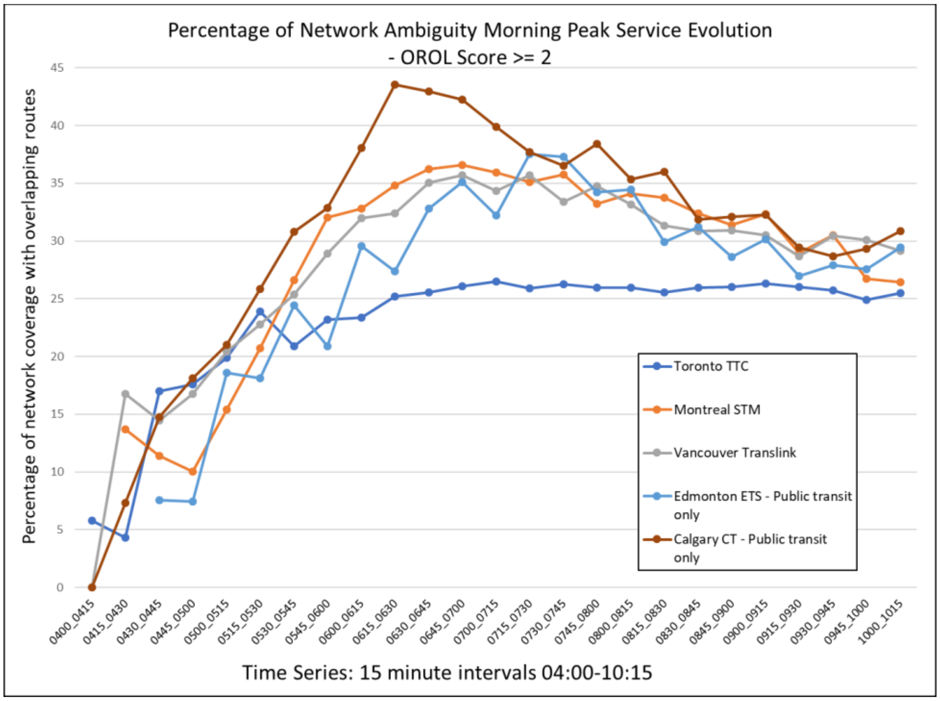


Figure 25: Overlapping-Routes-on-Links results: percentage of network length consisting of two or more overlapping routes

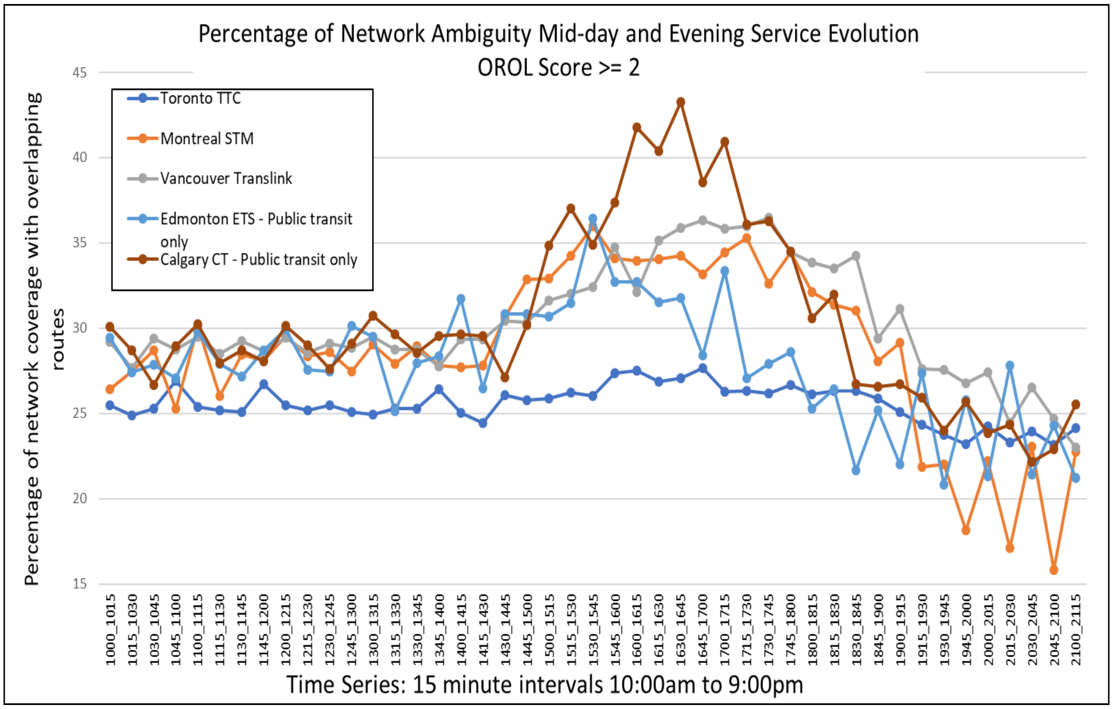


Figure 26: Overlapping-Routes-on-Links results: percentage of network length consisting of two or more overlapping routes

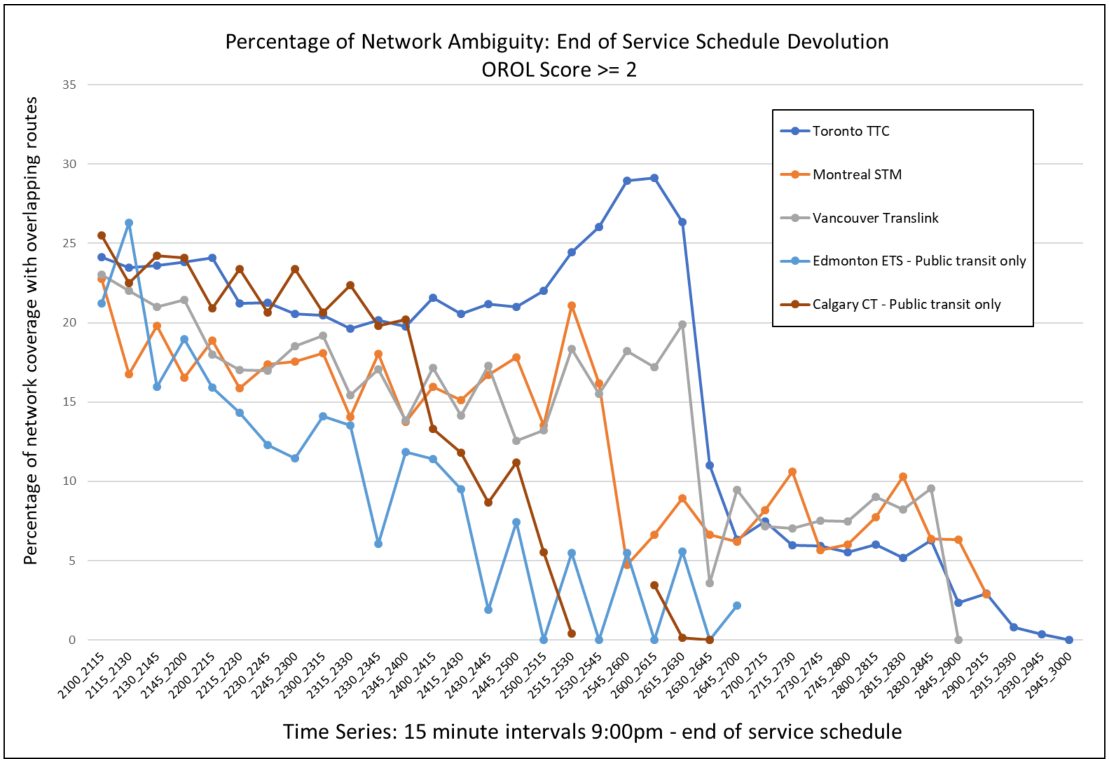


Figure 27: Overlapping-Routes-on-Links results: percentage of network length consisting of two or more overlapping routes

Also of note in these results are the zero values reported for several cities in the early morning hours. It was observed from the Line-Overlapping index that at no point over its sample periods did the network hit a 1.0 level of overlap; indicating that during each hour there was at least some instances of overlap. Yet the OROL score on the other hand does in fact report periods where there is 0% of the network length consisting of overlapping routes. This curious result is a direct result of the choice of examination period duration combined with the long headway times between departures on night service bus lines.

In the cities that report a degree of line-overlap in the early morning hours, yet no percentage of overlapping routes via the OROL calculation, it was discovered that the long headway times between departures of connecting lines will cause the static GTFS record to report that the lines *do not* overlap or connect within the 15-minute window. In other words, for a night bus trip that involves transfers, the second line departs more than 15 minutes after the first line is occupying its route. When examined over sample periods of one hour the bus lines appear to overlap in space, but when examined with shorter sample periods the OROL score reveals that these routes do not in fact coincide in space.

In the context of transit itinerary inference comparing smartphone travel survey data to the GTFS record, these results reveal the implications of time-window selection have on route attribution rates. Rider data compared to the one-hour window results will not result in reliable route inference in areas of overlap, while rider data compared to the 15-minute windows will be 'unambiguous'. These results underscore the importance of creating route comparison queries that attempt to match rider GPS points to the GTFS record on as fine a spatial and temporal scale as possible, a task that I have discovered to be resource intensive and often beyond the scope of projects that analyze large volumes of rider data. In order to facilitate such itinerary inference procedures with large volumes of rider data, it would be preferable to conduct a pre-screening of the GTFS data in order to reduce the level of route ambiguity found in the static GTFS record. The Probability of Passage score, discussed further below, allows for such reduction in route ambiguity by effectively producing new bus route layers that can substitute the GTFS layers used in the TII.

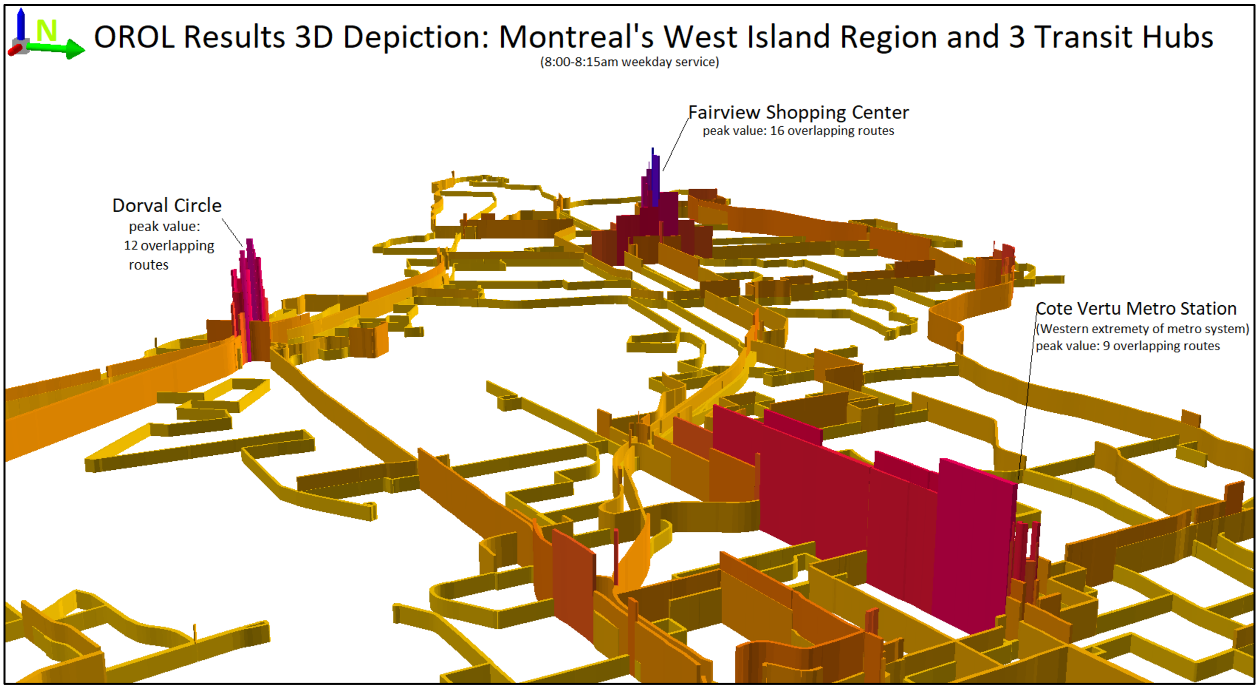

N
OROL Results 3D Depiction: Montreal's West Island Region and 3 Transit Hubs
(8:00-8:15am weekday service)
Fairview Shopping Center
peak value: 16 overlapping routes
Dorval Circle
peak value:
12 overlapping
routes
Cote Vertu Metro Station
(Western extremety of metro system)
peak value: 9 overlapping routes

Figure 28: 3D depiction of OROL results showing 3 transit hubs in western Montreal

In what is perhaps the clearest depiction of route overlap and how it changes with the form of the street network, figure 28 shows the route count as vertical elevation values for street shapes near three transit hubs located in Montreal's West Island. In the foreground, the arterial road approaching Cote Vertu metro station collects more routes the closer it is to Cote Vertu metro station, depicting how routes converge on the approach to the terminus. In the distance, the highest overlap value on record for Montreal is visible as a peak at Fairview shopping center. This single road link, discovered to be only 41 meters long can have as many as 19 routes passing through it in a 15-minute period during rush hour. Finally, the highway interchange known as Dorval circle, that connects transit riders to Montreal's international airport as well as a regional rail station, experiences a high degree of overlap where transit routes converge at a terminus. The ability to specifically locate instances of overlap within the network is an advantage this procedure contributes to the current body of literature over the existing Line-overlapping method. Should a researcher need to refine the GTFS record in order to reduce overall overlap, the OROL procedure will reveal which areas should be addressed and modified or "cleaned".

What follows below are maps showing a portion of each network in the study group with the classes of ambiguous vs unambiguous clearly defined. Following the observations made during the TII research conducted in Montreal, the results confirm that route ambiguity occurs most often in areas of high building density (downtown business districts) as well as on arterial and collector roads. The locations in the maps below were chosen to include the downtown district and to display several road links belonging to both classes (overlap vs no-overlap), the maps are included for reference only.

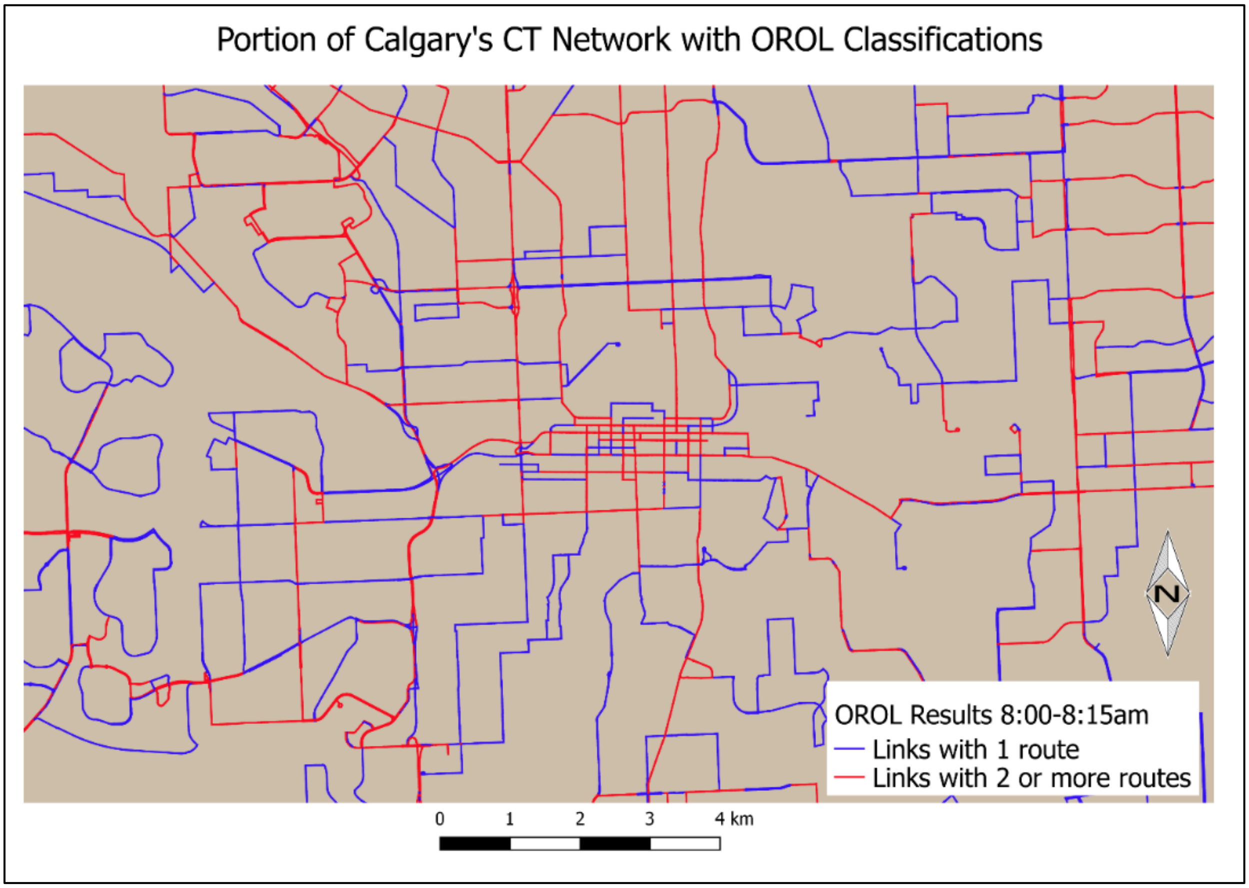

Portion of Calgary's CT Network with OROL Classifications
OROL Results 8:00-8:15am
Links with 1 route
Links with 2 or more routes
N
0
1
2
3
4 km

Figure 29: Portion of Calgary (CT) resultant OROL map

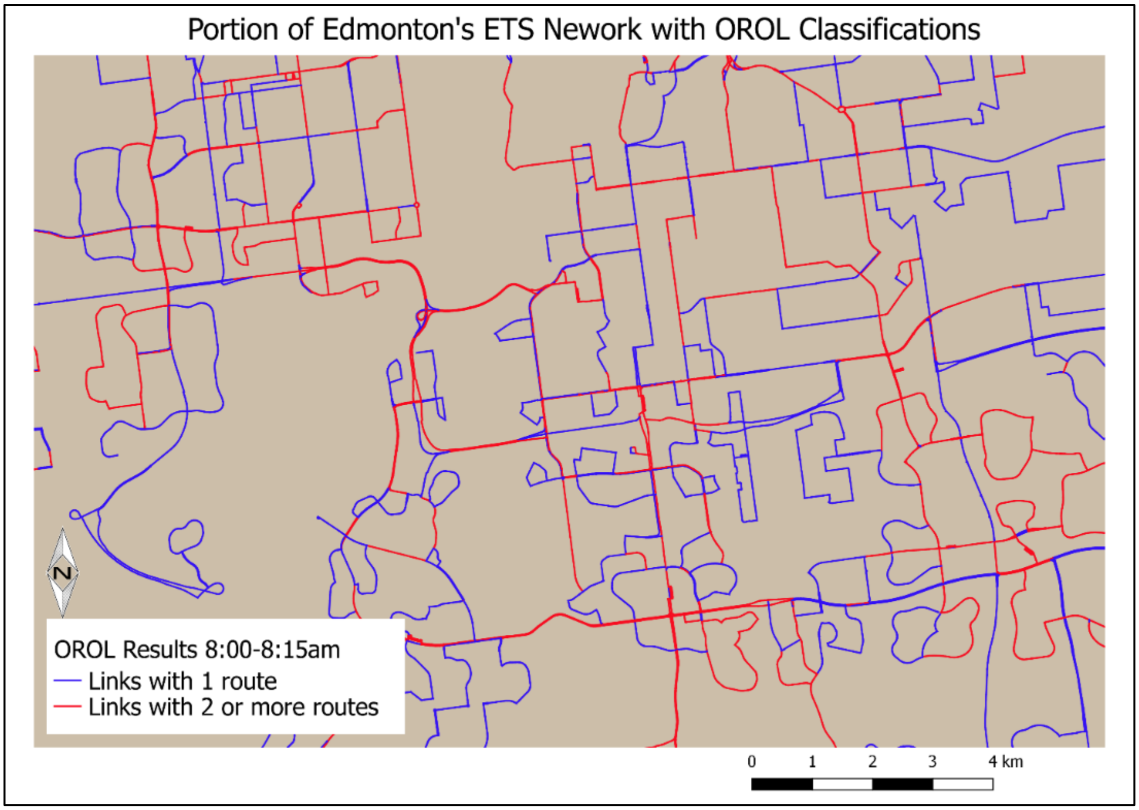

Portion of Edmonton's ETS Nework with OROL Classifications
N
OROL Results 8:00-8:15am
Links with 1 route
Links with 2 or more routes
0
1
2
3
4 km

Figure 30: Portion of Edmonton (ETS) resultant OROL map

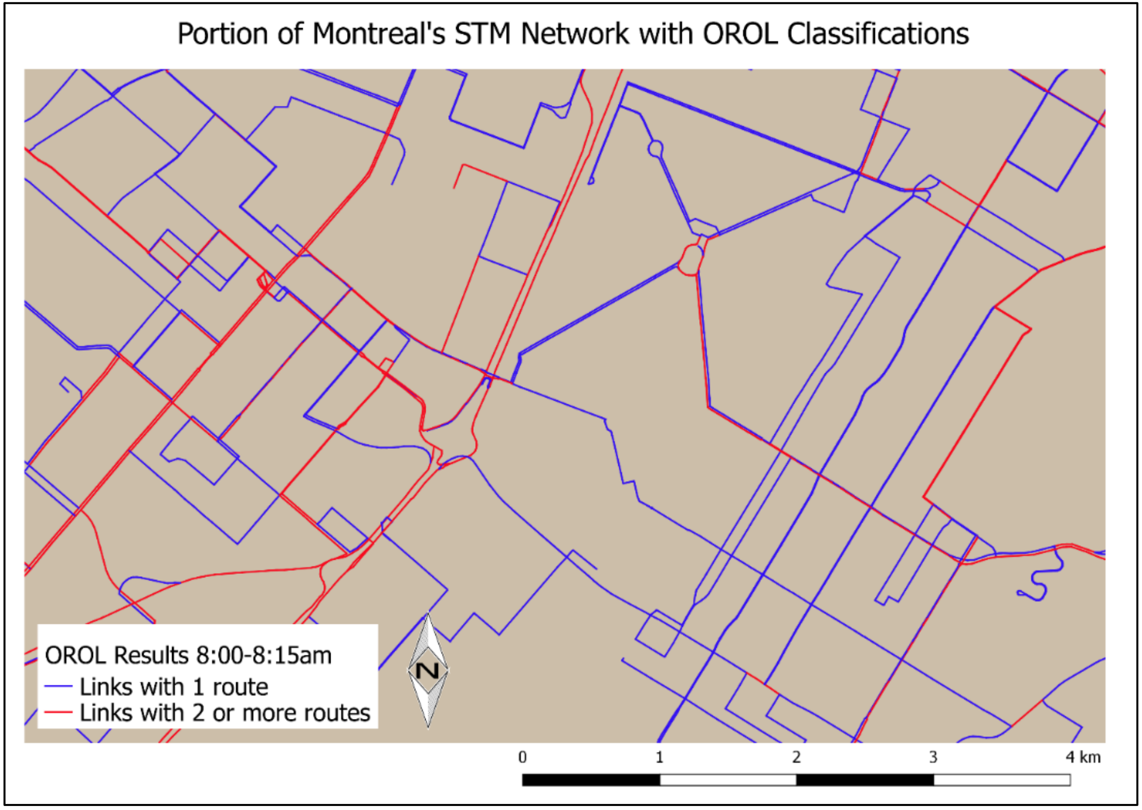

Portion of Montreal's STM Network with OROL Classifications
OROL Results 8:00-8:15am
Links with 1 route
Links with 2 or more routes
N
0
1
2
3
4 km

Figure 31: Portion of Montreal (STM) resultant OROL map

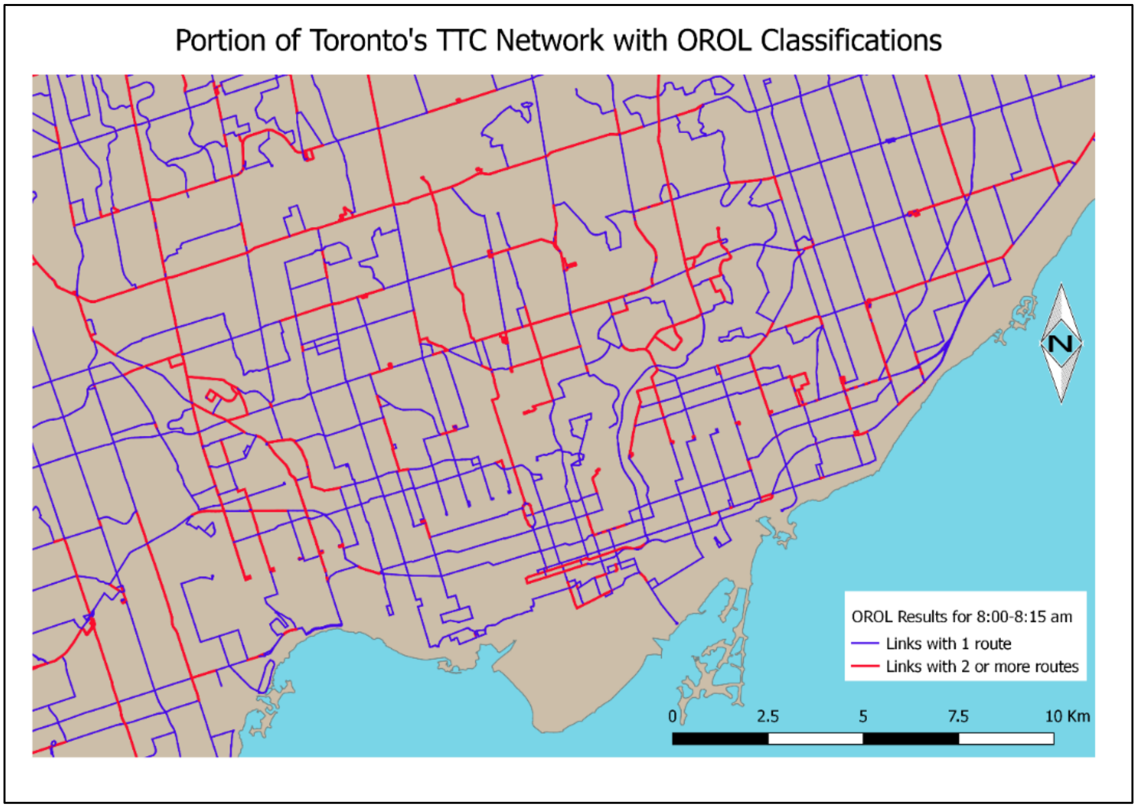

Portion of Toronto's TTC Network with OROL Classifications
N
OROL Results for 8:00-8:15 am
Links with 1 route
Links with 2 or more routes
0
2.5
5
7.5
10 Km

Figure 32: Portion of Toronto's (TTC) resultant OROL map

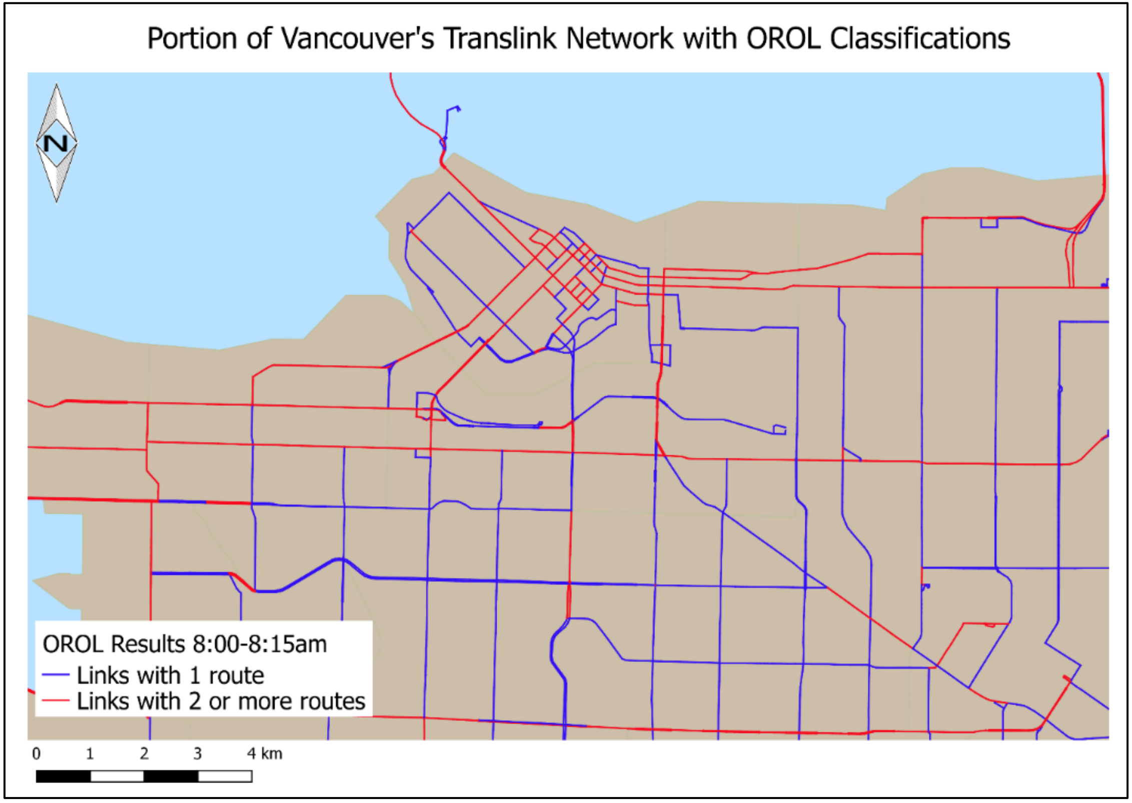

Portion of Vancouver's Translink Network with OROL Classifications
N
OROL Results 8:00-8:15am
Links with 1 route
Links with 2 or more routes
0
1
2
3
4 km

Figure 33: Portion of Vancouver's (Translink) resultant OROL map

4.4 Probability of Passage Score (POP)

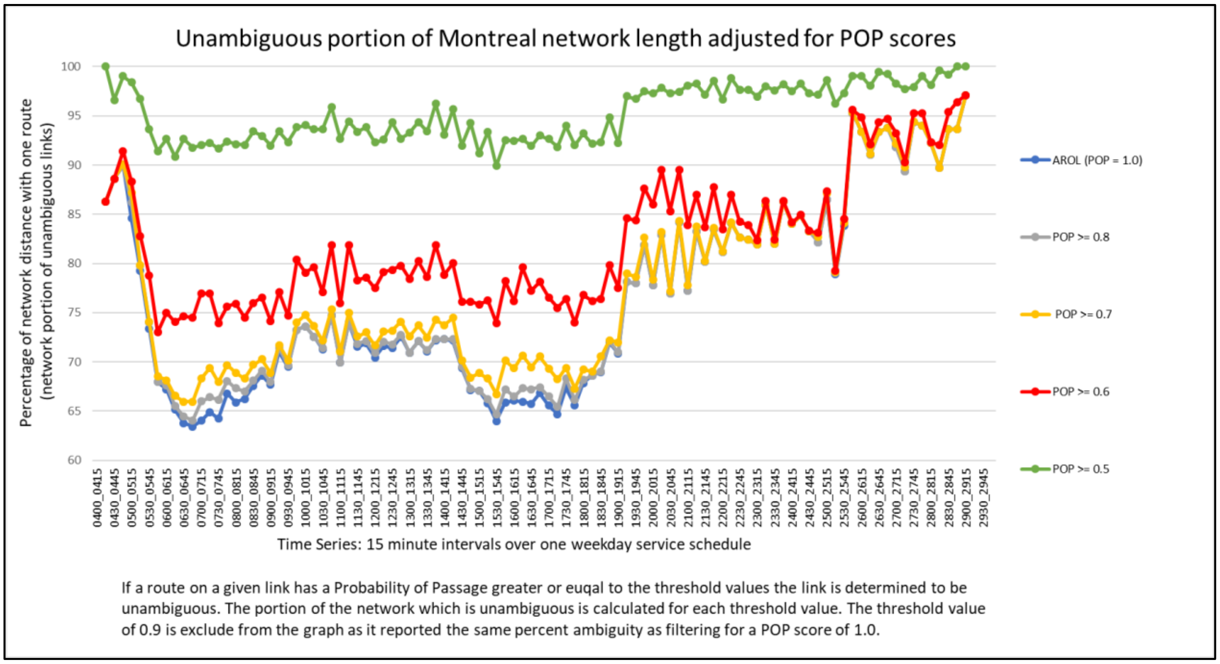
The Probability of Passage score (POP score) expands upon the OROL methodology by looking at the collection of overlapping routes recorded on each link and determining which route is most likely to be present during the 15-minute window. A POP score is generated for each route recorded on a link and is expressed as the ratio of that route's departure count over the total count of departures on that link. Once each route is assigned a POP score the network is filtered according to a POP threshold chosen by the user. Just as the OROL procedure produces a map with links categorized as ambiguous or un-ambiguous, the final step of the POP procedure allows for the same categorizing of links to be produced, but now rather than applying a Boolean categorization system, fuzzy logic can be introduced. A previously ambiguous link, now known to contain a route that passes 3 times more often than the others, can now be deemed unambiguous by the user by filtering for links with POP scores equal to or greater than that of the route in question. Thus, the POP methodology produces a more nuanced representation of the network by reducing the *ambiguity* level of the links flagged during the OROL process.

The graph presented below (figure 34) presents the inverse values of the OROL scores for Montreal. That is to say, where the OROL graphs communicate ambiguity, the POP graphs instead focus on un-ambiguity. Figure 34 shows how the unambiguous portion of the network increases as lower POP threshold values (α) are selected. As should be expected, filtering for a POP threshold of 1.0 produces the same network attributes as the Boolean approach used in the OROL procedure. For all study regions filtering for a POP value of 0.9 produced the same network as filtering for 1.0, for this reason the 0.9 results have been omitted from the graph.

With regards to TII procedures the graph demonstrates how pre-processing the GTFS data before employing for analysis can help overcome the hindrance of dense spatiotemporal routing information and effectively reduce the degree of overlap in the network. A theoretical result of 100% unambiguity across the entire schedule would allow for a TII procedure to correctly match routes to GPS survey trips with complete success. As is visible below, filtering the network for POP values equal to or above 0.5 produces the "cleanest" network depiction out of the alpha values selected for this graph.

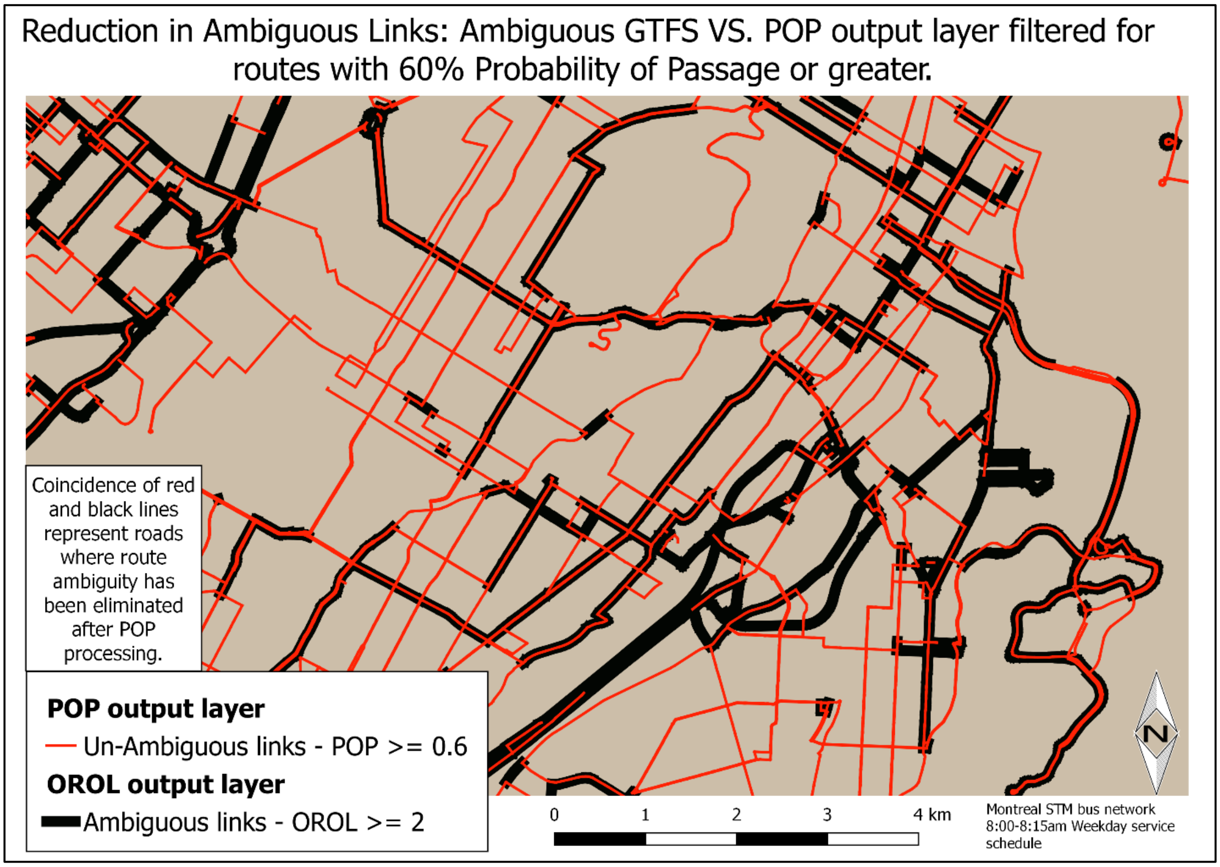
During the course of analysis, a range of alpha values from 0.1 to 1.0 were tested on the network, but after some consideration the decision was made to omit all values below 0.5 as the 0.5 links may very well represent a 50/50 chance of the algorithm selecting the correct route.

The POP of a link can be 0.50 if there are two routes with an equal number of departures, producing the above mentioned 50/50 odds of correct route selection, or the POP of a route can also be 0.50 if it has two departures, and shares the link with two other routes with one departure each (given the route in question a POP of 2/4). In this instance the 2/4 route is still the clear *winner* on the link since it has the highest POP (and therefore the most passages). Identifying which of the above overlap scenarios results in a 0.5 POP score on a link, however, is a procedure that is missing from the methodology at this moment. Identifying selecting winners based on POPs below this value become problematic as those values were found to occur most often on very crowded links. For example, a link exists in Montreal's STM network that has a total of 13 routes passing through it, of which the "winning" route contributes 3, giving this route a POP of 0.23. While this route is the winner according to this procedure, it is theorized that a 3/13 chance of correct route attribution will still hinder TII processes. When the temporal distribution of the 3 departures belonging to this route are considered, it seems even less likely that is the correct route based on the 3/13 ratio alone. One would assume that the 3 departures will be evenly spaced out during the 15 minute departure window, and given that this link is a single lane of road way, one can assume the other 10 departures are evenly spaced out over the entire 15 minute window. For the purpose of this example it should be noted that this link is at the exit of a terminus where routes for the western branch of Montreal's bus network begin their routes. With the winning route's 3 departures evenly distributed amongst the 10 others, the confidence in the selection of the winning route is diminished. For these above reasons the decision to limit a "trustworthy" POP score to a minimum of 0.5 was agreed upon and all future work employing such techniques should begin with network layers representing 0.5 and 0.6 POP scores to test their impact on TII procedures.



If a route on a given link has a Probability of Passage greater or euqal to the threshold values the link is determined to be unambiguous. The portion of the network which is unambiguous is calculated for each threshold value. The threshold value of 0.9 is exclude from the graph as it reported the same percent ambiguity as filtering for a POP score of 1.0.

Figure 34: Unambiguous portion of Montreal's STM network adjusted for different probability of passage scores (α)



Reduction in Ambiguous Links: Ambiguous GTFS VS. POP output layer filtered for routes with 60% Probability of Passage or greater.

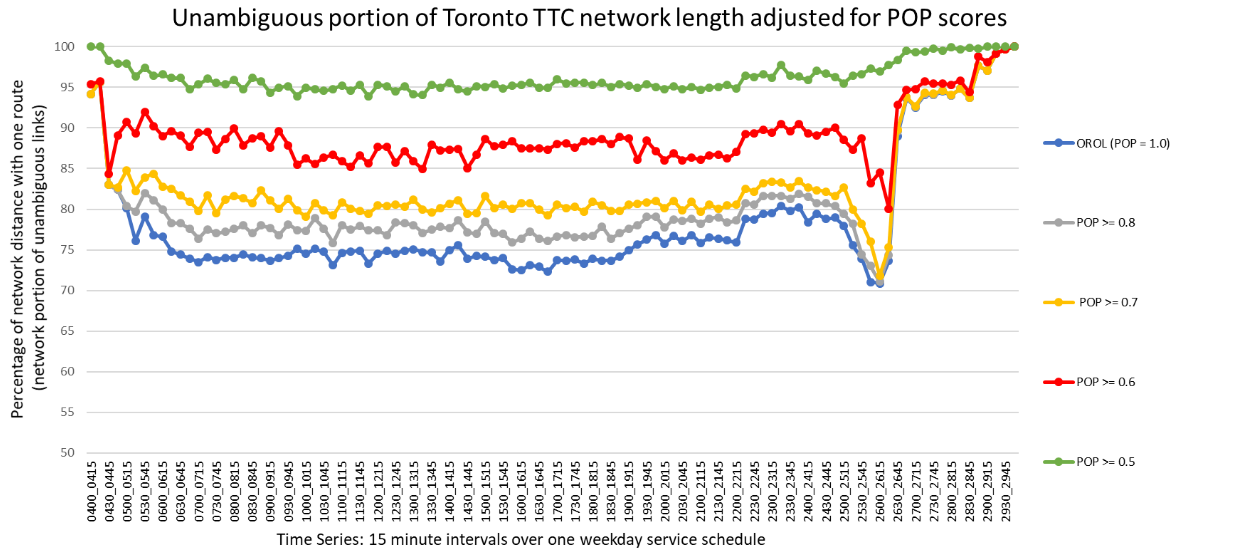
Coincidence of red and black lines represent roads where route ambiguity has been eliminated after POP processing.

POP output layer
Un-Ambiguous links - POP >= 0.6
OROL output layer
Ambiguous links - OROL >= 2
0
1
2
3
4 km
N
Montreal STM bus network
8:00-8:15am Weekday service schedule

Figure 35: Demonstration POP scores reducing route ambiguity when compared to OROL layers

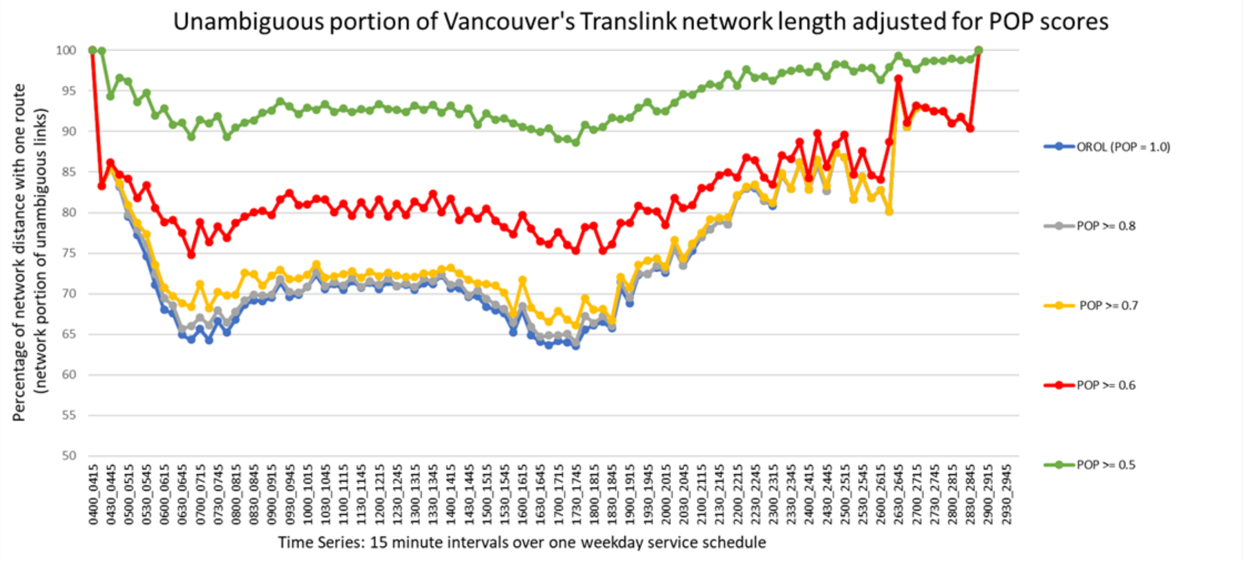
Figure 35 depicts the culmination of this research. By combining the OROL and POP output layers for Montreal during the 8:00-8:15am time window the map displays areas where POP processing has reduced the ambiguity of the network. The black lines represent road links that have deemed ambiguous with an OROL score >= 2, and the red lines show *un*-ambiguous links after having been filtered for POP scores >= 0.6. Where the two colors coincide represents roads where the ambiguity has been eliminated thanks to the POP score methodology. It should be noted that the complete route record as well as road way features are absent from this map for clarity, and only the two classes of links named above were included.

Included below are graphs depicting the degree of overlap reduction for each of the remaining study regions. As expected, the POP methodology reduced route ambiguity in the output spatial layers for all networks. For most regions with the exception of Toronto the difference in the portion of unambiguous network between the 1.0, 0.8, and 0.7 values is minimal, and the 0.6 results show a larger reduction in ambiguity. 0.5 produces the cleanest depiction of the network, but as noted above, the success rate of a TII using 0.5 input layers is open for experimentation.



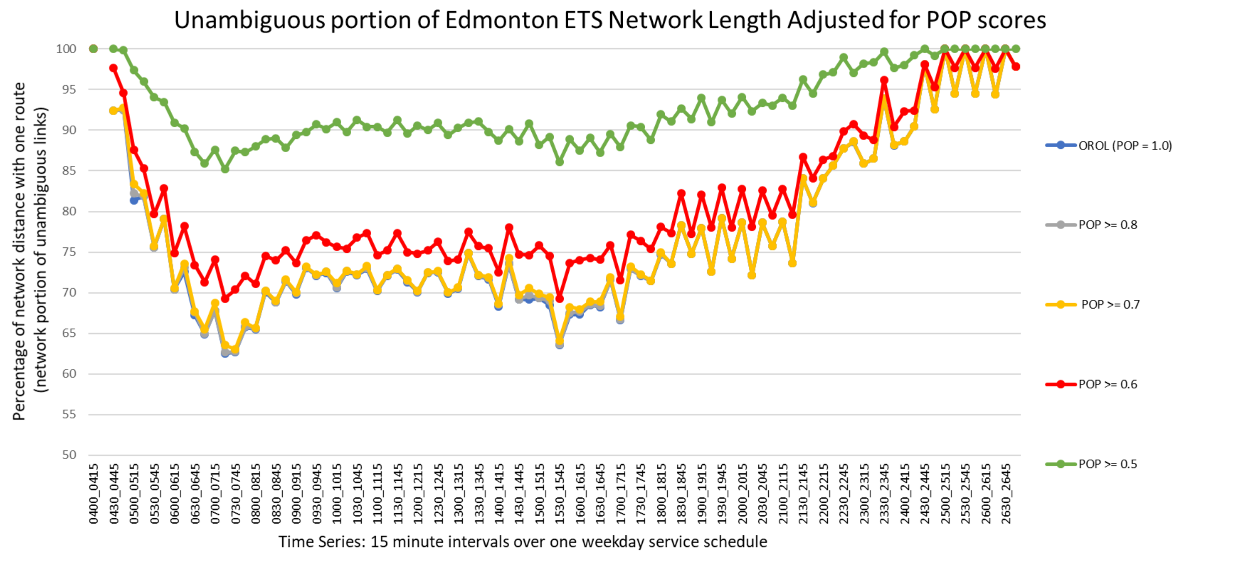
If a route on a given link has a Probability of Passage greater or euqal to the threshold value, the link is determined to be unambiguous. The portion of the network which is unambiguous is calculated for each threshold value. The threshold value of 0.9 is exclude from the graph as it reported the same percent ambiguity as filtering for a POP score of 1.0.

Figure 36: Unambiguous portion of Toronto's TTC network adjusted for different probability of passage scores (α)



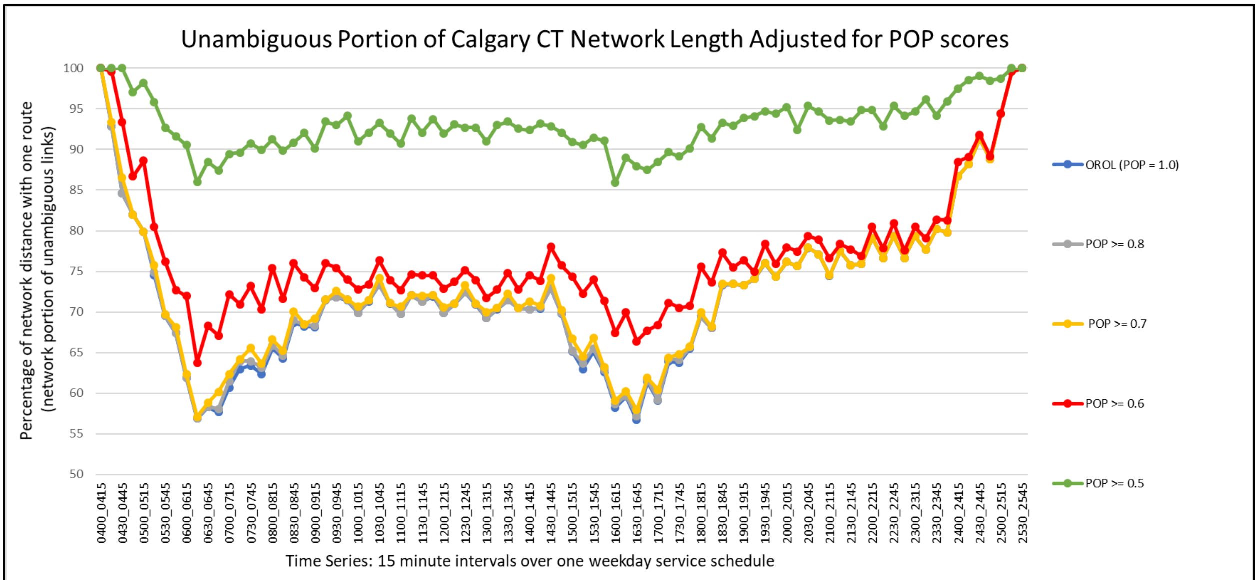
If a route on a given link has a Probability of Passage greater or euqal to the threshold values the link is determined to be unambiguous. The portion of the network which is unambiguous is calculated for each threshold value. The threshold value of 0.9 is exclude from the graph as it reported the same percent ambiguity as filtering for a POP score of 1.0.

Figure 37: Unambiguous portion of Vancouver's Translink network adjusted for different probability of passage scores (α)



If a route on a given link has a Probability of Passage greater or euqal to the threshold value, the link is determined to be unambiguous. The portion of the network which is unambiguous is calculated for each threshold value. The threshold value of 0.9 is exclude from the graph as it reported the same percent ambiguity as filtering for a POP score of 1.0.

Figure 38: Unambiguous portion of Edmonton's ETS network adjusted for different probability of passage scores (α)



If a route on a given link has a Probability of Passage greater or euqal to the threshold value, the link is determined to be unambiguous. The portion of the network which is unambiguous is calculated for each threshold value. The threshold value of 0.9 is exclude from the graph as it reported the same percent ambiguity as filtering for a POP score of 1.0.

Figure 39: Unambiguous portion of Calgary's CT network adjusted for different probability of passage scores (α)

4.5 Ranked Results

What follows below is a discussion of the results produced for each metric for each study region. Where applicable, extra graphs are provided to illustrate how the need to pre-filter the GTFS route types can impact the results of the analysis. Unique SQL route selection queries were scripted for Toronto, Calgary, and Edmonton to account for each agency's particular GTFS encoding and contents. Since this analysis is only concerned with the overlap of routes on roadways, Toronto's data was filtered to select all surface vehicles, i.e.: busses, bus rapid transit, and street cars. Similarly, underground metro systems were omitted from analysis since survey location data pertaining to metro trips can reliably be detected by trip breaking algorithms and/or itinerary inference procedures (Chen, Gong, Lawson, & Bialostozky, 2010; Zahabi et al., 2017). Calgary's and Edmonton's transit agencies both include school bus service in their GTFS datasets, coded with the same route type as regular bus vehicles (GTFS route type 3), the similar coding of these routes necessitates filtering the active routes lists using route id values to avoid the misrepresentation of the network. Without this pre-filtering of the data the results of each metric are exaggerated: Calgary's active route count, for example, reported 60% more routes at peak service, and a LOI over 2.0. A LOI over 2.0 means the distance covered by overlapping routes is at least equal to the length of the geometric union itself, this result was very surprising and spurred the investigation that revealed these alternate service types. Similarly, the inclusion of any metro route information will cause overlap to be counted on roads which may in fact have none as metro lines will be attributed to road links. Provided below is a summary of results with networks ordered from highest value to lowest. For each metric, the value reported below is the maximum value recorded over a full service schedule.

Number of Active Routes	Ranked Line Overlapping Index (max value over 24hr)	Ranked OROL Index (max value of examination periods)	Ranked ambiguous percentage POP < 0.6
Vancouver 206	Edmonton 1.76	Calgary 43.28%	Calgary 36.35%
Montreal 194	Calgary 1.65	Edmonton 37.50%	Edmonton 30.71%
Edmonton 193	Toronto 1.54	Montreal 36.57%	Montreal 26.99%
Toronto 170	Vancouver 1.51	Vancouver 36.47%	Vancouver 25.23%
Calgary 162	Montreal 1.34	Toronto 29.14%	Toronto 19.95%

Table 6: Ranked results

One might assume that the network with the most active routes would inherently contain the most overlap, yet the ranking of networks by active route count is different than the rankings according to overlap measures. As mentioned in the methodology, differences between study regions such as total network distance and the geography of each respective region produce network characteristics unique to each region and therefore an hypothesis such as the above cannot be assumed via active route counts alone. Figure 7 on page 21 reveals that Vancouver's network coverage is vast compared to the other study regions. This additional context helps explain how Vancouver can have the highest active route count while having the second to lowest OROL and POP percentages (many routes, little overlap.) Conversely, Calgary's second to highest LOI score indicates it contains more overlap than Vancouver, and this high degree of overlap is achieved despite having the lowest active route count of all the networks. This implies a greater degree of convergence exists between Calgary's routes when compared to Vancouver's; routes converge for greater portions of their distances, potentially allowing for more transfers within the system. Calgary's higher LOI score combined with lower route counts also suggests that its regions with overlap have higher degrees of overlap than the overlapping regions in Vancouver. The geographic distribution of each network backs up this assumption as with Vancouver we can observe long routes servicing periphery regions that overlap at the ends to facilitate transfers only. This is contrasted by the overlap experienced in Calgary's downtown region where 5 bridges feed routes to the downtown core. It was observed in both Calgary and Edmonton that the presence of bridges close to the downtown core contributes to higher degrees of overlap since even if few routes service these areas, the bridges oblige routes to share roadways for distances longer than what would be planned to facilitate transfers alone. While Vancouver does have bridges near its downtown core, the majority of its network is laid out away from the downtown region causing the few bridges close to downtown to have little impact on the LOI score. The hypothesis that Calgary's regions of overlap contain higher degrees of overlap than Vancouver's can be tested using the results of the OROL procedure since the resultant layers allow for the count of categories of links according to degree of overlap.

Comparing Calgary and Edmonton's LOI ranking to their OROL rankings also reveals an interesting difference between their networks. One should note how the rankings of Calgary and Edmonton are reversed when comparing the LOI to the OROL. Calgary reports a maximum overlapping length or 43.28% of its network, while Edmonton reports 37.5%. One might expect Calgary to have a higher LOI due to these results, yet it is in fact Edmonton that has the highest LOI. It is also interesting to note how Montreal's ranking changes from LOI to OROL. Both of these occurrences are related to how each measure incorporates the length of links with overlapping routes. With the LOI, the numerator contains the lengths of both categories of links: with overlap, and without, meaning the component of the equation that handles the overlapping length will always be longer than the length of overlap recorded via OROL. The OROL formulation effectively separates the two categories of lengths with their sum being the total network length. Since Montreal reports an OROL of 36.57%, yet has the lowest LOI of the group, we can determine that the *degree* to which routes overlap must be lower in Montreal than in the other networks. In other words, while Montreal has a longer distance of overlapping routes than Vancouver and Toronto, the average number of routes per instance of overlap is lower than the cities with higher LOIs.

Similarly, the change of rankings between Calgary and Edmonton indicates that Calgary has a greater distance of links with overlap, but on average the number of overlapping routes in those areas is lower than the average overlap count in Edmonton.

This illustrates a key difference between the LOI and OROL measures: LOI is concerned with *degree* of overlap averaged over the entire network, whereas OROL creates categories of links and allows for the calculation of length for each class. This is one of the key differences this approach brings to the current state of overlap measures.

The unchanged ranking between the OROL and POP is an expected result as it demonstrates that POP processing lowered the percentage of overlap by almost an equal amount for each network. With the exception of Calgary, each network's overlap was reduced by approximately 10%, Calgary's reduction was approximately 7.5%. This follows suit with the initial conclusion that Calgary's network must contain greater distances with higher degrees of overlap in comparison to the others: the smaller reduction in overlap compared to other networks indicates it has more links with a maximum POP score equal to or less than the 0.6 cut-off value employed in the study.

What follows below are concluding sections that summarize the contributions this research presents to the literature, the limitations of these methods discovered by the author, and finally, a perspectives section that outlines where this research should continue in the future.

5 Contributions

While the motivation for this work came from the need to understand the cause of reduced route attribution rates in TII procedures, the research conducted focused only on the development and verification of overlap measures and not on the testing of TII procedures themselves.

The necessity of consistent road layers between cities halted the research when erroneous results were discovered for each different type of road layers (OpenStreetMap, city Open data portals, DMTI Spatial inc). After hitting this procedural roadblock, efforts quickly turned to the development of the GTFS-to-roads procedure which occupied a large portion of the project's timeframe. Due to the lack of topological tools to properly address the myriad of geometric relationships between all the route linestrings, the project took on a new direction, evolving into an exercise in data-manipulation, GIS programming, and data verification procedures rather than a simple GIS overlap and counting procedure.

The overarching goal of the research was to provide analysts with a remedy to the route overlap issue faced in TII procedures so planners and researchers can extract the maximum value from survey data. The challenges faced during the development of these remedies instead steered the research into a new direction, evolving into an exploratory research project aimed at developing new GIS data handling techniques aimed at measuring network complexity with particular applications to GTFS linestring data, and eventually to the development of new network overlap measures.

It is still hoped that this research will go on to inform analysts how to increase the value retained from GPS survey datasets, which in turn will help planners make informed decisions once OD surveys are analyzed to recognize the needs of public transit users.

This research contributes a spatially disaggregate approach to network analysis that examines the overlap of network attributes via GIS procedures with the proposed Overlapping Routes on Links Score. In addition, this research offers a methodology for pre-processing GTFS datasets in order to reduce the degree of overlap of routes in the GTFS record via the proposed Probability of Passage score. This latter contribution has applications for transit itinerary inference methods from travel survey data when compared to the GTFS record. To conduct the analysis a novel python/SQL procedure was developed that converts a GTFS shapes.txt file into a simplified and accurate linestring GIS feature layer, replacing the need for a street network layer. This procedure has applications for any GIS analysis that works on clusters of linestring sample features and requires a baseline layer against which to measure them. While the process greatly benefits GTFS and network analysis in particular, the input features must not specifically be a depiction of a network.

The GTFS-to-roads procedure has been made publicly available via GitLab (https://github.com/TRIP-Lab/GTFS-to-Roads-converter) and offers a method of cleaning clusters of linestrings that is not readily available via the available open source GIS tools and software packages. In fact, developing this part of the procedure was the most time-consuming aspect of this research as no available GIS tools presented a reliable method of cleaning these linestring clusters.

An unexpected discovery during this research was how the OROL calculation can be repeated for different qualifications of overlap (2 or more routes, 3 or more routes, etc.) and the sum will equal the LOI score. The fact that a new overlap methodology can lead to the same scores enforces the conclusion that the developed methods are valid. The LOI index was developed at a time where network length calculations would be carried out using a scale-ruler and a paper map. The link-level analysis made possible by the OROL methodology would have been computationally inefficient if conducted during the 1980s. It is now thanks to modern GIS systems that this fine-scale view of the network can be generated.

6 Limitations

One limitation of the OROL methodology is related to how the SQL queries handle the bus route linestring shapes. In its current form, the GIS overlay procedure that intersects the active routes with road links recognizes a bus route's location to be *active across its entire length* when a departure is recognized at its first stop. This query structure was chosen to facilitate looping through the time series values as dictated by SQL formatting, as well as to follow the methodology of the TII pilot study that validated the Montreal survey data. In the query's current form, if two bus route shapes overlap far into their routes, in a location that may not actually be occupied by both busses at the same time due to each route requiring a different amount of time to reach that location, the current procedure will still recognize these routes as overlapping. Even if no overlap actually occurs on that road link at that time, the OROL procedure will report this as an instance of overlap. Different methods of addressing this short falling were attempted but it was determined that the computation resources required to refine the view of the network on such a scale were beyond the scope of this research. The most accurate querying procedure as conceived by the author is outlined in the meta-code below:

```
For a given time window:
    Identify stops with departures during this window.
    For each stop:
        Intersect this stop with the nearest road link
        Identify the road links between this stop and the proceeding stop
        IF the proceeding stop has a departure within the same window:
                Continue identifying consecutive links until another stop is reached and repeat IF statement.
        Else the proceeding stop has a departure outside the time-window:
                Assign each link between stops an “active route” classifier and write route info to the link.
Perform the regular OROL overlay and tabulating process to the above results.
```

This approach was attempted and abandoned once the processing demands of this method were realized. The task of identifying active road links between stops requires the creation of far more spatial layers than the OROL method outlined in this paper. Each active route needs to be broken into a sub-set of linestrings existing only between stops, effectively chopping and rebuilding the route shapes into new layers for each time window. On a typical service schedule, using 15-minute increments, there are 104 time windows to examine. To give an example of how much spatial data this query will generate, if a given time frame has 100 active routes, and on average each route has 20 stops, the result of this sub set process would produce 100 x 19 linestring shapes x 104 time windows, producing a total of 197,600 spatial layers (and/or SQL “windows”) simply to begin the analysis. After this step, buffers will need to be constructed around each stop to ensure intersection with each line string and then each of these 197,600 layers would need to be intersected with the stops, of which Montreal has 9,194. Only after this step can the route intersection process begin as described in the methodology section, which itself generates even more spatial layers. After much discussion it was determined that using the first outlined method, where a route is “active” across its entire length at the moment of departure, would be the most feasible approach, and also preferable for this research since it follows the methodology of the TII experiment in Montreal.

This approach was also abandoned due to the fact that the real time GTFS records (GTFS-r) provides a record of precise vehicle locations. Typically delivered in real-time via a web API, the GTFS-r data feed can be compiled into records representing a past date’s particular service. In regards to transit itinerary inference, obtaining the specific record of where each vehicle was on the given day a survey trip was recorded should greatly improve route attribution rates. I do not expect the GTFS-r data format to completely eliminate the challenge posed by route overlap however as research cited in the manuscript has already acknowledged overlap to be a problem in Active Vehicle Location data. As will be discussed further below in the perspectives section, performing TII with real-time GTFS data is a domain of transportation research that deserves much attention from scholars in the field as its applications for producing OD survey matrices are vast.

The limitations to the POP methodology are largely related to what has been described above. For example, when a given road link has a route with a POP of 0.5, and two other routes with POP scores of 0.25 each, the assumption that the first route has the highest probability of passage is based on the assumption that the bus is present on all related links at the moment of its departure.

Even given this limitation, the research conducted for this thesis demonstrates that filtering the network for specific POP values can reduce the degree of overlap in the GTFS record and it is my belief that applying the POP methodology to GTFS-r data will continue to diminish its already lower degree of overlap.

Although the limitation presented by this query structure results in an unrealistic depiction of network activity, it should be recognized that these processes are designed to reveal attributes present due to the way in which the GTFS data *is structured and packaged.* The measures are not intended to present a pure reflection of what happens on the roadway since the data does not permit for this. This aspect of the data handling underscores how GTFS is not a perfect data-structure for analysis in that the goal behind its construction is the reliable functioning of scheduling applications, not a precise depiction of the activities within the network. This project began with what should have been a simple GIS overlay procedure to produce a total count of overlapping routes, but it quickly turned into a data-management problem as the clustered nature of the routing shapes hindered accurate analysis, and the query structuring allowed for only certain structures of output results. What these results ensure is that if a GTFS feed is compared to travel survey data on a route by route basis, and the input GTFS data is not processed in any way to facilitate processing, then the methods described hold strong.

7 Perspectives

While real-time GTFS (GTFS-r) seems to be the ideal data source for TII procedures not all agencies have equipped their fleets with location tracking devices, and at the time of writing this over 1020 transit agencies are publishing static GTFS records. For this reason I believe static GTFS still represents a valuable data-source to researchers and methods built upon it should be explored further.

Future research that may result of the findings shared in this thesis should include the application of the POP output feature layers to TII processes in order to validate the method's impact on route attribution success rates in TII. If a correlation between network overlap and TII route attribution rates is to be developed additional validate survey trip data will need to be gathered for different regions. With the recent adoption of smartphone travel survey applications for as a method of conducting OD surveys, and some of this data making its way to the public, I hope that such work will be possible in the future.

To further explore the network density/overlap connection I feel that additional overlap indicators can be developed. I believe these indicators would benefit from integrating various surface area calculations and ridership counts at that such depictions of networks might even be able to portray *rider congestion* levels. By examining past real-time vehicle location records with ridership counts and even chip card alighting data I believe service models can be developed that will be able to predict rider congestion at stations as disturbances to the networks happen in real time. Agencies have long recognized the need to adjust their scheduling on an ad-hoc basis to accommodate changes they could not predict during the planning phase. Perhaps with the right overlay procedures and overlap measures statistical modelling will be able to produce better ad-hoc scheduling and adjust service levels on the fly to better address rider congestion before it even develops.

8 References

Allen, G. W., & DiCesare, F. (1976). Transit Service Evaluation: Preliminary Identification of Varaibles Chracterizing Level of Service. *Transportation Research Record*, *606*, 41–47.

Alter, C. H. (1976). Evaluation of public transit services: the level of service concept. *Transportation Research Record*, *606*, 37–40.

Antrim, A., Barbeau, S. J., & Others. (2013). The Many Uses of GTFS Data - Opening the Door Transit and Multimodal Application. *Location-Aware Information Systems Laboratory at the University of South Florida*, 1–24. https://doi.org/10.1.1.391.5421

Birant, D., & Kut, A. (2007). ST-DBSCAN: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering*, *60*(1), 208–221. https://doi.org/https://doi.org/10.1016/j.datak.2006.01.013

Carrel, A., Lau, P. S. C., Mishalani, R. G., Sengupta, R., & Walker, J. L. (2015). Quantifying transit travel experiences from the users' perspective with high-resolution smartphone and vehicle location data: Methodologies, validation, and example analyses. *Transportation Research Part C: Emerging Technologies*, *58*, 224–239. https://doi.org/10.1016/J.TRC.2015.03.021

Catala, M., Downing, S., & Hayward, D. (2011). Expanding the Google transit feed specification to support opertions and planning.

Chen, C., Gong, H., Lawson, C., & Bialostozky, E. (2010). Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the New York City case study. *Transportation Research Part A: Policy and Practice*, *44*(10), 830–840. https://doi.org/10.1016/j.tra.2010.08.004

Derrible, S., & Kennedy, C. (2011). Applications of graph theory and network science to transit network design. *Transport Reviews*, *31*(4), 495–519. https://doi.org/10.1080/01441647.2010.543709

Fielding, G. J., Glauthier, R. E., & Lave, C. A. (1978). Performance indicators for transit management. *Transportation*, *7*(4), 365–379. https://doi.org/10.1007/BF00168037

Fortin, P., Morency, C., & Trépanier, M. (2016). Innovative GTFS Data Application for Transit Network Analysis Using a Graph-Oriented Method. *Journal of Public Transportation*, *19*(4), 18–37. https://doi.org/10.5038/2375-0901.19.4.2

Gordon, J., Koutsopoulos, H., Wilson, N., & Attanucci, J. (2013). Automated Inference of Linked Transit Journeys in London Using Fare-Transaction and Vehicle Location Data. *Transportation Research*

Record: Journal of the Transportation Research Board, *2343*, 17–24. https://doi.org/10.3141/2343-03

Group, K. F. H. (2013). Transit capacity and quality of service manual.

Hadas, Y. (2013). Assessing public transport systems connectivity based on Google Transit data. *Journal of Transport Geography*, *33*, 105–116. https://doi.org/10.1016/j.jtrangeo.2013.09.015

Liao, F., & van Wee, B. (2017). Accessibility measures for robustness of the transport system. *Transportation*, *44*(5), 1213–1233. https://doi.org/10.1007/s11116-016-9701-y

Lu, Y., & Liu, Y. (2012). Pervasive location acquisition technologies: Opportunities and challenges for geospatial studies. *Computers, Environment and Urban Systems*, *36*(2), 105–108. https://doi.org/10.1016/j.compenvurbsys.2012.02.002

Musso, A., & Vuchic, V. R. (1988). Characteristics of metro networks and methodology for their evaluation. *Transportation Research Record*, *1162*, 22–33.

Nassir, N., Khani, A., Lee, S. G., Noh, H., & Hickman, M. (2011). Transit Stop-Level Origin–Destination Estimation through Use of Transit Schedule and Automated Data Collection System. *Transportation Research Record*, *2263*(1), 140–150. https://doi.org/10.3141/2263-16

Nitsche, P., Widhalm, P., Breuss, S., & Maurer, P. (2012). A Strategy on How to Utilize Smartphones for Automatically Reconstructing Trips in Travel Surveys. *Procedia - Social and Behavioral Sciences*, *48*, 1033–1046. https://doi.org/10.1016/j.sbspro.2012.06.1080

Nour, A. O. (2015). Automating and Optimizing a Transportation Mode Classification Model for use on Smartphone Data.

Shafique, A. M., & Hato, E. (2016). Travel Mode Detection with Varying Smartphone Data Collection Frequencies. *Sensors* . https://doi.org/10.3390/s16050716

Shen, L., & Stopher, P. R. (2014). Review of GPS Travel Survey and GPS Data-Processing Methods. *Transport Reviews*, *34*(3), 316–334. https://doi.org/10.1080/01441647.2014.903530

Thiagarajan, A., Biagioni, J., Gerlich, T., & Eriksson, J. (2010). Cooperative transit tracking using smart-phones. *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems - SenSys '10*, 85. https://doi.org/10.1145/1869983.1869993

Wong, J. (2013). Leveraging the General Transit Feed Specification for Efficient Transit Analysis. *Transportation Research Record: Journal of the Transportation Research Board*, *2338*, 11–19.

https://doi.org/10.3141/2338-02

Wong, J. C. (2013). Use of the General Transit Feed Specification (Gtfs) in Transit Performance Measurement Use of the General Transit Feed Specification (Gtfs), (December).

Zahabi, S. A. H., Ajzachi, A., & Patterson, Z. (2017). Transit Trip Itinerary Inference with GTFS and Smartphone Data. *Transportation Research Record: Journal of the Transportation Research Board*, *2652*, 59–69.

Zhao, F., Pereira, F. C., Ball, R., Kim, Y., Han, Y., Zegras, C., & Ben-Akiva, M. (2015). Exploratory Analysis of a Smartphone-Based Travel Survey in Singapore. *Transportation Research Record: Journal of the Transportation Research Board*, *2*, 45–56. https://doi.org/10.3141/2494-06