Exploiting Open Data for Public Transportation Analysis

A Report Prepared By:

Dr. Saeid Allahdadian
Lap-Tak Chu
Mina Park
William Qi

August 2017

Foreword

We came together during the summer of 2017 to work on a data science project for social good. Along with the rest of the Data Science for Social Good fellows, we participated in the inaugural year of this program at the University of British Columbia, with the goal of making a positive contribution using the tools of data science. Over the course of the past 4 months, we have had the pleasure of being exposed to new perspectives, ideas, and concepts while working as an interdisciplinary team. We are grateful to the Data Science Institute and Microsoft for making this opportunity possible and hope that this program will continue to inspire students for many years to come.

Sincerely,
The Transportation Team


Dr. Saeid Allahdadian


Lap-Tak Chu


Mina Park


William Qi

Table of Contents

Chapter 1 - Introduction

1.1 Data Science for Social Good

This project was undertaken as part of the inaugural Data Science for Social Good (DSSG) program at Data Science Institute (DSI) in the University of British Columbia. The DSSG program brings together teams of undergraduate and graduate students to work on data science projects in conjunction with partners from the public and nonprofit sectors.

Though the definition of “social good” can be broadly conceptualized, the program aims to leverage existing data to help communities make decisions that improve people’s lives and well-being. In its first offering this summer, a concerted effort was made to orient the program towards urban data science, with themes of transportation, tourism, business development, and education. As a result, this project is primarily concerned with how open data can be leveraged to explore how public transportation influences urban communities.

1.2 Public Transportation and Social Good

High quality public transportation is an important component of urban life. The ability of citizens to move around within their own cities - from places of residence to workspaces, social activities, and entertainment - contributes to the social and economic fabric of urban metropolitan centres. A necessary component of modern urban neighborhoods, public transit is also integrated closely with various domains that all contribute to the quality of life of city-dwellers (Figure 1).

Figure 1. Interaction between different sectors contributing towards urban quality of life.

In particular, improving the quality of public transit has a significant positive impact on several areas that can be broadly conceptualized as “social good”:

- Public Safety: Public transit reduces traffic congestion and occurrences of automobile-related accidents.
- Environment: By reducing use of personal automobiles, public transit reduces urban greenhouse gas emissions.
- Health: Mental health is ameliorated by improving connectivity to and within residential areas, resulting in lower levels of commute and transit-related stress. Public transit also promotes physical health by encouraging physical activity.
- Economy: Economic development is boosted by concentrating population density and providing economic opportunities near public transit corridors.
- Equity: Public transit provides greater accessibility to government, health, education, employment and other services. This is especially pertinent for vulnerable populations, such as low-income individuals or seniors, who have fewer transportation options and may be unable to drive for financial or physical reasons.

Improving public transit can also have important synergist and spillover effects on other dimensions of urban life. For instance, improving transit accessibility may enhance access to public health services for residents living in poorly-serviced areas. In turn, residents’ health may improve, resulting in a healthier and more productive city.

1.3 Public Transit in Surrey

Surrey is the second largest city in the province of British Columbia and the fastest growing municipality in the Metro Vancouver region, with a population of approximately 515,000 people. With rapid projected increases in population and economic growth, expansion of public transit is currently underway, with two new light rail transit (LRT) lines planned. However, for the time being, Surrey is currently served by the following public transit modalities:

- Skytrain
- Buses
- Bike share

Surrey also faces a number of unique challenges when it comes to public transit. Characterized as a municipality composed of “cities within a city”, there are 6 “town centres” that comprise the major population and commerce centres (Fleetwood, Guilford, Newton, Cloverdale, South Surrey, and Whalley/City Centre). Surrey also represents a mixture of urban/rural, residential/commercial/agricultural zones with varying population densities, demographic factors, and levels of economic development. A single bus route could start in a dense urban centre and terminate in an agricultural reserve with hundreds of meters between adjacent homes.

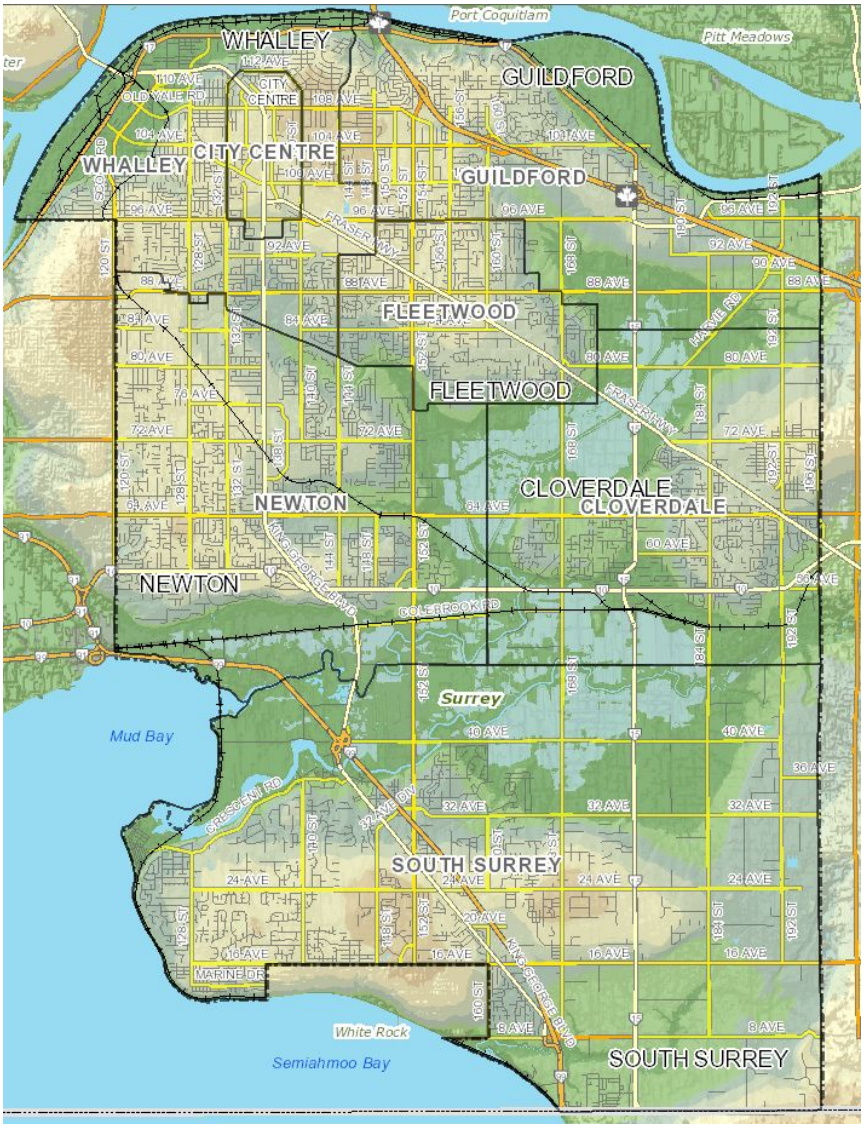


Figure 2. Map of the City of Surrey, with “town centres” indicated.

1.4 Data Science and Public Transit

Increasing efforts are being made to incorporate “smarter” ways of planning public transit service and delivery. With the advent of more data as well as novel ways of utilizing this data, cities now have the opportunity to exploit data science tools to help understand and inform transit service and planning.

1.5 Project Vision and Objectives

With many transit agencies sitting on a wealth of commuter data from fare collection and vehicle tracking systems, opportunities exist to leverage data science methods to enable more efficient transportation networks. By targeting the delivery of transportation resources to areas of greatest need, municipalities can achieve greater inter and intra-regional connectivity with minimal cost, taking advantage of the social benefits that come. Indeed, many transit agencies have recognized the potential of their proprietary data and have released open datasets to enable further analysis.

Our goal for this project was to explore how publicly available open data can be used to characterize and model public transportation networks, with a particular focus on the City of Surrey. Within this goal, we first developed a visualization tool to help transportation planners better understand how people interact with public transit in Surrey (Chapter 2). Using the same sources of data, we then explored how graph models can be used to represent the transportation system and perform a mathematical analysis of the network (Chapter 3).

Next, we attempted to characterize frequent transit networks in Surrey and the Metro Vancouver region by constructing a Bayesian model (Chapter 4), allowing us to better understand how and why certain routes are better serviced.

Finally, we explored a new method of obtaining data on commuter travel patterns, by collecting public geo-tagged data from social media (Chapter 5). This method proves to be an effective solution for providing low cost estimates of travel times and volume across the region.

Chapter 2 - Exploring Transit Availability with Interactive Data Visualization

2.1 Introduction

With more than 50 routes serving over 100,000 passengers across hundreds of bus stops every day, Surrey's bus network is a complex system with a challenging set of problems and constraints. A geographic visualization tool is one way to help enable transit planners and the general public alike to better understand the state of transit availability in Surrey. As part of our analysis of public transportation in Surrey, we set out to explore publicly available open transit data in order to develop a visualization tool that would help us better understand how and where people are interacting with the network. However, in addition to general transit indicators, we wanted to show how underlying demographic trends and socioeconomic factors relate to transportation systems in Surrey. Incorporating both types of information can help address questions targeted to transit but also to issues related to social good; examples are the following:

1. How and where are people in Surrey interacting with the bus network?
2. How does transit service availability vary across the Surrey region?
3. How does transit service availability change throughout the day?
4. How does transit service availability relate to underlying demographic trends and socioeconomic indicators?

2.2 Objectives and Methods

Our objective was to develop a multi-layered geographic information system (GIS) to present spatial and geographic attributes of the bus network. In order to reach the maximum number of potential users, we chose to package the tool as an openly hosted self-contained web application. From any internet-connected device, one can explore the extent of Surrey's public transit system and compare transportation metrics across the region.

2.2.1 – Data Sources

Translink publishes a comprehensive summary of public transit in Metro Vancouver in a standardized format conforming to the General Transit Feed Specification (GTFS). Originally envisioned as a method of supplying standardized data to Google's "Transit Trip Planner", GTFS has since been used to power everything from fare calculators to real-time bus tracking applications.

The final visualization tool incorporates GTFS data from the day of September 16, 2016 to plot the path of bus routes on the map, along with the locations of bus stops located along the way. GTFS data also provides scheduling information to allow measurement of service frequency and service quality. In addition to publicly available data, we incorporated data collected from automated passenger counters (APCs) obtained via the City of Surrey's Engineering Department. Located on entrance and exit doors of transit vehicles, APC data provides a measure for stop utilization and is accurate within a 5% margin of error.

In order to explore how demographic trends and socioeconomic factors impact public transportation, we also incorporated open data obtained from Census Canada and the City of Surrey. With information on zoning and population, we can explore how these factors interact with transit availability.

2.2.2 – Technical Details

In anticipation of new functionality and expanded coverage, the visualization tool was built using Javascript with Webpack. Each geographic layer is built as a separate module, allowing additional layers to be added with no modifications to existing code. In addition, the data module is also built with plug and play compatibility, allowing processed GTFS data from any region to be used without code changes.
At the conclusion of the DSSG program, we plan to release the code powering the transit visualization tool into the open domain on GitHub. We hope that other researchers or transit planners will find it useful and modify the project to serve their own needs.

2.3 Results

The visualization tool is accessible online. A list of customizable features for each layer of the visualization tool is presented in Table 2.

Table 1. List of customizable features for visualization tool.

Regional Boundary	Choropleth Features	Bus Stop Features
Census Tract	Number of Office Buildings	Service Count
Town Centre	Population	Boardings
	Population Density	Alightings
	Number of Jobs	Boardings and Alightings
	Employment Density	
	Jobs per Resident	
	Number of Bus Stops	
	Number of Bus Routes	
	Bus Routes Density	
	Number of Residential Buildings	
	Residential Occupancy	
	Number of Retail Buildings	
	Number of Office Buildings	
	Number of Industrial Buildings	

Within the visualization tool, users can select layers and options of interest on the left hand side of the page, with their selection visualized on the right. Hovering over a region on the map shows a popup with socioeconomic and transportation metrics computed for that particular census tract or town centre. Users can also click on individual bus stops or bus routes to view fine-grained statistics on service and utilization.

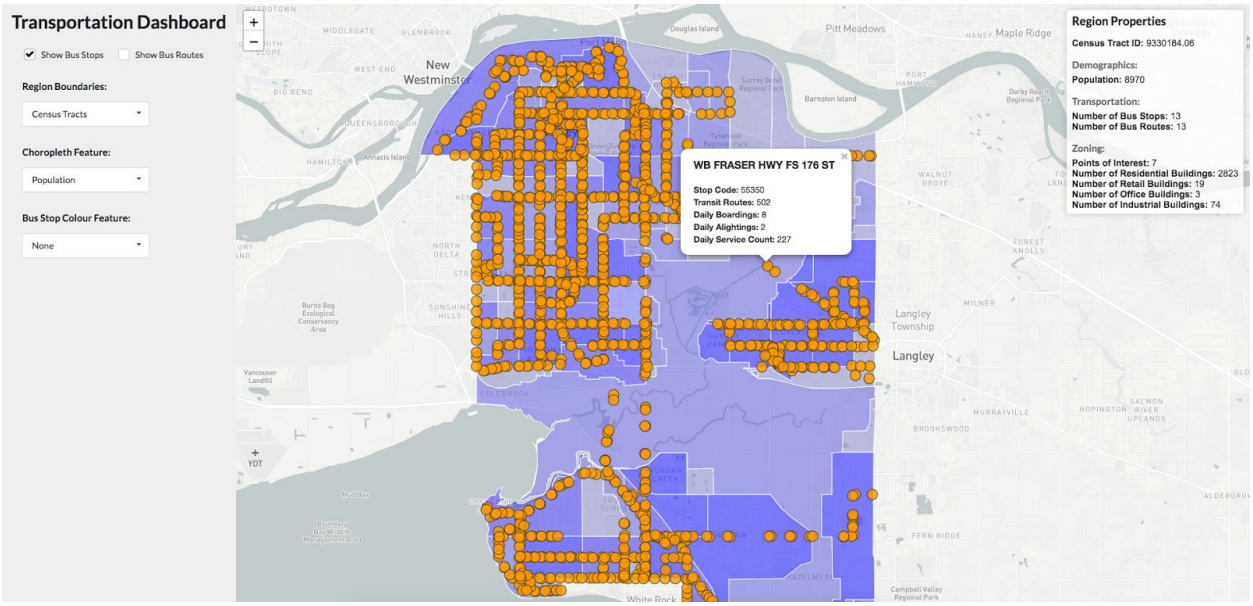


Figure 3. Example of visualization tool in use, with bus stops in Surrey plotted onto a choropleth map of population by census tract.

Users can also explore how various regional properties impact public transit by manipulating a choropleth map, which colours regions with varying intensity based on the value of a particular parameter of interest. For example, in the above figure (Figure 3), regions with lower population are light blue, while regions with higher population appear to be much darker. By choosing a particular socioeconomic measurement for the choropleth feature and overlaying bus stops coloured by a transportation metric, it is possible to explore how the two indicators interact. One interesting observation drawn from the visualization tool is shown in Figure 4, which demonstrates a correlation between regions of high industrial concentration and poor transit availability.

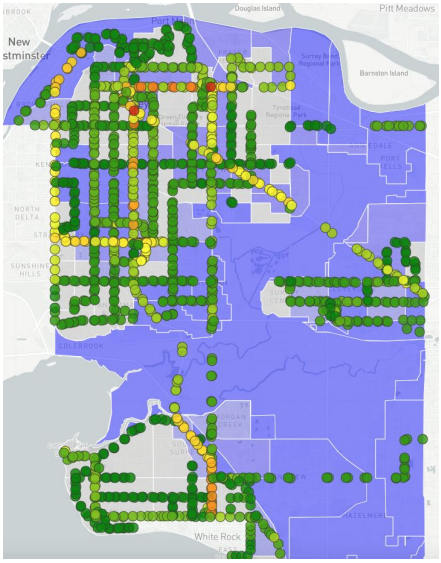


Figure 4. Bus stops coloured by service level (green is less, red is more) overlaid on top of a choropleth map visualizing the concentration of industrial buildings by census tract (darker is more).

Chapter 3 - Graph Network Analysis of Bus Transit Systems

3.1 Introduction

Understanding how public transit aligns with the manner people live, work, and play in Surrey is an important component of transportation planning. One way to measure the availability and connectivity of public transit is through network analysis, by representing the bus network using a graph model. Under this representation, bus stops are represented by nodes and routes connecting stops are represented with edges. By modelling a transportation network as a graph, mathematical characteristics of a network can be computed and compared against one another or even amongst subsections of a single graph.

3.2 Aims

The aim of the graph model is to gain a better understanding of the characteristics of the bus network in Surrey. To achieve this, the graph model is built to allow analysis on any arbitrary sub-component of the network; this allows greater insight to how well a certain area of the overall network compares against others. A partial list of metrics computed include complexity, degree of connectivity, degree of centrality, number of vertices and edges. An overall table of summary metrics can be found later on in the report.

3.3 Methods

Inspired by research conducted by Quintero-Cano et al. at UBC on operational characteristics of transportation networks[1], a modified approach from the paper was used. Because bus stops are only connected in a single direction (additional bus stops are placed on the other side of the

[1] "Bus networks as graphs: new connectivity indicators with operational ...." 21 Jul. 2014, http://www.nrcresearchpress.com/doi/abs/10.1139/cjce-2014-0054. Accessed 1 Aug. 2017.

street for the opposite direction), a directed graph was chosen to represent the network. In addition, only certain stops were kept in the initial analysis; this was done to improve computational time as well as to focus the analysis on network efficiency. In order to construct a graph model, the Python module NetworkX was chosen due to its extensive support and performance optimizations for common graph operations.

Only terminus or transfer stops were included. A terminus stop is defined as a bus stop at the beginning or the end of a route. A transfer stop is defined as a bus stop that serves two or more routes. This is further defined to be a conglomerate of bus stops that are within a 400 meter radius. As the analysis only concerns Surrey's bus network, only bus stops within Surrey were kept. However this analysis can be easily extended to include the rest of the Metro Vancouver Regional District bus network. It is also important to note that City Centre is defined as a separate area within the town centre of Whalley, in accordance with how it is represented in Surrey's GIS software, COSMOS.

Due to scheduling differences between weekdays, Saturdays and Sundays, this analysis is performed using only data for weekday schedules. However, this method is valid for weekends and holidays as well, and the analysis can be easily extended to include them. The process can be split between three distinct phases; preprocessing, graph building, and graph analysis.

3.3.1 Preprocessing

The first step of the analysis is to filter out all bus stops in the GTFS data that are not located within Surrey city limits. The remaining bus stops are then geocoded with their corresponding town centre and census tract. Next, the bus stops are labelled as a terminus, transfer or neither, following the definition given above. At this stage, bus routes are also labelled with their capacities.

In the second step, bus stops within walking distance are collapsed into a single node using the following algorithm. A dictionary is first set up to keep track of all bus stops that have already processed. The algorithm then iterates over every bus stop in Surrey, searching their vicinity for stops within 150 metres and adding them to a queue sorted by distance for secondary processing. Each stop in the queue will then go through a check to verify that it is within 400 metres of the initial stop, has not previously been marked as seen, and does not contain a

duplicate bus route. The last condition ensures that two stops in the same direction of a given route are not collapsed into a single node. If the stop passes all three checks, it then adds all stops within 150 metres of itself to the queue and is marked as seen. After iterating over all bus stops, the algorithm outputs a list of lists of bus stops, each representing the bus stops that belong to a particular node.

3.3.2 Graph Building

Using Python and NetworkX, a graph model is constructed from the output of the previous stage. We choose to use a directed representation for the in order to enable future support for incorporation of capacity and demand information. GTFS data also contributes additional properties to the graph representation; each node retains information on which census tract or town centre it belongs to and each edge retains information on which routes it holds and their capacity. The graph is then pruned of all stops that are neither transfer nor terminus, using the information gathered in the preprocessing stage.

Next, using the output generated in the preprocessing stage, stops within walking distance (400m radius) are joined together into a single node in the network. Finally, the network can be split into multiple subgraphs, with the corresponding nodes a part of each subgraph. Nodes that are not physically located in the subgraph, but have an edge leading into the subgraph are kept for analysis as well. For the purpose of this analysis, the 2011 census tract and Surrey town centre boundaries were used to create the subgraphs.

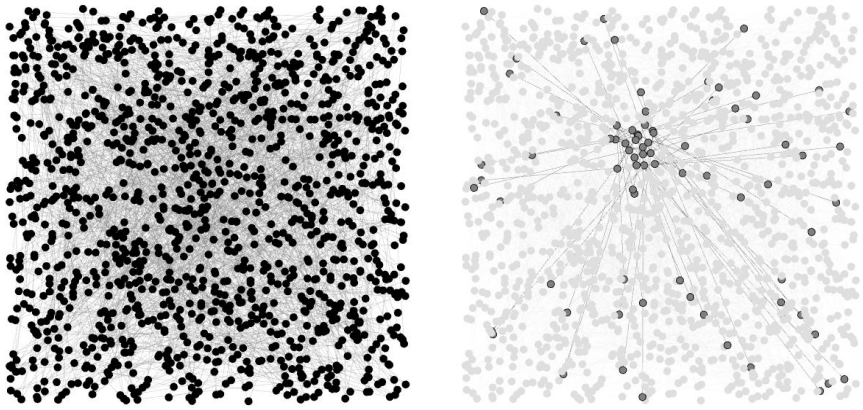


Figure 5. An illustration of the graph network for Surrey's bus network (left). A select number of nodes and their connections to other nodes (right).

3.3.3 Graph Analysis

Computation of measures of connectivity and complexity were implemented in the same manner as the paper that inspired this approach[2]. The degree of connectivity in this analysis is defined as the ratio between the number of edges present divided by the maximum number of edges possible in the graph.

$$\gamma = \frac{E}{E_{max}} = \frac{E}{3(V-2)}$$

In the above equation, E represents the number of edges present, E_{max} represents the maximum edges possible in the graph, and V represents the number of nodes (vertices) in the graph.

The mathematical complexity of the network is defined as the average number of edges per node:

$$\beta = \frac{E}{V}$$

In our analysis, we also use a modified version of the metrics above suggested by Quintero et al. in order to incorporate frequency information. In the modified version of the above metrics, E is replaced with a frequency normalized number of edges present, E^f:

$$E^f = \frac{0.5(\sum_{i=0}^{a}\sum_{j=0}^{b}\sum_{k=0}^{c} f_{ijk})}{f_{max}}$$

In the above equation, f_{ijk} is the frequency of the kth route connecting the nodes i and j and f_{max} is the maximum frequency between two nodes in the entire network (in this case, Surrey). This new definition for number of edges is substituted into the formulas for degree of connectivity and complexity to compute a new metric that includes frequency information.

[2] "Bus networks as graphs: new connectivity indicators with operational ...." 21 Jul. 2014, http://www.nrcresearchpress.com/doi/abs/10.1139/cjce-2014-0054. Accessed 1 Aug. 2017.

A summary of additional metrics computed for each type of graph is given in the table below. Note that computation of metrics for the overall graph and for average clustering within town centres was performed by first converting the graph into an undirected graph, as these metrics do not exist for a directed graph.

Table 2. List of metrics computed over the select types of graphs

Overall Graph (Surrey)	**Census Tract Sub-Graphs**	**Town Centre Sub-Graphs**
- Degree of Centrality - Dominating Set Approximation - Rich-club Coefficient - Average Clustering Coefficient	- Degree of Connectivity - Complexity - Number of Nodes - Number of Edges - Number of Stops - Number of Routes	- Degree of Connectivity - Complexity - Number of Nodes - Number of Edges - Number of Stops - Number of Routes - Average Clustering Coefficient

Degree of Centrality: Degree of centrality for a node is the fraction of nodes it is connected to.
Dominating Set Approximation: An approximate subset of the graph such that any node that is not in the set, is adjacent to at least one member of the set. In this analysis, it is used to identify potential nodes that are significant in importance to the network.
Rich-club Coefficient: A metric designed to measure robustness in the network. It measures the extent nodes with a certain number of connections (degree) also connect to other nodes with the same degree.
Average Clustering Coefficient: A metric that measures the tendency for nodes in a graph to cluster together. In this analysis it is used to measure the ease of moving within a network or subnetwork.

3.4 Results

3.4.1 Overall Graph Summary Statistics

Table 3. Summary statistics for the entire Surrey bus network

Metric	**Value**

Degree of Centrality	Max = 0.038 Node: King George Station, Guildford Exchange Average = 0.012
Rich-Club Coefficient	*Degree of Node: Rich-Club Coefficient* 0: 1.0 1: 1.0 2: 1.5 3: 2.3 4: 2.6 5: 2.0 6: 6.0 7: 3.0
Average Clustering Coefficient	0.13

By using the degree of centrality, we are able to pick out major hubs in the network. Below is a list of the top 5 nodes with the highest centrality:

1) King George Station: 0.038
2) Guildford Exchange: 0.038
3) White Rock Centre: 0.035
4) Surrey Central Station: 0.035
5) Scottsdale Exchange: 0.032

The Rich-Club coefficient also seems to suggest that as the degree of a node increases, it is more likely to be connected to other nodes with high degrees. This shows that the network is quite robust, with highly trafficked bus exchanges having many connections to other bus exchanges. However, in order to fully make sense of this metric, comparison with other networks in other regions must be done. It would be interesting to see how Surrey's bus network compares to the rest of Metro Vancouver. The average clustering coefficient is another metric that would benefit from comparison with other networks.

Looking at the set of bus stops in the dominating set approximation, it is also interesting to see that it matches very closely with the plot of highly serviced bus stops from the visualization tool and the map of FTN routes in Surrey (Figures 6 and 7).

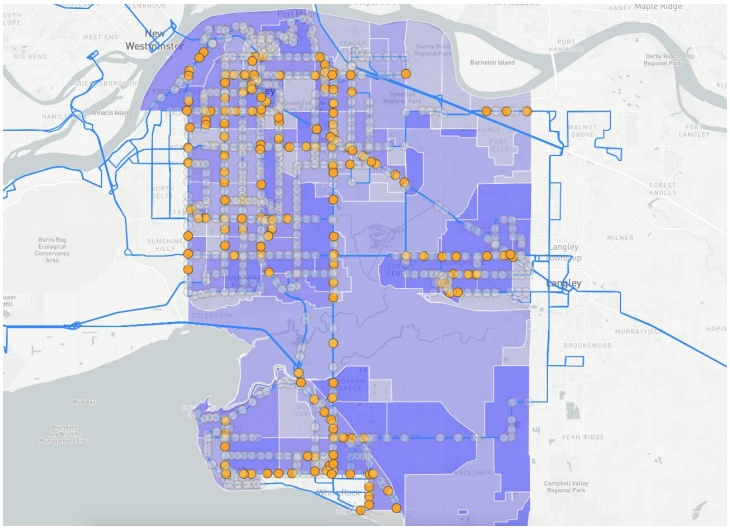


Figure 6. Set of bus stops in the dominating set shown using the visualization tool (orange stops are in the dominating set and grey stops are in the complementary set).

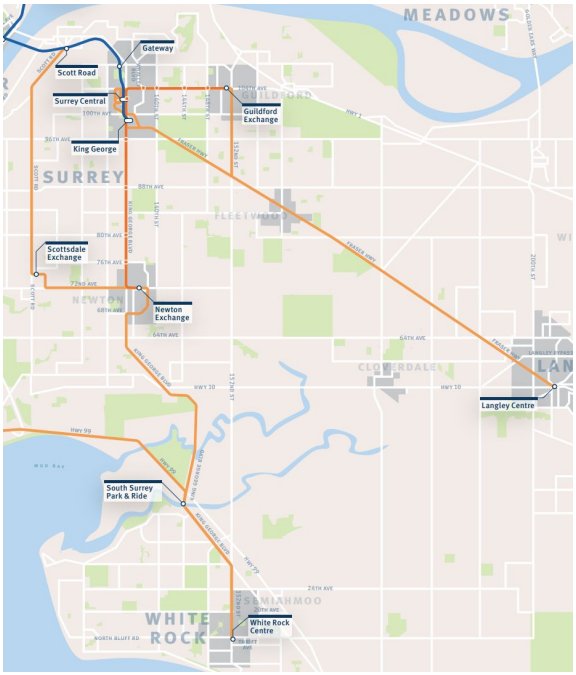


Figure 7. Map of FTN routes within the City of Surrey (indicated with orange lines).

3.4.2 Census Tract Sub-Graph Summary Statistics

Table 4. Summary statistics for the census tract level analysis

Metric	Averaged	Max	Min	Standard Deviation

Degree of Connectivity	0.06	0.35 (CTUID: 9330189.06)	0.02 (CTUID: 9330183.01)	0.04
Complexity	0.13	0.36 (CTUID: 9330190.03)	0.02 (CTUID: 9330181.12)	0.07
Number of Nodes	9.01	26 (CTUID: 9330191.04)	0 (CTUID: 9330188.07 9330188.04 9330188.02 9330182.02 9330191.06 9330185.18 9330185.08 9330187.15 9330188.01 9330180.04 9330188.06 9330187.07 9330190.04 9330188.08 9330187.06)	6.61
Number of Edges	10.66	36 (CTUID: 9330191.04)	0 (CTUID: 9330188.07 9330188.04 9330188.02 9330182.02 9330191.06 9330185.18 9330185.08 9330187.15 9330188.01 9330180.04 9330188.06 9330187.07 9330190.04 9330188.08 9330187.06)	9.36
Number of Bus Stops	14	49 (CTUID: 9330192.00)	0 (CTUID: 9330182.04)	9
Number of Routes	6	21 (CTUID: 9330191.04)	0 (CTUID: 9330182.04)	3

3.4.3 Town Centre Sub-Graph Summary Statistics

Table 5. Summary statistics for the town centre level analysis

Metric	Averaged	Max	Min	Standard

				Deviation
Degree of Connectivity	0.05	0.07 (City Centre)	0.02 (Cloverdale)	0.02
Complexity	0.13	0.20 (City Centre)	0.07 (Cloverdale)	0.05
Number of Nodes	59.9	89 (South Surrey, Whalley)	37 (Guildford, Newton)	24.8
Number of Edges	97.6	163 (Whalley)	57 (Guildford)	44.4
Number of Bus Stops	226	450 (Whalley)	107 (City Centre)	138
Number of Routes	18	31 (Whalley)	8 (Cloverdale)	9
Average Clustering Coefficient	0.10	0.14 (City Centre) 0.13 (Whalley)	0.09 (Newton)	0.02

The results in the above table suggest that Cloverdale is the town centre that is least well serviced by public transit, with the lowest degree of connectivity, complexity and number of routes. These metrics also suggest that it may be harder to get around within Cloverdale relative to the other town centres. Looking at the average clustering coefficient, it appears that most town centres have similar intra-regional ease of movement, with the exception of City Center and Whalley, which are positive outliers.

By cross-referencing the above results with the visualization tool, we can see that Cloverdale appears to be more isolated than the other town centres, as it is surrounded by the agricultural land reserve (ALR). Looking at the socioeconomic factors, we can see that Cloverdale also ranks near the bottom in population density, number of jobs, and number of office and retail buildings.

However, it is also clear from the visualization tool that Cloverdale is much closer to the City of Langley's urban centre than Surrey Central. Since our analysis is focused strictly on the transportation network within Surrey, the metrics calculated do not take into consideration any bus stops or routes outside of the city. It is possible that the connectivity is low for Cloverdale because most routes connect to Langley City Centre rather than Surrey Central. .

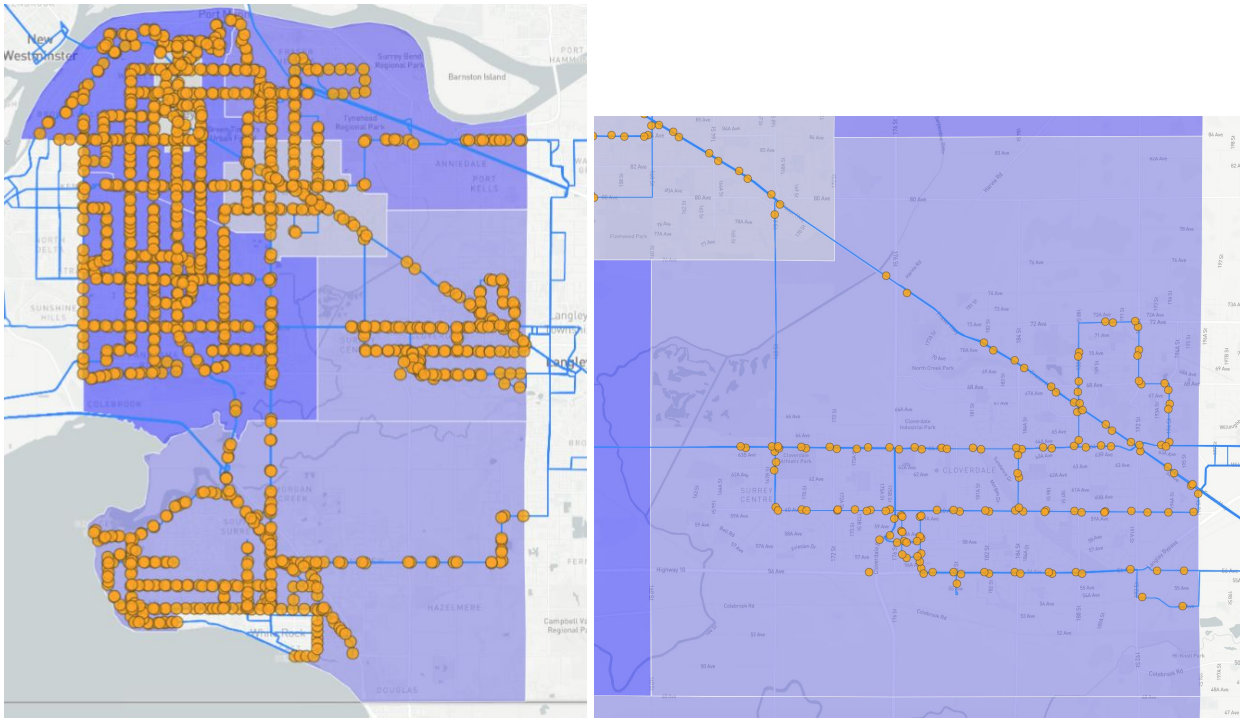


Figure 8. Map generated from the visualization tool with town centre boundaries (left), a zoomed in view of bus stops in Cloverdale (right).

3.5 Future Work and Limitations

Although our graph model provides some interesting insight into Surrey's public transportation network, it does have some limitations. In its current state, it does not incorporate any information about the flow of people through the network, meaning that it is not taking into account where people are entering and exiting the system. Such analysis would require stop-level boardings and alightings information to be integrated with the model, potentially sourced from data collected through the Compass Card system.

Once Compass Card data is integrated, information from tap-ins and tap-outs can be combined with route-level capacity to model passenger flow through the network. This could allow analysis of situations such as local spikes in transit demand and road closures on certain routes, increasing the real-world usefulness of graph models exponentially.

Another improvement that could be made to the current graph model is to represent the capacity of routes in a more realistic fashion. As we are currently computing capacity in a naive manner by scaling the capacity of the most common bus type on the route by the number of trips per day, future work could examine individual trips at different times of day to obtain a more accurate estimate. By incorporating more accurate information about passenger flow and capacity, complex graph analyses of the transit network will be much more effective.

One final thing to keep in mind is that while numerical characterizations of transit networks were computed in the course of this analysis, they are, by themselves, not very useful. These metrics are best framed by comparing against other networks in different regions and/or comparing sub-networks within a larger overall network, as demonstrated in this chapter. Future work could expand on the analysis performed in this project by extending the scope to cover networks in other cities within Metro Vancouver and by constructing an overall graph representation of the Translink system.

3.6 Conclusions

Graph representations of transportation networks are a useful tool in public transit research and planning. They allow for mathematical characterizations of transportation systems that can be used to compare and benchmark networks and subnetworks against one another. With additional data on passenger movement and more complex representations of frequency and capacity, graph networks become exponentially more useful, allowing for modelling of complex situations such as spikes in demand and road closures.

Chapter 4 - Characterization of the Frequent Transit Network in Surrey

4.1 Introduction

One way to improve public transit and network connectivity is to target the frequent transit network (FTN). As a crucial component of public transit, the FTN is a network of transit corridors where service commences by 6 AM (weekdays) or by 7/8 AM (weekends) and lasts until at least 9 PM, with frequency occurring every 15 minutes or less. Due to the regular frequency and predictability of routes contributing to the FTN, the FTN enhances connectivity within and across regions while facilitating use and reliance on public transit. Targeting FTN routes for improvement or identifying new/existing routes to integrate into the FTN may be able to enhance areas where network connectivity is shown to be lower (Chapter 3). Further, understanding the demographic characteristics of regions underlying the FTN and performance-related metrics of these routes can help city planners determine and quantify the effects of the FTN on transit performance as well as on other sectors.

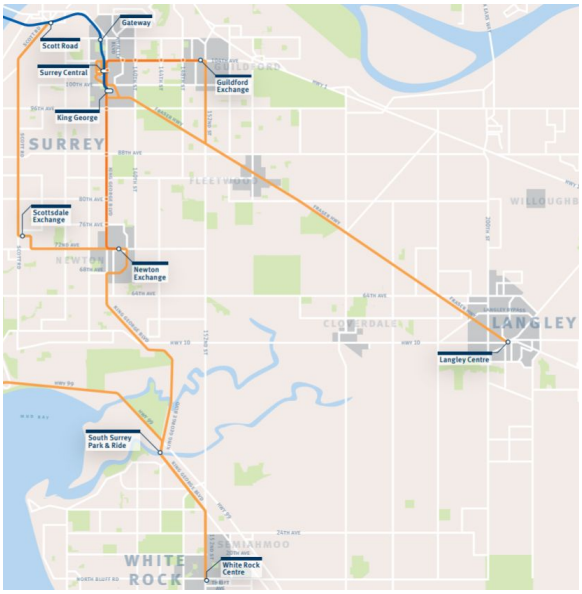


Figure 9. Map of frequent transit network in Surrey (outlined in orange).

4.2 Aims

We sought to address the following aims:

1. Identify characteristics differentiating FTN routes from other routes. This aim allows us to quantify differences of FTN routes and other routes in terms of transit performance and service utilization.
2. Explore differences in the characteristics of FTN routes in Surrey versus those in other regions. This aim allows us to determine how FTN routes in Surrey fare in comparison to the rest of Metro Vancouver.
3. Predict FTN routes in Surrey based on the characteristics of FTNs in the rest of Metro Vancouver. This aim allows us to identify which FTN routes in Surrey differ or match those of the rest of Metro Vancouver, for further route-level analysis.

4.3 Methods

Data source: Data on bus routes within Surrey and the Metro Vancouver region were manually extracted using Tabula from Translink's Transit Service Performance Review for routes numbered 001 – 799. The review provides information on demographic and service-related factors such as annual service cost and peak passenger load for all bus routes in Metro Vancouver from 2011 to 2015 (when available).

Data transformation: After the initial extraction, data was first cleaned using Pandas in Python, with all subsequent analyses conducted in R. Given high rates of missing historical data for bus routes with no apparent pattern, we impute data using averages and medians for each year. As both metrics were highly consistent with each other ($R > 0.97$ for all variables based on Pearson correlations), we proceeded with analyses using means.

Statistical analyses: Significant differences between FTNs and other routes were identified using t-tests. We predicted FTNs in Surrey using a Naïve Bayesian classifier; training and test datasets were all routes excluding those in Surrey and routes in Surrey, respectively.

4.4 Results

4.4.1 – Differences Between the FTN and Other Routes

A total of 145 bus routes were represented in our data. Of these, 46 (31.7%) contributed to the FTN. Of the features that we examined, the majority were significantly different between FTNs and non-FTN routes. From a demographic perspective, FTN routes provided service to areas of greater population and employment density (64750 vs 41868.69, p=0.00001; 68695.65 vs 35419.19, p=0.0004; Table 1). In addition, they had significantly higher revenue hours and passenger boardings across the week (Table 1). The overall service cost for FTN routes was greater than that of non-FTNs (4683765 vs 1223019, p=1.99 E-12, Table 1); yet, the cost per boarded passenger was lower for FTNs than other routes (1.42 vs 2.49, p=0.00008, Table 1). In terms of performance, FTNs and non-FTN routes were comparable in terms of being on time, though FTNs had significantly higher rates of bus bunching (Table 1).

Table 6. Operational characteristics of bus routes sourced from Translink's Transit Service Performance Review, separated by FTN status and sorted in order of statistical significance.

	Route Status		
Metric (mean from 2011 – 2015)	**Non-FTN (n=99)**	**FTN (n=46)**	**P-value**
Bus Bunching	0.89	4.86	8.90E-13*
Annual Service Cost ($)	1223019	4683765	1.99E-12*
Annual Revenue Hours	12430.29	46838.62	2.39E-12*
Average Sunday Daily Boardings	1287.82	7357.36	4.00E-10*
Average Saturday Daily Boardings	1593.41	9220.66	8.41E-10*
Average Boardings per Revenue Hour	50.52	79.82	1.97E-09*
Annual Boardings	672976.4	3935699	3.29E-09*

Average Weekday Daily Boardings	2217.45	12221.67	6.33E-09*
Average Passenger Turnover	60.53	105.64	9.62E-08*
Peak Passenger Load	21.49	29.97	2.56E-06*
Average Speed (km/hr)	28.22	21.54	5.80E-06*
Peak Load Vehicle Occupancy	42.46	54.29	6.92E-06*
Population (400m buffer)	41868.69	64750	1.27E-05*
Cost per Boarded Passenger ($)	2.49	1.42	7.56E-05*
Employment/Jobs (400m buffer)	35419.19	68695.65	0.000423*
Average Capacity Utilization	26.52	32.09	0.005371*
Revenue Hours with Overcrowding	1.6	5.08	0.018645*
Average Route Length (km)	14.63	14.12	0.708547
On Time Performance (%)	49.6	49.7	0.941331
**indicates statistically significant at a raw p<0.05.*			

4.4.2 – Characteristics of FTNs in Surrey Versus the Rest of Metro Vancouver

Of the 46 FTN routes in Metro Vancouver, 8 (17.4%) were routes that serviced Surrey. Comparing the metrics of FTN routes in Surrey versus the rest of Metro Vancouver revealed a number of significant differences. FTN routes in the rest of Metro Vancouver had significantly higher employment density than FTN routes in Surrey (88039.47 vs 29062.5, p=0.00001; Table 2); however, this was not the case for population density (Table 2). Overall, FTN routes outside of Surrey had higher amounts of passenger boardings across the week, though FTN routes in Surrey performed better in terms of being on time and average speed (40.73 vs 51.59, p=0.002; 30.07 vs 19.74, p=0.01; Table 2). Finally, an additional metric of consideration that is near

statistical significance (p=0.055) and of note is the higher cost per boarded passenger for FTN routes in Surrey versus elsewhere (2.16 vs 1.27, Table 2).

Table 7. Operational characteristics of FTN routes located in Surrey compared with FTN routes located in other municipalities within Metro Vancouver, sorted in order of statistical significance.

	FTN Route Location		
Metric (mean from 2011 – 2015)	**Non-Surrey (n=38)**	**Surrey (n=8)**	**P-value**
Employment/Jobs (400m buffer)	77039.47	29062.5	1.00E-05*
On Time Performance (%)	51.59	40.73	0.00212*
Average Passenger Turnover	111.89	76	0.00488*
Average Saturday Daily Boardings	9915.83	5645.48	0.00638*
Average Sunday Daily Boardings	7888.33	4626.67	0.0071*
Annual Boardings	4285242	2275371	0.00893*
Average Weekday Daily Boardings	13311.84	7043.33	0.00912*
Average Speed (km/hr)	19.74	30.07	0.00972*
Average Boardings per Revenue Hour	84.58	57.2	0.01658*
Cost per Boarded Passenger ($)	1.27	2.16	0.0554
Peak Passenger Load	30.85	25.77	0.06513
Bus Bunching	5.21	3.21	0.0682
Average Route Length (km)	12.78	20.5	0.07998
Annual Revenue Hours	48892.63	37082.08	0.11498
Annual Service Cost ($)	4889163	3708125	0.11513
Population (400m buffer)	66671.05	55625	0.18647

Peak Load Vehicle Occupancy	55.26	49.69	0.26249
Average Capacity Utilization	32.41	30.57	0.55722
Revenue Hours with Overcrowding	5.32	3.93	0.59474
**indicates statistically significant at a raw p<0.05.*			

4.4.3 – Prediction of FTN Routes in Surrey

Finally, in order to better understand which attributes are significant in determining whether a route belongs to the FTN, we sought to build a model to predict the FTN status of routes in Surrey using a statistical model trained on data from the rest of Metro Vancouver. Results from a Naïve Bayesian Classifier correctly classified 5 out of 8 (62.5%) FTN routes in Surrey (Table 3); table 4 shows the corresponding route numbers. From this analysis, we explored the 5 correctly classified FTN routes versus the 3 that were not predicted to be part of the FTN (Figure 1).

Table 8. Number of routes in Surrey predicted to be FTN routes vs actual FTN routes from a Naïve Bayesian Classifier model.

		Predicted	
		No	Yes
Actual	No	27	0
	Yes	3	5

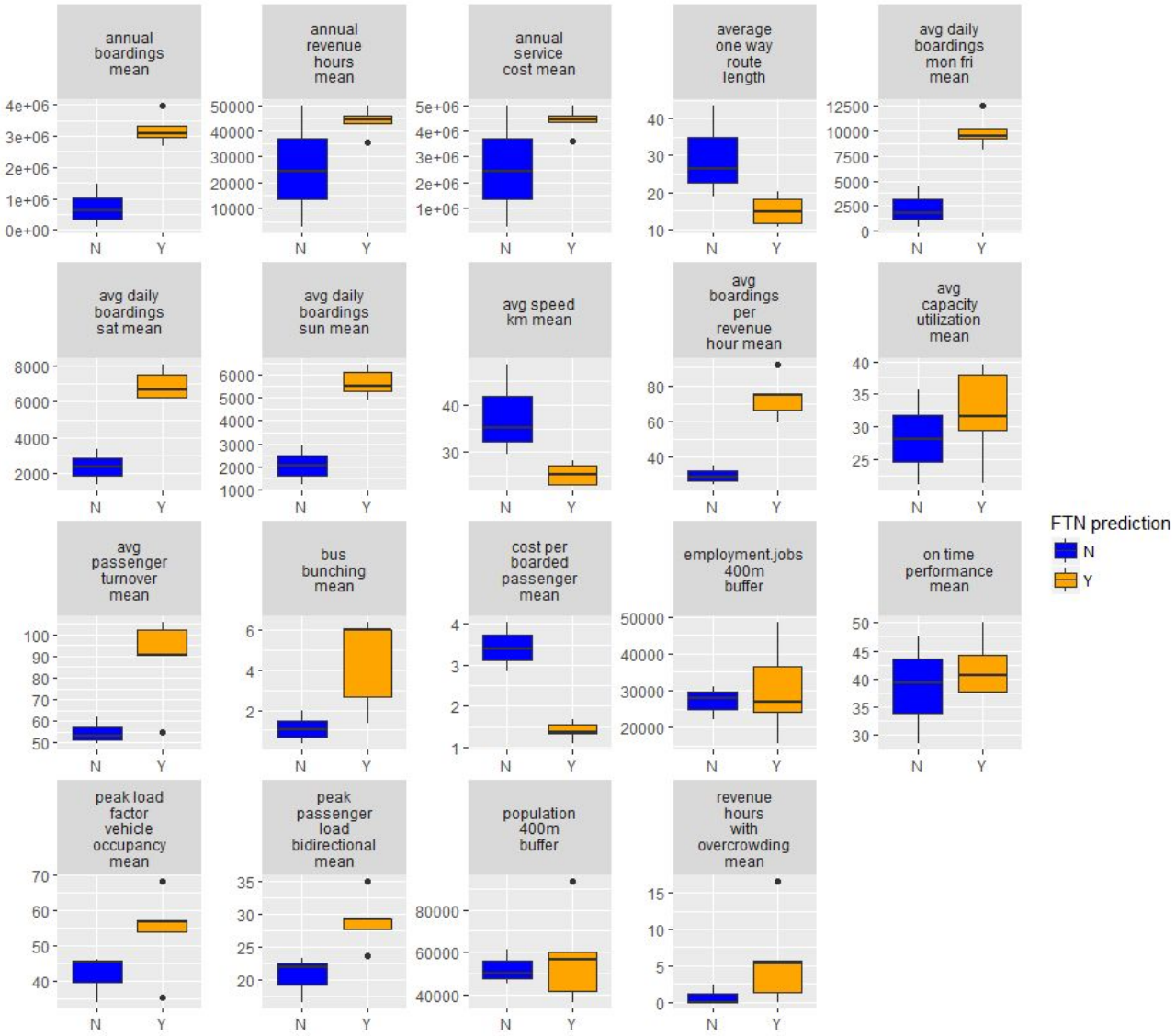


Figure 10. Boxplots depicting differences in route characteristics (Table 6) based on whether they were correctly or incorrectly classified as being part of a FTN using a Naive Bayesian Classifier.

Table 9. Summary of predictions from Naïve Bayesian Classifier of FTN status on FTN routes in Surrey.

Surrey FTN Route #	Prediction
96	FTN
319	FTN
320	FTN
321	FTN
502	FTN
351	Non-FTN
503	Non-FTN
394	Non-FTN

A close examination of differences in correctly and incorrectly classified routes (Table 9) showed a number of differences related to utilization, with correctly identified routes having higher rates of boarding and passenger turnover, as well as lower cost and speed relative to the incorrectly identified routes (note, we did not perform statistical tests given the very small sample size). Furthermore, employment and population densities were identified as similar for both groups. Examining the physical location of the routes themselves revealed that most of the the routes that were correctly predicted stayed within or near the boundaries of Surrey, whereas the incorrectly classified routes extend deep into neighboring municipalities (351 - White Rock, 394 - White Rock, 503 - Langley/Aldergrove).

Another interesting observation about these routes that is not reflected in our model is that these routes are all operated by a different model of bus from most other routes. These particular routes are served by Orion highway coaches, which offer greater passenger comfort at the expense of seating capacity on routes with long distances between stops. As a result, the cost per boarded passenger may be inflated on these routes due to limited passenger turnover and low capacity, which could influence the prediction from our model.

4.5 Future Work and Limitations

One of the primary limitations in our analysis was a lack of data, as we were unable to secure access to open transportation data for all regions. Future work could expand on the model described in this chapter by incorporating such data, as well as information from other modes of transit (Skytrain, Seabus, bike paths, etc.). It is likely that additional data could provide a significant boost to accuracy and show additional insights.

The use of non-public data, for instance, from Translink directly, that provides more detailed service utilization metrics, would be a valuable addition. Another option to explore is the use of other sources of public data from social media (Chapter 5). The small amount of data is also a limitation in our statistical approach, as power is limited.

4.6 Conclusions

The results of our analyses are consistent with what would be expected. As high quality public transit promotes higher utilization of transit, the differences in boardings we observe with FTN routes compared to other routes serve as confirmation of this expectation. The differences observed in FTN routes in Surrey compared to the rest of Metro Vancouver are also unsurprising, as Surrey represents a growing and emergent city. An interesting finding was that for FTN routes in Surrey relative to those in the rest of Metro Vancouver, there was lower employment density along these routes, though population density was not different, suggesting that FTN in Surrey are required to service the high population in the city. Perhaps one implication from these results is that there is room for greater economic development along FTN corridors in Surrey.

Examining the routes in Surrey that were incorrectly classified as not being part of the FTN showed some key differences. One key observation was that the cost per passenger was far higher in the routes that were incorrectly classified; however, we also saw that these routes were important in connecting Surrey to other municipalities which may justify their inclusion in the FTN. In addition, a key similarity between these routes was the density of population and employment in the areas surrounding the routes, further suggesting their importance in

providing service to key population and economic corridors despite having overall lower utilization.

In the future, it is possible that with enough and appropriate data, the use of predictive modelling approaches can be used as an additional decision making tool when planning for the FTN and other public transit routes. Further, modeling the potential effects of introducing FTN routes on transit network connectivity, as seen in the previously reported graph analysis, may help city planners understand how different regions may be better served.

Chapter 5 – Understanding Public Transportation Patterns Using Social Media

5.1 Introduction

With the rapid spread and increasing affordability of sensor-laden smart devices, users are generating more public content than ever before. Information shared online by commuters allows for direct and instant feedback about their transportation experiences, helping to pinpoint problems and bottlenecks in real time. With a wealth of accurately geo-tagged data available to mine, social networks have become a valuable resource for retroactive analysis, enabling greater understanding of how and where commuters interact with public transit systems.

Of the major social networks used in North America, Twitter users tend to be particularly active, with over 6000 tweets sent every second worldwide. Of the million tweets sent every week in Vancouver, approximately 1% are geotagged, meaning that over 10000 data points are generated on a weekly basis. These tweets could provide valuable insight into how, where, and when commuters interact with the public transit system, allowing for more accurate measurement of utilization.

5.2 Objectives

Analysis of public Twitter posts can help us better understand the transportation system in three ways: 1) how the population is distributed throughout a region, 2) how people move around and commute within a region, and 3) how people feel about the state of their transportation systems.

The first objective was to use public Twitter data to help transit planners better understand how and where people live and work, in order to inform better demand models for transportation systems. The patterns recognized from Tweets could also help to create more efficient routes for public transit by minimizing the overall distance travelled by each individual to their nearest

access point. Additionally, service frequency be adjusted by real world usage in order to better reflect how people are interacting with the network.

Our second objective was to measure demand on popular public transit routes by analyzing historical locations of Twitter users. By measuring across a large sample of public geotagged Tweets in Metro Vancouver, it is possible to obtain an estimate for relative demand on each route throughout the day. Transportation planners can then take advantage of this information by optimizing schedules to match commuter and traveller demand. Comparison of demand estimates with existing timetables also helps to identify potential areas of concern.

Finally, by analyzing Tweets directed at the transit authority's official Twitter account (@Translink), locations and routes in the public transit network that need additional attention can be identified.

5.3 Analysis of Twitter Data

Using the Twitter API, more than 3 million public tweets were gathered in Metro Vancouver over a period of 21 days in July and August, 2017. Of these tweets, approximately 1% were geo-tagged, providing more than 30,000 data points for analysis. These geo-tagged tweets provide insight into the movements of 3440 Twitter users over a three week time period.

During this time, 4800 tweets were also directed towards @Translink. While this amount of data provides sufficient accuracy for an exploratory analysis, it should be noted that the accuracy can be significantly improved by collecting data over a longer period of time.

A plot of all geo-tagged tweets is shown with blue markers on a map of Metro Vancouver in Figure 11. From the distribution of points in this figure, several route shapes are identified as part of the first objective can be inferred.

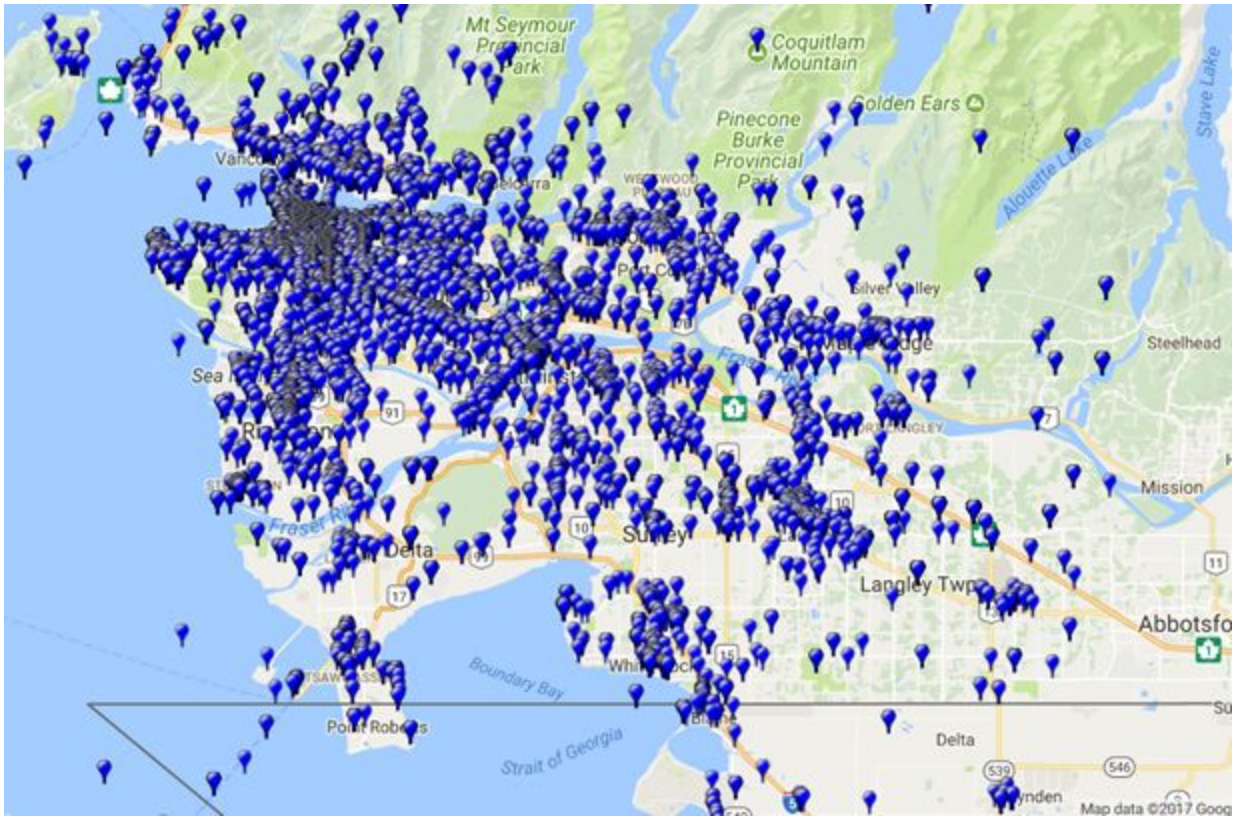
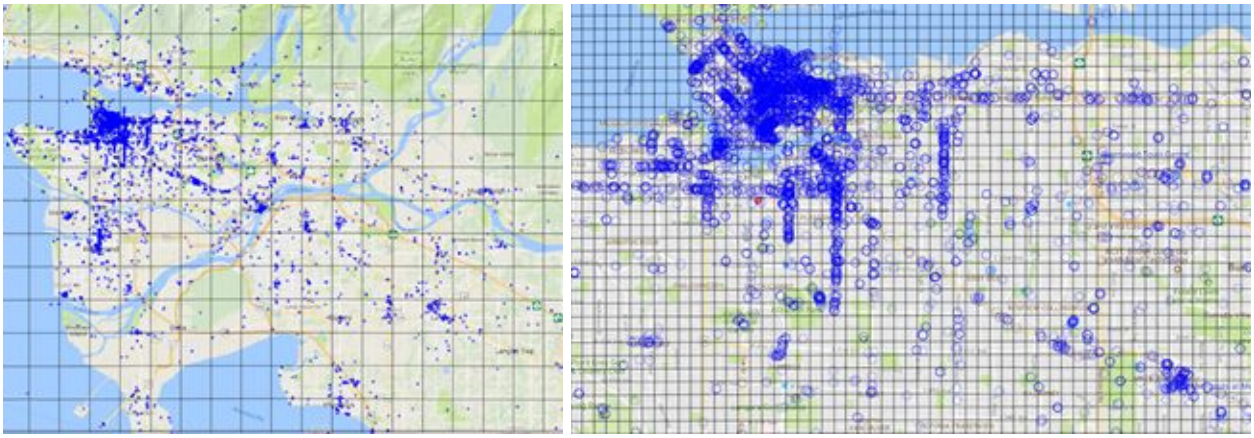


Figure 11. Distribution of geotagged Tweets across Greater Vancouver; each blue marker represents a single Tweet.

5.3.1 Routing From Distribution of Tweets to Increase Service Availability

In order to create optimal routes based on the geographic distribution of Tweets, the map must first be discretized into a mesh. Under this discretization, route segments can be approximated using a vector in each cell of the mesh, which in turn connects with segments in neighboring cells to form a larger route. The distribution of discretized Tweet locations is visualized in the following figure for two different mesh sizes, determined based on the number of Tweets in Surrey and Downtown Vancouver.



a) **b)**

Figure 12. Discretized mesh adjusted for (a) Surrey and (b) Downtown Vancouver overlaid on top of a map of the region (blue circles show the location of tweets).

Using the mesh shown in Figure 12, the distribution pattern can be approximated with a line in each cell. The direction of these lines are calculated by a singular value decomposition (SVD) analysis, which reduces the effects of random noise from the distribution. Moreover, the number of Tweets in each cell can be visualized by the adjusting the thickness of each line. After repeating this analysis for the Metro Vancouver region, the joined routes are shown in Figure 13.

a) **b)**

Figure 13. Routes from SVD analysis with mesh size adjusted for (a) Surrey and (b) Downtown Vancouver.

5.3.2 Measuring Transportation Utilization Using Tweets

By analyzing the movements of users using the locations of their past tweets, general patterns of movement throughout the day can be identified. In order to perform this analysis, the map is again first discretized with a mesh. By counting the number of users who travel between each cell of the mesh, volume of traffic between any two points on the map can be approximated. The average interaction between cells in the Downtown Vancouver area over three weeks is shown in Figure 14. The thickness of lines between cells is proportional to the strength of their interaction (transportation demand).

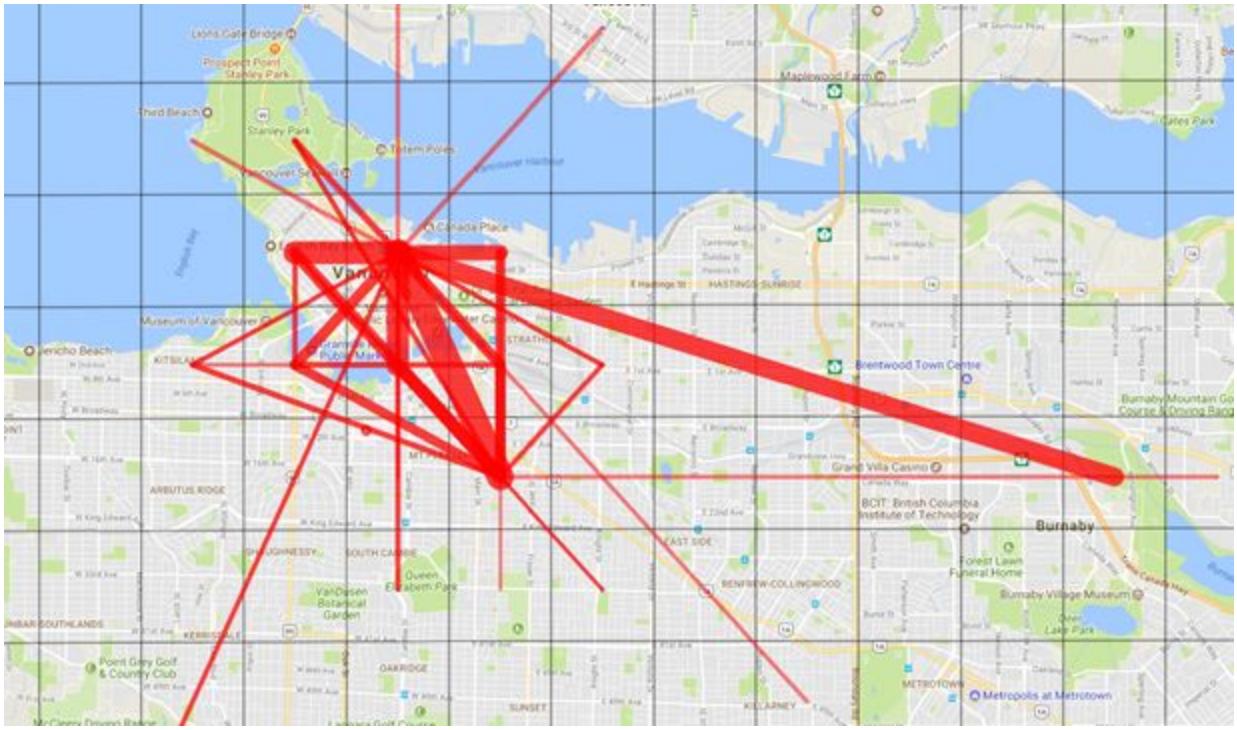


Figure 14. Average interaction between regions in the Downtown area over three weeks.

It should be noted that the lines in Figure 14 are connecting the centroids of each cell to one another. The above figure shows that a large amount of people travel between the Main Street corridor and Downtown. Moreover, there is also a large volume of traffic between Burnaby and Downtown; this is expected as many suburban residents commute to work in the Downtown core on the Skytrain. It should also be noted that less significant interactions (low demand) have been omitted from the above figure for the sake of clarity. Areas with high interaction (high

demand) in the Downtown Vancouver area also show a correlation with existing FTN routes (Chapter 4). This shows that the existing public transit network serves the needs of commuters well and that the routes are well-used by the general public.

By changing the size of the mesh and restricting the data for specific time intervals, these interactions can be computed for different times of day (or year) in each region and with various levels of spatial resolution.

In order to produce a visualization of interaction between areas with sparsely distributed tweets, such as Surrey, major points of interest within the region must first be identified using a clustering approach. For the purposes of this study, *k*-means clustering with *k*=30 is chosen. Temporally and spatially significant trips from/to/within the Surrey area are shown in Figure 15. The origin and destination of each trip is shown with markers colored by their respective clusters. The centre of each cluster is also marked using a blue circle.

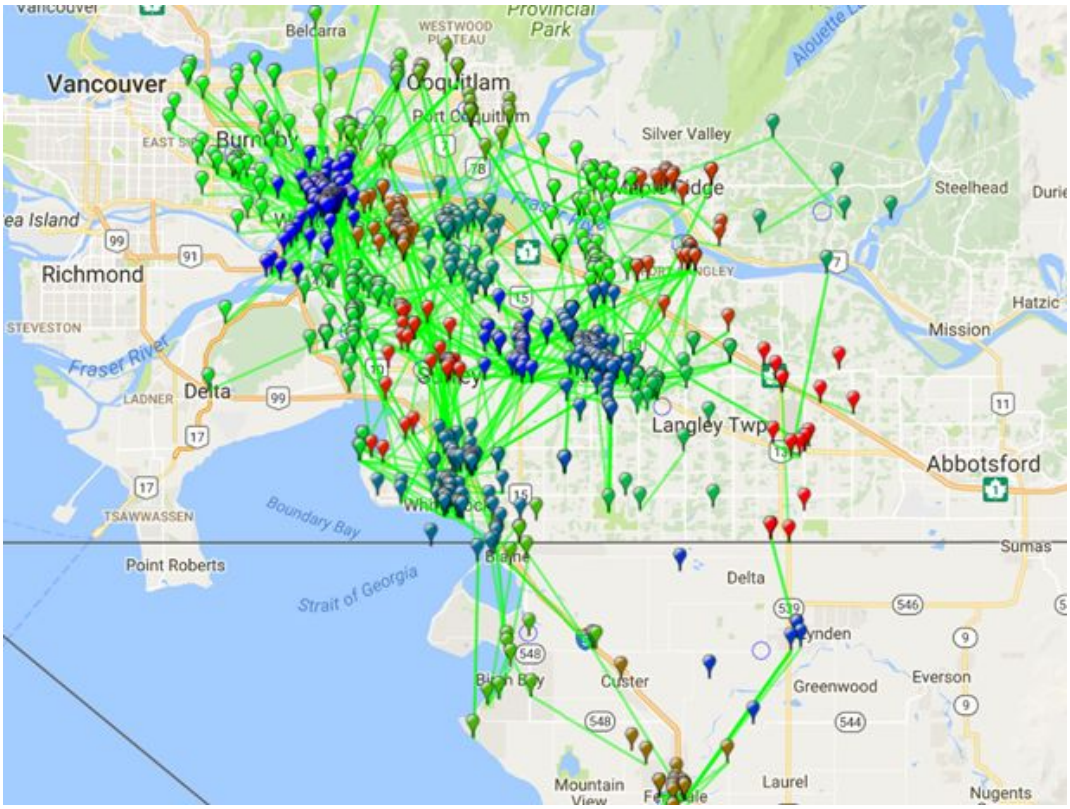


Figure 15. Visualization of travel between clusters of points of interest in the Surrey region (markers are coloured by cluster).

The interaction between clusters can then be calculated by determining the number of people travelling between each point throughout the day. Computed interactions with cluster centres in the Surrey region are visualized in Figure 16, in which the thickness of each line is proportional to the strength of interaction. One interesting observation is that residents of Langley Township appear to have greater intra-regional transportation interaction relative to their neighbours, indicating that residents of Langley may prefer to work and play closer to home.

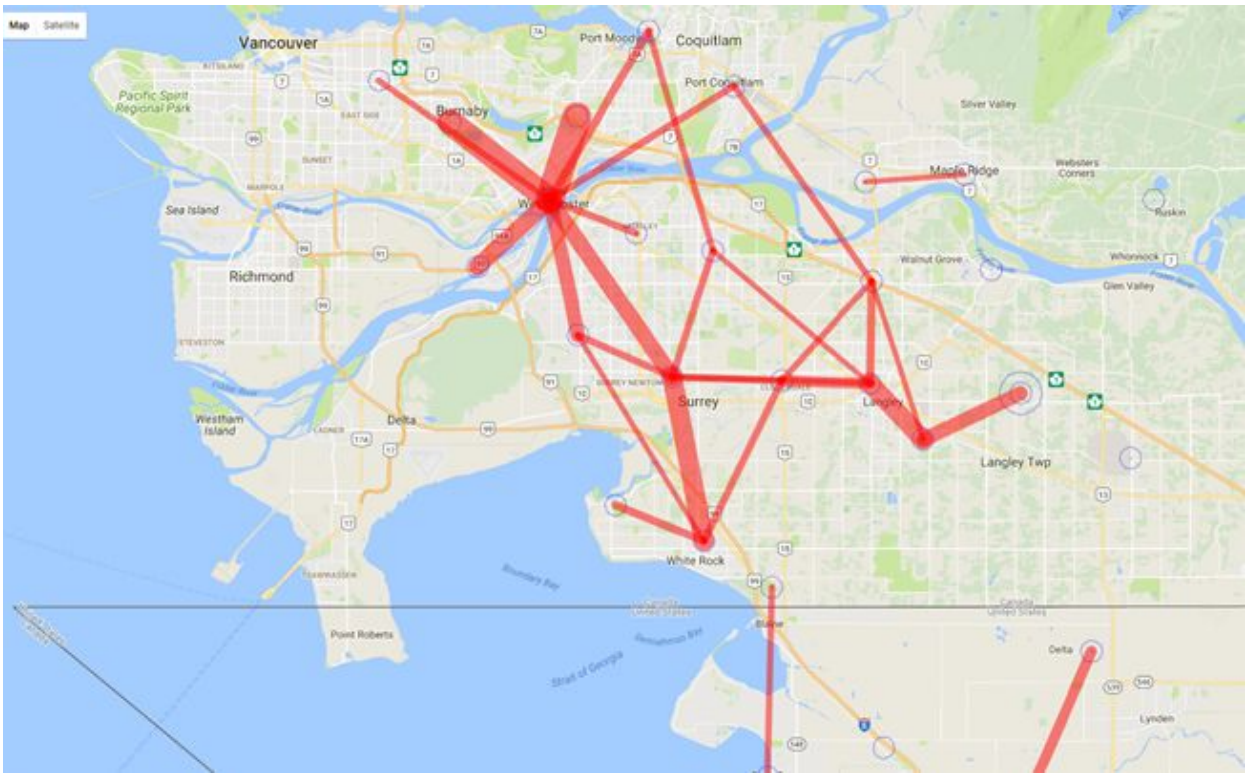


Figure 16. Visualization of transportation interactions between clusters in the Surrey region.

Next, the public transit travel time between each cluster centre shown in Figure 16 is calculated using Google Maps. The minimum travel times for each trip (starting at noon on a Friday) are shown and compared in Figure 17. Highly trafficked connections with long travel times are circled.

Figure 17. Visualization of transportation interactions between clusters in the Surrey region, marked with their respective travel times on public transit.

From the above figure, it can be seen that some connections within Surrey and many connections within Langley appear to be poorly serviced, with long public transit travel times on highly trafficked routes. With inefficient transit networks in these areas, many people are likely commuting via personal automobile, contributing to pollution and congestion in the region. By upgrading the transit network with increased frequency and more direct connections, it would go a long way towards improving quality of life in these regions.

Other candidates for transit network improvements are connections between Coquitlam/Port Coquitlam and Surrey/Langley, as there is significant demand for transportation but the current public transit network ill-equipped to serve commuters. Additional areas of interest for transit planners include the Newton area within Surrey and New Westminster, as they are critical transfer points for travellers in the region.

By comparing Figure 17 with existing FTN routes in Surrey (Figure 9), it can be seen that the poorly connected areas are located some distance away from rapid transit; reaching an access point for fast and reliable public transit requires either a long walking connection or an additional bus connection. To meet the demand for transportation in underserved regions, transit planners may want to consider extending the FTN deeper into Langley.

5.3.3 Analysis of Commuter Feedback

The final component of this analysis was to explore how social media can be used to evaluate commuter feedback and identify problematic areas of the network. For the purposes of this exploration, Tweets directed towards Translink's official Twitter handle (@Translink) were collected over the course of three weeks. As Twitter users generally only send Tweets to Translink when they encounter issues in their travels, it is safe to assume that the vast majority of messages will be negative.

By extracting stop IDs and route numbers from these Tweets using various filters, it is possible to obtain a measure of how problematic a particular element of the network is. Locations mentioned in Tweets to @Translink are visualized in Figure 18, with markers indicating the concentration of Tweets from each area (blue is lowest, red is highest). Markers coloured in grey indicate the geo-tagged location of each Tweet.

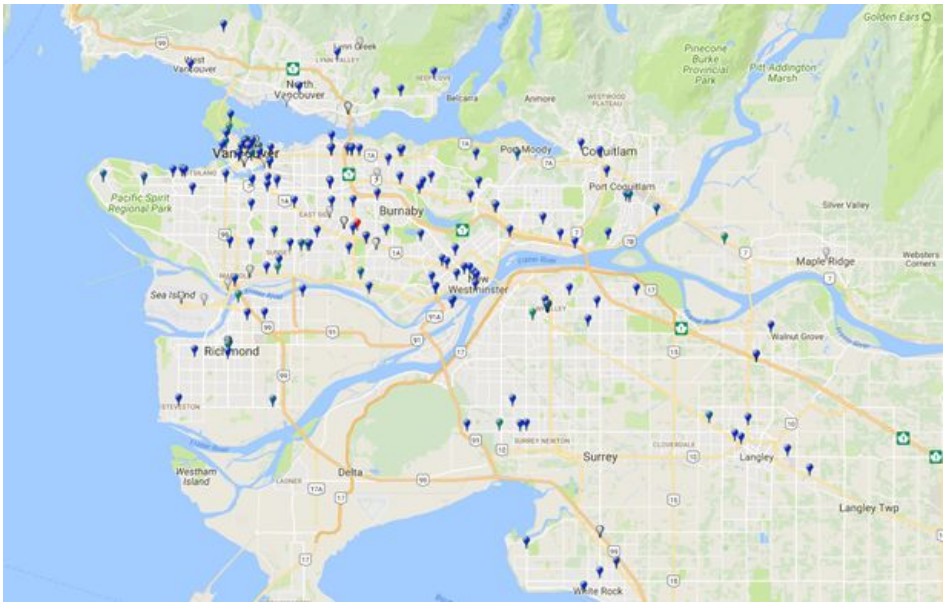


Figure 18. Visualization of locations mentioned in Tweets to @Translink.

Table 10. Bus route number ranked by number of mentions in Tweets to @Translink over 3 weeks

Route Number	Number of Occurrences
340	22.0
502	17.0
555	15.0
301	13.0
96	10.0
319	10.0
351	7.0
345	7.0
503	6.0
335	6.0
326	5.0
388	5.0
501	5.0
352	4.0
354	4.0

In the above table, a list of problematic routes with the highest number of mentions in Tweets to @Translink is summarized. It should be noted that these results can be significantly more accurate when categorized by the time of day and analyzed over a period longer than three weeks. However, these initial results show the tractability of using Tweets as a method of collecting feedback and identifying potential problems in the transportation system.

5.4 Discussion and Conclusion

In this chapter, public data was gathered from a social network over the course of three weeks and used to provide insight into the transportation patterns of commuters in Greater Vancouver. First, geographical information from Twitter users was used to identify areas of interest throughout the region, in order to determine where people work and live. Next, by analyzing the movements of Twitter users over time, it was possible to identify highly trafficked routes between previously identified regions of interest. The results were adjusted based on population density and travel time on public transit was computed for each route. From the findings, several

routes and connections in the Surrey region were then proposed as candidates for future upgrades.
As a final exploration, Tweets directed towards the official account of the regional public transit authority were analyzed to help identify problematic areas in the network. Bus numbers and stop IDs were extracted and ranked based on number of mentions, providing a starting point for transit planners to investigate transportation issues. These methods proved to be an effective and efficient way of passively collecting commuter feedback.

It should be noted that data from social media can be biased towards the demographics active on these platforms. Results collected from analysis of data from these platforms are best used as a cost-effective approximation of real-world transit patterns, rather than a comprehensive survey of how all travellers are interacting with the network. However, as mobile internet penetration increases, increasing numbers of people from all walks of life will join social media and begin producing public content, meaning that these methods will only get more accurate with time.

Finally, although this analysis was performed on Twitter data collected over the course of three weeks, utilizing a larger data set with increased temporal coverage should improve accuracy significantly. By collecting data over the course of an entire year or longer, it is possible to analyze how commuters behave at different times of the year or how commuters behave in adverse weather conditions. Moreover, feedback collected in real-time can also be compared to historical data in order to determine if certain problems in the transportation system are temporary or chronic. This can be extremely valuable when planning changes to routing and in adjusting timetables dynamically to meet the demands of the people.

Chapter 6 - Conclusions

6.1 Summary

To better understand the availability and connectivity of public transit in Surrey, and how public transit aligns with the population's needs, we sought to 1) visualize and 2) characterize various aspects of the bus network. We focused on bus routes as they represented the public transit modality on which the vast majority of intra-regional travel occurs. As we were unable to obtain access to detailed boardings/alightings data from Translink's smart card system, we also addressed an additional aim, which was to 3) explore Twitter as a potential data source for public transit analysis.

An open-source visualization tool incorporating information on transit availability as well as socioeconomic and demographic factors was developed. By showing the frequency of service at the level of every bus stop in Surrey, our tool can readily provide an indication of the amount of service that is available throughout the city. Mapping service onto regional characteristics can further reveal insights into how well the underlying population is being served in addition to highlighting any disparities.

From the graph analysis, it was found that the bus network in Surrey was well-integrated, overall. However, we also identified certain regions where there were notable differences in network complexity and connectivity. These were Cloverdale and City Centre/Whalley, which fared poorer and better, respectively, in terms of network characteristics. The higher connectivity of City Centre/Whalley was unsurprising, given its prominence as a key economic hub and as the seat of municipal control. The lower metrics for Cloverdale suggest that it is harder to move around within that community compared to others.

The characteristics of FTN routes were explored as they represent a key component of transit network connectivity and reliability. We found that routes contributing to the FTN had higher utilization, served more densely populated and more developed economic regions, and were cost-effective, as would be expected. Comparing FTN routes in Surrey to the rest of Metro

Vancouver suggested future economic growth along FTN corridors in Surrey is possible (and likely, as the city grows and matures). Finally, looking within Surrey, we found some evidence to suggest that the FTN may help facilitate long commutes and regional interconnectedness by serving population centres across municipalities, despite the higher cost of providing service along inter-municipal corridors.

Finally, social media data from Twitter was explored as a proxy data source to characterize the location, movement patterns, and feedback of people using the public transportation system in a region. The methods from this analysis proved to be a novel and effective solution for approximating transportation patterns of urban populations, providing important data that could be used to inform development of more efficient transit routes and less congested cities. Tweets to @Translink were also identified as a real-time solution to find problematic areas in the transportation network, with several bus stops and routes proposed for further analysis.

6.2 Relevance to Social Good

A number of different ways of analyzing public data to understand public transit service in Surrey were explored. We hope that our project provides additional insight into public transit network dynamics and performance in Surrey. Further, we hope that our visualization tool will allow the City of Surrey and the general public to address questions relevant to transit availability and its relationship to the underlying socioeconomic and demographic characteristics of underlying regions. We hope that these insights can enhance understanding of transit in Surrey and its relevance to sectors beyond transportation such as health, economic growth, and equity.

6.3 Future Directions

Data-driven approaches to understanding public transit in Surrey were outlined. We anticipate that much value can be generated by applying these and other methods to rich data sources held by transit authorities, such as Compass Card data. We expect that integrating these data with other public sources of data, and applying such analytic methods, will facilitate deeper

understanding of transit dynamics, and facilitate future transit planning to better serve the needs of cities and their citizens.