Navigating Bus Bunching problem in Brooklyn with GTFS realtime data

Hanbyul Jo

Intro

Public transportation has long been a critical component of transportation infrastructure in New York City. While the subway system has typically received more attention than buses, the COVID-19 pandemic has resulted in daily bus ridership exceeding that of the subway for the first time in over five decades. Buses have demonstrated their value in connecting different parts of the city, particularly in comparison to the subway, which primarily facilitates travel to Manhattan. Despite the potential of the bus system, it has been beset by declining ridership over several years, which has resulted in insufficient funding to support system improvements, thereby perpetuating a negative cycle.

The introduction of alternative modes of transportation, such as ride-sharing options, has increased the complexity of urban mobility. Determining the factors contributing to the decline in bus ridership is thus not straightforward. However, it can be argued that poor service quality, including bus reliability and speed, may be contributing to the issue. The unpredictability of bus trips and the slowness of the service may discourage passengers from choosing the bus as their preferred mode of transportation.

Bus-bunching is the phenomenon when the headways between buses decrease from the scheduled value and are magnified over time. This occurrence often leads to frustration among passengers who experience longer-than-expected wait times, ultimately undermining the reliability of public transit services. This research is dedicated to the examination of the bus bunching phenomenon, a problem that frequently compromises the reliability of scheduled transportation services. The aim of this study is to develop a method for quantifying bus bunching using real-time General Transit Feed Specification (GTFS) data and identify areas in Brooklyn, New York, that are prone to bus bunching. Furthermore, the study seeks to better understand the spatial factors that contribute to bus bunching through spatial analysis.

Literature

Previous studies have approached the bus bunching problem in three main categories, first, using the bus bunching as a way of measuring service reliability, second, finding the cause of bus bunching, and the last visualizing the operational data. (Figliozzi et al. 2012) Even though the categories are not mutually exclusive to each other, such as Visualization techniques that make use of multiple variables are able to provide additional insight into the potential causes of transit service problems (Kimpel, 2006), this study aims to build upon existing research regarding visualizaion and finding the cause of bus bunching. Key studies that I reviewed closely are discussed below.

Visualizations

Data Visualization as a Tool for Improved Decision Making within Transit Agencies

Time-distance diagrams have been widely adapted to visualize the schedule data. This visualization technique offers insight into the relationship between actual and scheduled service on a per-trip basis as well as the spacing of vehicles between successive trips—at any point in time and space (e.g., at each time point) as well as over time and space (e.g., between time points) (Kimpel, 2006). While highlighting the utility of the time-distance matrix, Kimpel suggests the incorporation of additional dimensions of trips using 3D space, depending on the agencies' requirements.Kimpel also provides other examples of transit data visualization, including the general mapping of quantity information and linear referencing.

Study of Headway Maintenance for Bus Routes: Causes and Effects of ”Bus Bunching” in Extensive and Congested Service Areas

Despite the convenience that time-distance diagrams provide, Figiliozzi et al. (2012) argue that they may not be the optimal method for comprehending the finer details of what is occurring within a system, and thus overlook the opportunity to offer insights from each trip. As an alternative, the authors propose interactive applications that allow users to select performance measures at the time point or route level.

Finding contributing factors

Factors affecting bus bunching at the stop level: A geographically weighted regression approach

With the availability of real-time data such as Automatic Vehicle Location (AVL) and Automatic Passenger Count (APC) data, new methods for addressing transportation issues have arisen. Chioni et al. (2020) utilized AVL data to determine the headways or the time intervals between consecutive arrivals of same-line buses at each stop in central Athens. Subsequently, bunching events per stop were aggregated for the period considered, and this figure was used as the model’s dependent variable. The authors ran Global Moran analysis to show the cluster patterns of the independent variables and bunching events. Then a Geographically Weighted Regression (GWR) model is estimated and compared with a conventional global regression model to demonstrate the effectiveness of the proposed approach. The model included various factors on bunching, such as bus lanes, the number of lanes, and nearby transits.

Empirical findings of Bus Bunching Distributions and Attributes Using Archived AVL/APC Bus Data

This study identified specific time periods and segments where bus bunching frequently happens in Portland, Oregon. It further identifies the relationships between bus bunching and relative attributes from consecutive bus trips through Bus Detection System(BDS) data. The data includes stop-event data where the headway can be obtained by subtracting two consecutive departure times. The front bus and the following bus are labeled for each headway. They used the arbitrary value of 3 mins, which means if a departure headway is smaller than three minutes, this headway is identified as a bus bunching event. After analyzing where bunchings happen often, they try to identify the attributes contributing to the problem. They focus on the combination of the bus pair on the same route, front-bus, and following-bus. They

tested the various combinations of possible scenarios and ran sensitivity analysis to identify the most problematic pair.

Data and Methodology

Data Collections

The General Transit Feed Specification (GTFS) is a data specification that allows public transit agencies to publish their transit data in a format that can be consumed by a wide variety of software applications. Today, the GTFS data format is used by thousands of public transport providers,[1] including Metropolitan Transit Agency (MTA) in New York. GTFS has the specifications for both static schedules and the real time. Static GTFS data consists of various components such as stops, routes, trips, and calendars. The static GTFS data that is used for this study was collected through TransitLand (https://www.transit.land/), which offers the archive of GTFS data from many transit agencies in the world. The real-time data used for this study were collected by scraping the real-time data every minute through MTA's developer's api. The real-time data consists of trip updates, alerts, and vehicle positions, with trip updates being the main component used in this study to identify bus bunchings based on bus stops. This study uses the real-time dataset collected from 7 am to 9 am on November 14th Monday. Due to the request failures, a total of 88 realtime data set was collected for the period of time. This dataset is relatively limited in size due to technical difficulties associated with parsing realtime data. However, it is expected that the available data will still provide valuable insights into bus bunching patterns in New York City. Bus lane data that was used in the further spatial analysis was acquired from NYC Open Data Portal (https://opendata.cityofnewyork.us/).

Data Preparation

The trip updates from GTFS-realtime data include the information of the trip, route, and the changed arrival time for each stop of the trips. This includes updates for all subsequent stops in the event of a delay at an earlier stop. In order to obtain a complete record of trip schedule changes, updates for a given trip are merged based on their sequence, as depicted in Figure 1. The scheduled arrival times for each stop were retrieved from the static GTFS using the trip and stop IDs. It should be noted that real-time GTFS data only includes information for trips with schedule changes, which indicates the trips made without the change of the schedule were not included in this study. The resulting bunched data were aggregated for each bus stop, with the aggregated results presented as follows. At the end of the processing, the data had a record of 335 trips from 53 routes across 2,2235 stops. Trips often have two directions, in-bound and out-bound which are represented as direction_id as 0 and 1. This study only considered the trips with direction_id 1 for the convenience of analysis.

[1] (n.d.). *GTFS: Making Public Transit Data Universally Accessible*. General Transit Feed Specification. https://gtfs.org/

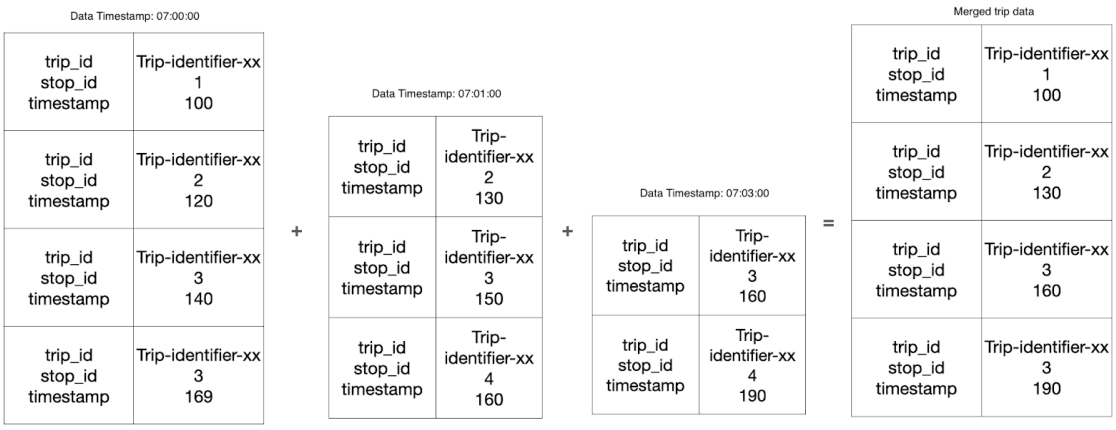

Figure 1. The process of consolidating scraped trip data

Geographically Weighted Regression(GWR) required the calculation of supplementary attributes for each bus stop. This process includes spatially joining the bus stop data with the bus lane data to determine whether the stop is located on the bus lane. In addition, the study evaluated the density of nearby bus stops by counting the number of stops within an 800-meter radius of each stop. Lastly, the number of routes associated with each stop is determined through the GTFS data and incorporated as an attribute in the GWR analysis.

Identifying Bus Bunching

Upon data preparation, the occurrence of bus bunching was calculated in this study. This study measures the bus-bunching on stops employing the following steps:

1. Measure realtime headways between the vehicles with real-time GTFS, based on the arrival time of vehicles on the stops.
2. Compare the realtime headways to the scheduled headways by matching trip id from the realtime data to static GTFS.
3. Identify instances of bus bunching by marking stops where the real-time headway is less than 70% of the scheduled headway.
4. Aggregate the bunched events at each stop.

The result is shown in Figure 2.

Figure 2. Bus bunching events aggregated on stop level

Analysis

Visualizations of bus bunching

Bus bunching is a complex phenomenon that involves both spatial and temporal components. Comprehensive visualization of the problem can assist researchers and planners in developing and evaluating strategies to alleviate the issue. The time-distance diagram of a selected route (B15) is examined below. We can observe that there are three vehicles getting closer than scheduled through the visualizations. The visualizations were produced using the Matplotlib Python Library.

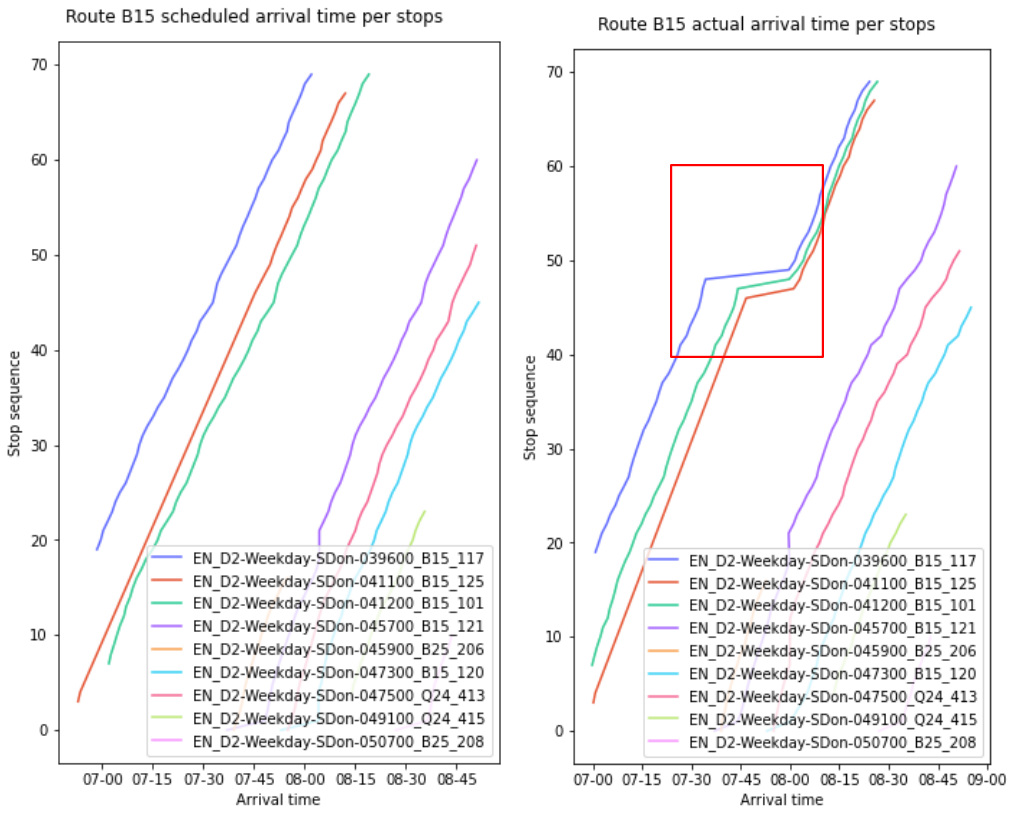

Figure 3-1 (left). Time-distance diagram of scheduled trips from route B15. Figure 3-2 (right). Time-distance diagram of realtime trips from route B15, with bunched trip segments marked with a red rectangle.

Below is the visualization of the same data but with time-geography approach. The stop coordinates were plotted on x-axis and y-axis, with their scheduled arrival time on the z-axis. Although this approach may not be scalable, the 3D plots allow for a more comprehensive visualization of the geospatial context of bus bunching. The resulting 3D plots reveal that several vehicles are getting closer together than scheduled at the end of the route, indicating the occurrence of bus bunching as we observed in time-distance diagram. However, this visualization offers more geospatial context of where this happens. For example, the plots show that bunching occurs in segments of the routes before the curve. This information can give good insights to planners about specific route configurations. The study provides several additional examples of these types of visualizations in the appendix, created using the Plotly library in Python.

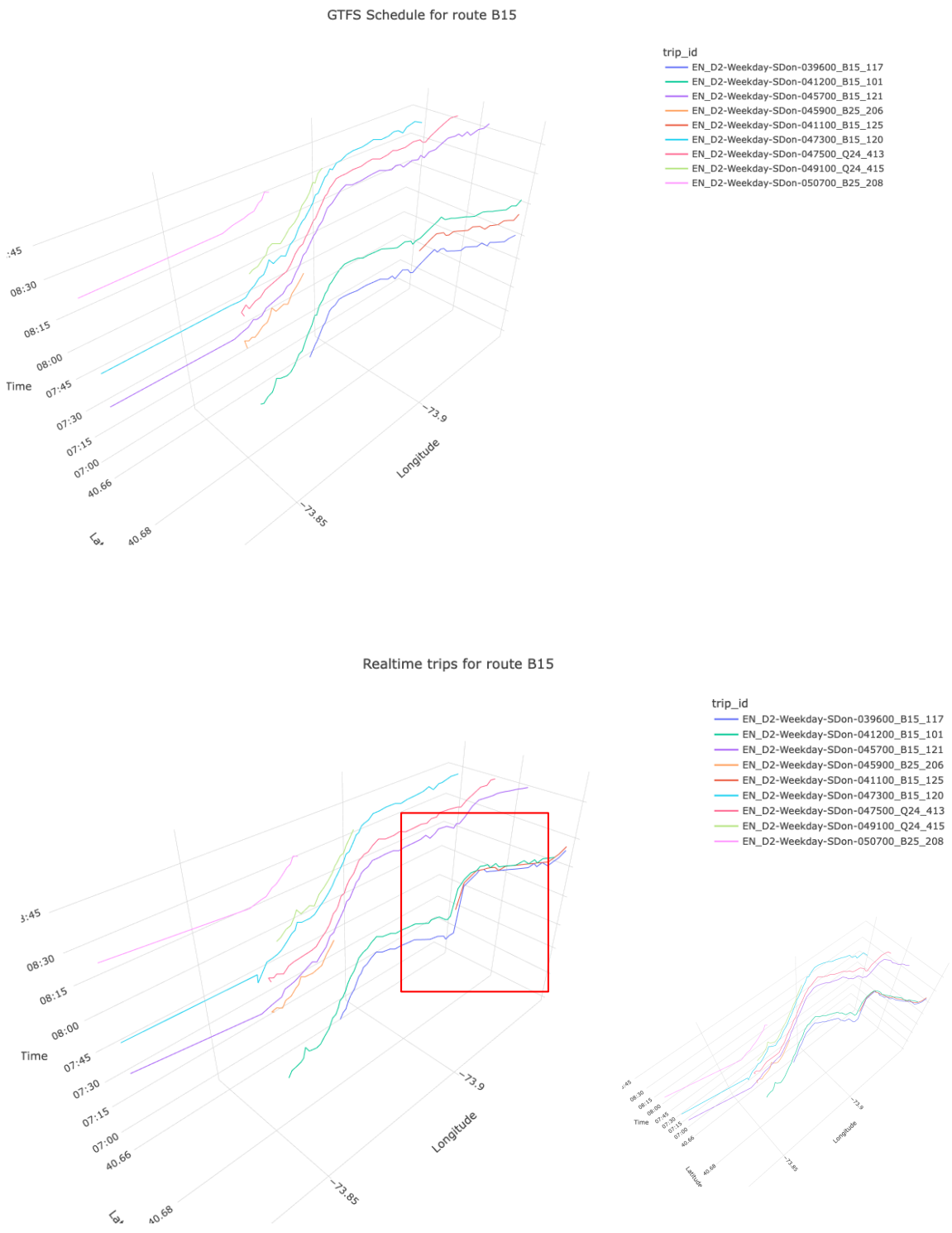

Figure 4-1 (top) depicts the time-geography approach used to visualize the scheduled trips from route B15. Figure 4-2 (bottom left) shows the time-geography approach used to visualize the real-time trips from the same route, with bunched trip segments marked with a red rectangle. Figure 4-3 presents a different angle of the visualization of real-time trips.

Geographically Weighted Regression (GWR)

Spatial Autocorrelation through Moran's I

In order to examine the spatial autocorrelations of bus bunching events, the study calculated the Global Moran's I value using the PySal ESDA package. The resulting Global Moran's I value was 0.40, indicating a moderate level of autocorrelation. However, in the Local Moran's test, statistically significant hotspots and cold spots were identified. Specifically, the hotspots were found to be concentrated in the southern part of Brooklyn and along the border with Queens, while cold spots appeared to be clustered in the northern area of Prospect Park. These findings suggest that bus bunching is not randomly distributed across the study area, and that there may be underlying spatial patterns that contribute to the occurrence of bus bunching events.

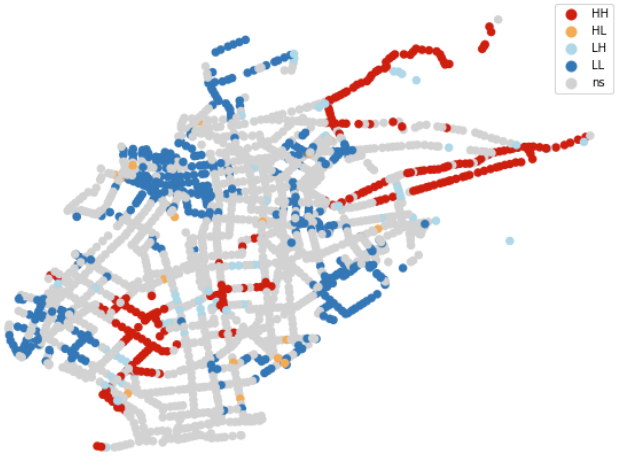

Figure 5. Bus bunching LISA cluster map

GWR

Local Moran's result shows some levels of spatial autocorrelations of bus bunching events. To identify potential spatial factors contributing to bus bunching, three independent variables were chosen based on previous research. The first variable was stop density, measured by counting the number of stops within specific thresholds. A threshold of 300 meters was selected, taking into account the average distance between stops in Brooklyn. The second variable was whether the stop was located on a bus lane or not, determined by checking whether the stop intersected with a bus lane. The third variable was the number of routes using the stop, obtained through joining the trip and stop data from the static GTFS. A GWR analysis was conducted with PySal's mgwr package.

	GWR (Bisquare kernel)	Global Regression
Akaike Information Criterion (AIC)	5394.738	6313.265
Adjusted R2	0.362	0.015

Table 1. AIC and Adjusted R2 value from GWR and Global Regression

	Mean	STD	MIN	Median	Max	p-value
Intercept	827787673517.812	491178863801 18.617	-431813723674620.812	-0.072	228191917398 6893.500	1
BusLane	244406203847 3.084	145021682917216.062	-12749398952026 31.000	0.057	673741855118 9778.000	0.005
Number of nearby stops	-0.032	0.108	-0.484	-0.016	0.189	0.000
Nubmer of Routes	0.175	0.194	-0.160	0.143	0.883	0.000

Table 2. GWR Parameter Estimates

The higher adjusted r2 score in GWR indicates that the model is able to explain more of the variation in the data at the local level. However, the fact that the stops with high R2 scores don't overlap much with the hotspots identified by the Local Moran's test suggests that the factors contributing to bus bunching may be complex and influenced by a variety of factors. The results of the GWR analysis also indicate some issues with the parameter estimates. The BusLane variable showed a very large value for both the mean and standard deviation, which doesn't make sense. Additionally, the relationship between the number of nearby stops and bus bunching seems to be counterintuitive, with more stops apparently leading to a decrease in bus bunching. This could be due to the small sample size of the data, particularly for stops on bus lanes. Also, counting the stops inside of the radius of a stop might not be the most optimal way of showing how dense the stop is, considering this method doesn't differentiate the stops on different routes. Only the number of routes using the stops showed a significant and reasonable relationship with bus bunching.

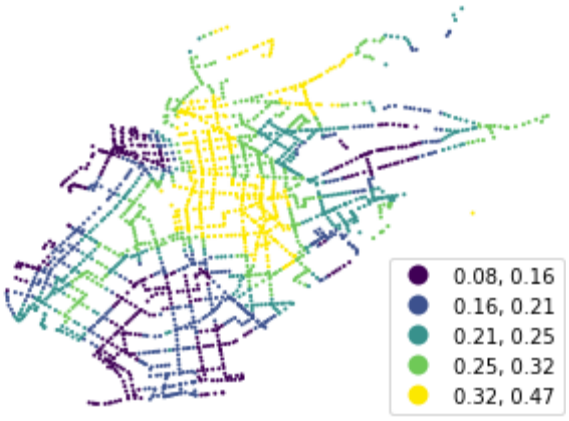

Figure 6. Local R2 Value mapped

Connecting the visualizations and the findings from GWR

One of the benefits we can get the visualizations based on time-geography approach is that it is not limited to one route. After identifying the number of routes as an attribute of bus bunching, the visualization of the trips from two routes that share the stops is examined.

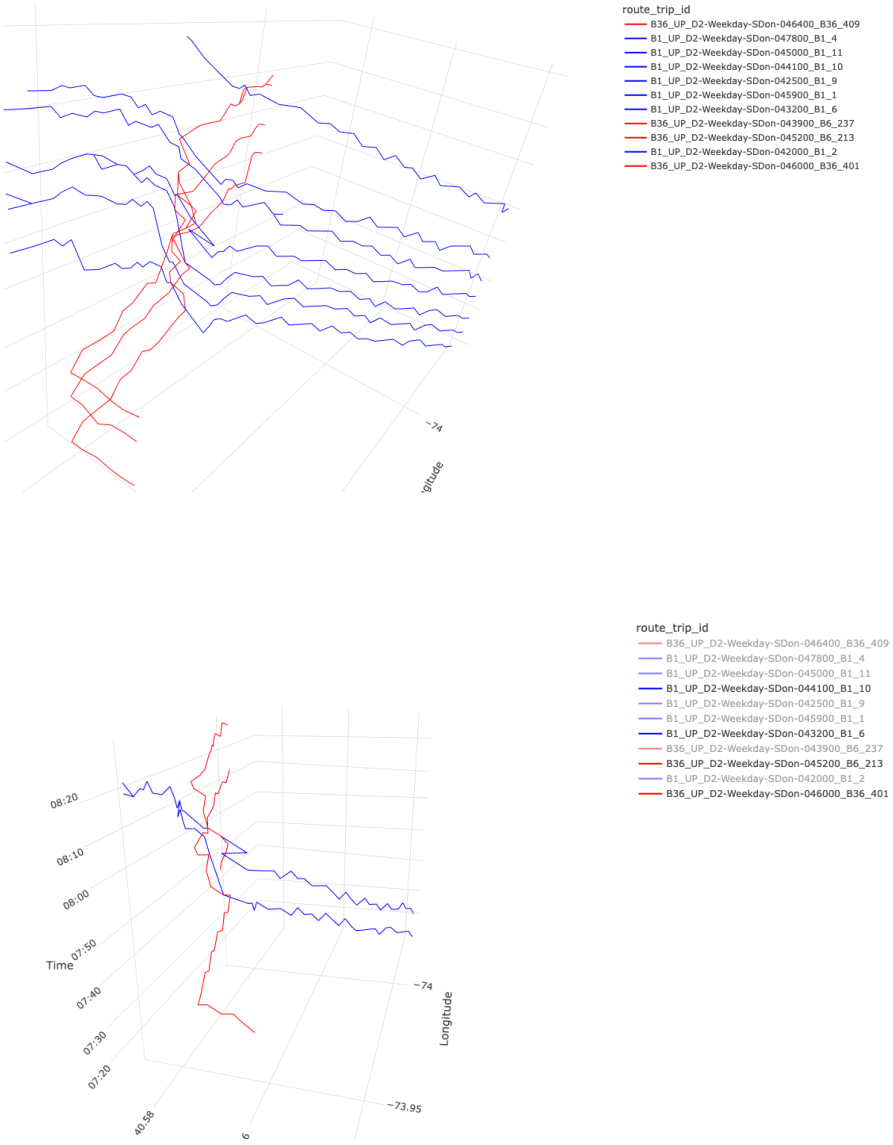

Figure 7-1 (top). The trips from two bus routes (B1, B36) that share the stops were visualized. Figure 7-2 (bottom). Some trips are filtered to highlight some trips. The interval between vehicles becomes smaller after the intersection of the other route. This suggests that the interaction between different routes at the intersection may be a contributing factor to the occurrence of bus bunching, which aligns with the finding from GWR.

Conclusions

It is important to acknowledge that the dataset used in this study was subject to certain limitations. Given the highly dynamic nature of public transit, it would have been beneficial to incorporate data from other time segments, such as evening peak time and night time, in order to obtain a more comprehensive understanding of bus bunching phenomenon.

As mentioned in GWR section, bus stops have different spatial and temporal weights inside of the routes. Unfortunately, these nuanced attributes were not comprehensively captured in this study, as evidenced by the relatively weak outcomes produced by the GWR analysis.

Despite the limitations of the dataset, this study provides a useful time-geography approach for visualizing bus bunchings that incorporates geospatial context and is applicable to multiple routes. Time-geography visualization showed its strength by demonstrating the finding of this study's GWR section that the number of routes on the stop might have an impact on bus bunchings. The approach, in combination with findings from other studies, can provide valuable insights to transport planners and riders on the quality of transit services and aid in improving route configurations and stop designs.

Bibliography

Kimpel, Thomas. (2006). Data Visualization as a Tool for Improved Decision Making within Transit Agencies. Transportation Northwest.

Figliozzi, Miguel ; Feng, Wu-chi ; Laferriere, Gerardo (2012). *A Study of Headway Maintenance for Bus Routes: Causes and Effects of "Bus Bunching" in Extensive and Congested Service Areas,* Oregon Transportation Research and Education Consortium.

Chioni, E., Iliopoulou, C., Milioti, C., & Kepaptsoglou, K. (2012). Factors affecting bus bunching at the stop level: A geographically weighted regression approach. *International Journal of Transportation Science and Technology*. https://doi.org/10.1016/j.ijtst.2020.04.001

Feng, Wei; Figliozzi, Miguel (2011). ICCTP 2011: Towards Sustainable Transportation Systems, *Empirical Findings of Bus Bunching Distributions and Attributes Using Archived AVL/APC Bus Data.* https://doi.org/10.1061/41186(421)427