1

Bus Reliability Metrics using Public MTA Bus Time Data

Jiaxu Zhou, Yuqiao Cen, Matthew Urbanek, Bonan Yuan, and Sara Arango-Franco

Abstract—With the growing demand for public transportation, bus becomes an important issue. New York UniversityâĂŹs Center for Urban Science and Progress (CUSP) collaborates with the New York City Department of Transportation (DOT) to provide user focused reliability metrics relevant to the agencyâĂŹs capabilities to improve user experience. This paper summarizes and analyses common bus reliability metrics. Based on the public Bus Time Advanced Vehicle Location (AVL) data from the public Metropolitan Transportation Authority (MTA) and the schedule data (GTFS), this paper discusses data extraction and parsing, data quality, big data techniques, and inference of time at stops to aim at measuring bus reliability. Several extraction, parsing and interpolation exercises are critically discussed. By comparing with Bus Service Measurement Standards (NYCT), this paper presents a method using the AVL system to improve bus operations. Finally, it aims at aiding the DOT with some bus operation, traffic, and some other potential elements that affect bus reliability to help the agency control and improve bus performance on the planning level.

Index Terms—reliability; MTA Bus Time data; Big Data; AVL systems; bus performance.

1 Introduction

Despite a growing demand for public transportation in New York City, bus ridership levels are declining. This can be explained by drops in vehicle speeds and customersâĂŹ perceptions of dependability. The New York City Department of Transportation wishes to engage CUSP to explore the use of public data from the MTA open vehicle location system Bus Time to generate operational information relevant to NYC DOT planning decisions. This information will be provided in the form of reliability metrics for the bus service. Based on the MTA Bus Time data, this project uses big data techniques and data analysis methods and algorithms. At the end, it will provide methods for estimating bus travel times, measuring reliability, and potentially identifying the distribution of those measurements as a functions of factors regarded as relevant.

The MTA is the designated authority for transit operations, and it has internally defined metrics that are used for schedule planning and analysis of the bus network. Despite of this, the MTA pays more attention to the bus level but not to the whole system level, which is more of the interest of the DOT. The Department of Transportation has already collected the Bus Time data, but they lack a formal process for compiling and using it for decisions that are more relevant to them, such as those related with road design and traffic management. Unlike the MTA, the DOT takes more responsibility on traffic planning instead of bus operation management and concerns more about customersâĂŹ expectation. To help the DOT improve their planning efficiency, this project aims to figure out the dependent variable metrics related to bus performance and reliability, so in a further analysis the agency can analyse the effect of independent variables that affect service quality. CUSP will collaborate with the DOT to develop methods for estimating bus travel times and measuring reliability, while performing data quality analysis of the Bus Time data and the MTA schedule data (referred to as GTFS in this document for its format, General Transit Feed Specification).

In the first phase of the project, CUSP will develop algorithms to measure or estimate certain types of events associated with bus vehiclesâĂŹ trips using publicallyÂŋ available data, such as departures from terminals, vehicle arrivals at stops and interruptions between stops. In the second

• *Center for Urban Science and Progress, New York University.*

phase, CUSP will analyze the resulting data and calculate metrics related to the performance of the buses with respect to their planned schedule. Finally, in consultation with the DOT, CUSP will develop hypotheses about the significance of contextual factors on those performance metrics. The entire process has been and will continue to be documented and flexible in the code, so both data quality assessment and reliability measurements are more clearly evaluated.

2 LITERATURE REVIEW AND RELATED PREVIOUS WORK

BUS service reliability has been widely noticed since it has a great influence on the daily life of public. Large number of studies have been conducted to measure bus reliability as well as analyze factors with influence. The work of Sterman and Schofer (1976) was among the early studies on bus service reliability in the United States. Using data from bus services in the Chicago area, the study aimed to test the inverse of the standard deviation of point-to-point travel times, which is a particular measurement of reliability. Although found to be useful and easy-collected, the form of reliability measurement is significantly degraded by increasing the route length, intensity of intersection control, traffic volumes, and, with less certainty, bus passenger loadings. Abkowitz and Engelstein (1983, 1984) studied factors affecting the running time on transit routes and methods for maintaining transit service regularity. The proposed method for maintaining service regularity through improved scheduling and real-time control was found to be a reasonable solution to increase reliability. Based on bus data in Portland, Oregon, Strathman and Hopper (1993) presented an empirical assessment of factors affecting the on-time performance of the fixed route bus system. The logit model results showed that the probability of on-time failures increased during PM peak periods, with longer headways and higher levels of passenger activity and as buses progress further along their routes. Nakanishi (1997) described NYCTâĂŹs new bus performance indicators.The Performance Indicator (PI) program was established in 1994 in response to the MTA Inspector. GeneralâĂŹs research recommending the need for measures of service reliability other than the traditional Terminal On-Time Performance (TOTP).

An exhaustive list of transit reliability measure examples was shown in the Transit Capacity and Quality of Service Manual (TCQSM) (Kittelson et al., 2003). On-time performance and headway adherence, the most widely used reliability measures in the transit industry, were especially discussed. Taking into account the interaction between the network performance and passengersâĂŹ route choice behavior, Yin et al. (2004) developed a generic simulation-based approach to assess transit service reliability.Three types of the reliability, system wide travel time reliability, schedule reliability, and direct boarding waiting-time reliability, were defined from perspectives of the community or transit administration, the operator, and passengers. Lu and Ismutulla (2006) set up a model that contained the transferring via three public transport routes with different running time reliabilities. The study suggested that the on-time/punctuality performance and headway evenness are primary focuses in the practice of bus reliability analysis. Xumei Chen, Lei Yu, Yushi Zhang and Jifu Guo (2009) established a system to measure bus reliability on stop, route and network levels based on bus service data in Beijing. The results indicate low service reliability for buses in Beijing and a high correlation between service reliability and route length, headway, distance from the stop to the origin terminal, and the provision of exclusive bus lanes.

3 DATA DESCRIPTION

3.1 Schedule

MTA develops bus schedules for 3-4 months at a time. Schedule planning works with operating departments to ensure schedules align with resource constraints, meet the organizationâĂŹs operating metrics, and maximize service to customers. Schedules are valid for date ranges included in the publications, or until a superceding schedule is released for the same time period.

Schedule data is published according to the General Transit Feed Specification, a standard established in 2006 and now widely used by transit agencies and developers. One transit feed is essentially a small relational database, containing a minimum of six tables and any of seven optional tables. Basic required data in a transit feed file are routes, trips, stops, stop times, and effective date ranges. MTA does not include optional metadata that can be used to distinguish multiple publications covering the same schedule period. The median duration of a trip in the schedule is 44 minutes, but approximately 5 percent are shorter than 10 minutes.

3.2 Operations

Buses are equipped with Automated Vehicle Location systems, which combine GPS and other positioning technology with a wireless transponder in order to report vehicle information at some frequency back to a central database.

In 2012, Dead Reckoning Units on earlier buses was upgraded. Dead-reckoning sensors use direction/bearing and distance/speed to determine relative location from a fixed point. Compasses, odometers, and inertial platforms (gyroscopes and accelerometers) are all dead-reckoning sensors. The newer installed units by Cubic were more accurate than the earlier model by Veriphone.

The vehicle movement Operator login includes route and headsign. However trip reference and phase are inferred automatically on the server side.

"MTA BusTime tries to assign buses to blocks- a sequence of trips that start and end at a depot. This allows the system to make a statement about what a bus will do after it reaches the end of its current trip. However, there is not always enough affirmative and corresponding evidence to make such a strong statement. In this case, MTA BusTime falls back to a trip-level assignment, where it just picks a trip from the schedule that is representative of the route and stopping pattern that the bus is likely to pursue."

The SIRI API now reflects this distinction as described here and in other items below. If the assignment is block-level, the new BlockRef field of the MonitoredVehicleJourney is present, and populated with the assigned block id.

Table 1: List of elements in real-time response for each vehicle.

Element	Used
RecordedAtTime	Yes
LineRef	Yes
DirectionRef	No
DataFrameRef	Yes
DatedVehicleJourneyRef	Yes
JourneyPatternRef	Yes
PublishedLineName	No
OperatorRef	No
OriginRef	No
DestinationRef	No
DestinationName	No
SituationRef	No
VehicleLocation	Yes
Bearing	No
ProgressRate	No
ProgressStatus	Yes
VehicleRef	Yes
MonitoredCall.StopPointRef	Yes
MonitoredCall.VisitNumber	No
MonitoredCall.StopPointName	No
MonitoredCall.Distances.StopsFromCall	No
MonitoredCall.Distances.CallDistanceAlongRoute	Yes
MonitoredCall.Distances.DistanceFromCall	Yes
MonitoredCall.Distances.PresentableDistance	Yes
OnwardCalls	No

Occasionally, no data is available for any vehicles during a long period of time, indicating a problem with the real-time database or the interface used by CUSP for extraction of the data.
Short trips are obscured disproportionately by the long breaks in the data feed. 72% of days contained at least one break of longer than 30 minutes, which almost always occurred during weekday peak traffic times of 8am-10am and 6pm-8pm; 30% of days contained three or more such breaks. These long breaks results in no data for more than 25% of scheduled trips, on certain days.

3.2.1 Holes in the data

Many elements of the reported data are inferred automatically upon processing and storage to MTAâĂŹs database. As these inferences are made real-time, there are irregularities in that may be improved with additional information, or by processing longer periods analytically rather than transactionally.
Major disruptions to service are clearly identifiable based on the number of vehicles reporting throughout the day. A large area of the Northeast was hit by a storm on January 26, 2015, with blizzard conditions lasting overnight and leaving more than 2 feet of snow in some cases. On January 27, 2016, the number of vehicles reporting

throughout the day was less than 50% of the average over comparable time periods. News outlets and press releases are available to corroborate the cause of the large deviation observed. This allows a straightforward decision to eliminate the time period entirely from any analysis.
However, small disruptions are more difficult to detect and research.

4 METHODOLOGY

The project has been structured to have the following milestones:

- Bibliography review;
- Data extraction;
- Data quality assessment;
- Estimation of the departure and arrival times at bus stops and other locations (in other words, interpolation);
- Measurement of bus performance and reliability metrics, with a flexible implementation;

4.0.2 Details on interpolation

The Bus Time data framework includes a variable that indicates the distance along each route where each station is located, and the distance from each ping to the next station. It is fairly simple to calculate the location of each ping along the route by subtracting these two features. After this calculation, a linear interpolation and extrapolation exercise was done by first interpolating the bus pings (time and position) for all trips along one route in a sample day, and then inferring, with extrapolation, the time at which each bus was at each stop.
The AVL system that generates the data automatically detects the nearest station to a each point. At the beginning and end of each trip there is a ”tail” and a ”head” accounting for the time that the bus takes to leave the warehouse and actually start the route, and to return. This was treated by choosing only the first point of the tail and the last point of the head.

4.0.3 Details on Big Data techniques

The total size of our Dataset is around 3 terabytes for SIRI API data and 5 Gigabytes for the schedule data . In order to fully utilize the data, we must deem it as Big Data Challenge since the more data we use, the more accurate result and less bias we can achieve.

Terminology

- HDFS
  - HDFS (Hadoop File System) is a Java-based file system that provides scalable and reliable data storage, and it was designed to span large clusters of commodity servers.
  - The data must be first uploaded into HDFS to perform big data operations.
- Apache Spark
  - Apache Spark is an open source cluster computing framework that performs parallelized stream computing using multiple CPU cores. It is proved to be the fastest open source framework (100 times faster than Hadoop).
- Spark SQL
  - Spark SQL is a Spark module for structured data processing. Spark SQL relies on Spark Dataframe to operate, while Spark Dataframe is a column structured data collection that can be easily saved to csv for further analysis.
  - Spark SQL is efficient and reliable and SQL files are easy to share and cooperate.
- Data Schema
  - The original data is stored in nested json format with lots of redundant information. The schema function based on Spark Dataframe allows us to understand the data structure in a more distinguishable fashion.
  - We can manually set data schema along with data type to fasten the operation
- CSV to Dataframe Tool by Databricks
  - Databricks published the tool that can directly read CSV file into Spark DataFrame.
- Time Series for Spark(interpolation)
  - A Scala / Java / Python library for interacting with time series data on Apache Spark.

Operation

- Data parsing and Extraction
  - After uploading the data onto HDFS, we can then submit Spark Script to perform data parsing and extraction.

- The first step of data parsing is learning the data schema in order to select the right elements to parse by after reading the JSON files into Spark DataFrame.
- The details of element selected are stored in separate SQL query script, which will be submitted along with the Spark Script simultaneously
- After we extract the elements we need, we temporarily keep the data in DataFrame.

- Data Cleaning and Storing
  - Because the data was originally stored in JSON array format, the data extracted for each element is a list of arrays. In order to use those data, we must first flatten the arrays by Applying Flatmap function on the DataFrame.
  - The data extracted from SIRI has prefixes such as *MTA NYCT_*. In order to further merge with the schedule data, we must remove those prefixes by applying map on replacing the prefixes as empty strings.
  - After the aforementioned processing, we save its outcomes into CSV format by mapping and joining each item.
- Measuring Data Coverage
  - As previously mentioned in section 3, We noticed that a portion of trips had not been recorded in the SIRI dataset. With the extracted dataset, we could easily check the data coverage for each line for each day.
  - We achieved the measurement by groupby on data and bus line and count on trips to get how many trips actually get recorded for each bus line and day.
- Data Merging and manipulation
  - By applying interpolation using spark time-series tool on the extracted SIRI data, we can calculated the actual stop times of each stop for each bus line. With the calculated stop times for each bus line, we can then calculate the actual headways on each stop for each line.
  - Finally, we merged the actual stop times with scheduled stops. We then calculated the delays and differences of headways at each stop between actual time and scheduled time.

4.0.4 Performance indicators - reliability

Based on formal research, we choose the most common indexes to measure the reliability of bus performance: Wait Assessment, On Time Performance, Running Time Adherence and Headway Regularity.

Wait Assessment is a metric used by New York City Transit, defined in the Transit Capacity and Quality of Service Manual as the percentage of actual headways between successive vehicle arrivals that are less than or equal to a given standard. On Time Performance, which compared the actual arrival time with the scheduled arrival time directly, is a universal wide indecator for bus reliability performance. Both running time aherence as well as headway regularity are defined as the average difference between scheduled and actual, normalized by the schedule data. The higher the running time metrics the worse the running time adherence. A high headway metric value indicates poor headway regularity adherence. Bus bunching is an extreme example of short headway.

1) **Wait Assessment (PI):**
   It is defined as the percentage of observed service intervals that are no more than the scheduled interval plus 3 minutes during peak (7 a.m. - 9 a.m., 4 p.m. - 7 p.m.) and plus 5 minutes during off-peak (12 a.m - 7 a.m., 9 a.m. - 4 p.m, 7 p.m. - 12 a.m.). Wait Assessment is a simple calculation that can be performed after all headway calculations have been performed for a given location.

$$WT_i = \frac{WT_i}{T_i} \times 100\%$$

   where WT_i is number of actual intervals between buses that are no more than the scheduled interval plus 3 minutes during peak hours (7 a.m. - 9 a.m., 4 p.m. - 7 p.m.) and plus 5 minutes during off-peak hours of bus route i. That is:

$$WT_i = card\left(\{bus : bus \in route_i, t_{bus_0} - t_{bus} < T\}\right)$$

   where,
   $card$ is the dimension of a set.
   $route_i$ is the set of all buses in a bus route i.
   $t_b us$ is the time of arrival at a particular station.

bus_0 is the nearest bus ahead of bus.
T equals 3 minutes during peak hours (7 a.m. - 9 a.m., 4 p.m. - 7 p.m.) and 5 minutes during off-peak hours.

2) **On Time Performance (OTP):**
It is defined as the positive difference between actual arrival time and schedule arrival time. However, During low-frequency period, on-time performance is more important while during the high frequency period, the headways matter more.

$$OTP_{ij} = |aat_{ij} - sat_{ij}|$$

where,
aat_{ij} is the jth actual arrival time of bus route i.
sat_{ij} is the jth scheduled arrival time of bus route i.

3) **Running Time Adherence (RTA)**:
(measured in %) is defined as the average difference between the actual and the scheduled running times relative to the scheduled running time.

$$RTA_i = \frac{1}{n} \times \left(\sum_{j=1}^{n} \frac{|art_{ij} - srt_{ij}|}{srt} \right) * 100\%$$

where,
art_{ij} is the jth actual running time of bus route i.
srt_{ij} is the jth scheduled running time of bus route j.
n is the number of stop route i has.

4) **Headway Regularity (HR)**:
(measured in %) is defined as the average difference between the actual and the scheduled headways relative to the scheduled headway.

$$HR_i = \frac{1}{n} \times \left(\sum_{j=1}^{n} \frac{|ah_{ij} - sh_{ij}|}{sh} \right) * 100\%$$

where,
ah_{ij} is the jth actual headway of bus route i.
sh_{ij} is the jth scheduled headway of bus route j.
n is the number of stop route i has.

5 Results and implications for practice

5.1 Data extraction and comparison to current DOT process

Daily files of stored response content from the Bus Time API, totalling approximately 3 terabytes, were successfully parsed and converted into one table containing only the useful elements, totalling approximately 63 gigabytes. It is important to highlight that data extracted from Bus Time is different than the data contained in flat files used by DOT, and not purely supplemental. Table 1 lists the differences along several metadata dimensions.

Table 2: Differences between DOT and Bus Time data.

	DOT flat file	Bus Time API
Source database	Archived	Real-time
Sample frequency	30 seconds	Limited by reliability of interface (max 30 seconds)
Spatiotemporal elements	Raw NMEA, including speed*	Only time and location (projected onto shapeline)
Trip elements	Route and status only	Includes inferred elements, like Next Stop and Trip ID

5.2 Data quality assessment

The following is an overview of the reliability of the whole dataset. By visualizing the size of the data (for example, in terms of JSON length or number of vehicle records) as a time series and correlating with data from the schedule, we can identify unreasonable variations. Finally, some tactical examination of data elements reveals irregularities that must be considered during the performance measurement phase.

Missing data

The data covers a total number of 318 different bus lines and 340 days in a year. So, although this project tried to focus on the whole year data (from 2015-01-01 to 2015-12-31), some days (listed in Figure 1) are entirely missing.
From the list of missing days, December is the month with the most missing data, which contains 21 days without data. It is necessary to find out what factors cause this problem. If it is caused by some uncontrollable factors such as weather, some mitigation plan may be devised. But if it is caused by some human factors of systems factors, it should be avoided in the future.

```
['3/8/15', '7/12/15', '7/13/15', '10/11/15', '12/9/15', '12/10/15', '12/11/15', '12/12/15', '
12/13/15', '12/14/15', '12/15/15', '12/16/15', '12/17/15', '12/18/15', '12/19/15', '12/20/15'
, '12/21/15', '12/22/15', '12/23/15', '12/24/15', '12/25/15', '12/26/15', '12/27/15', '12/28/
15', '12/29/15', '12/31/15']
```

Figure 1: Missing days in the dataset.

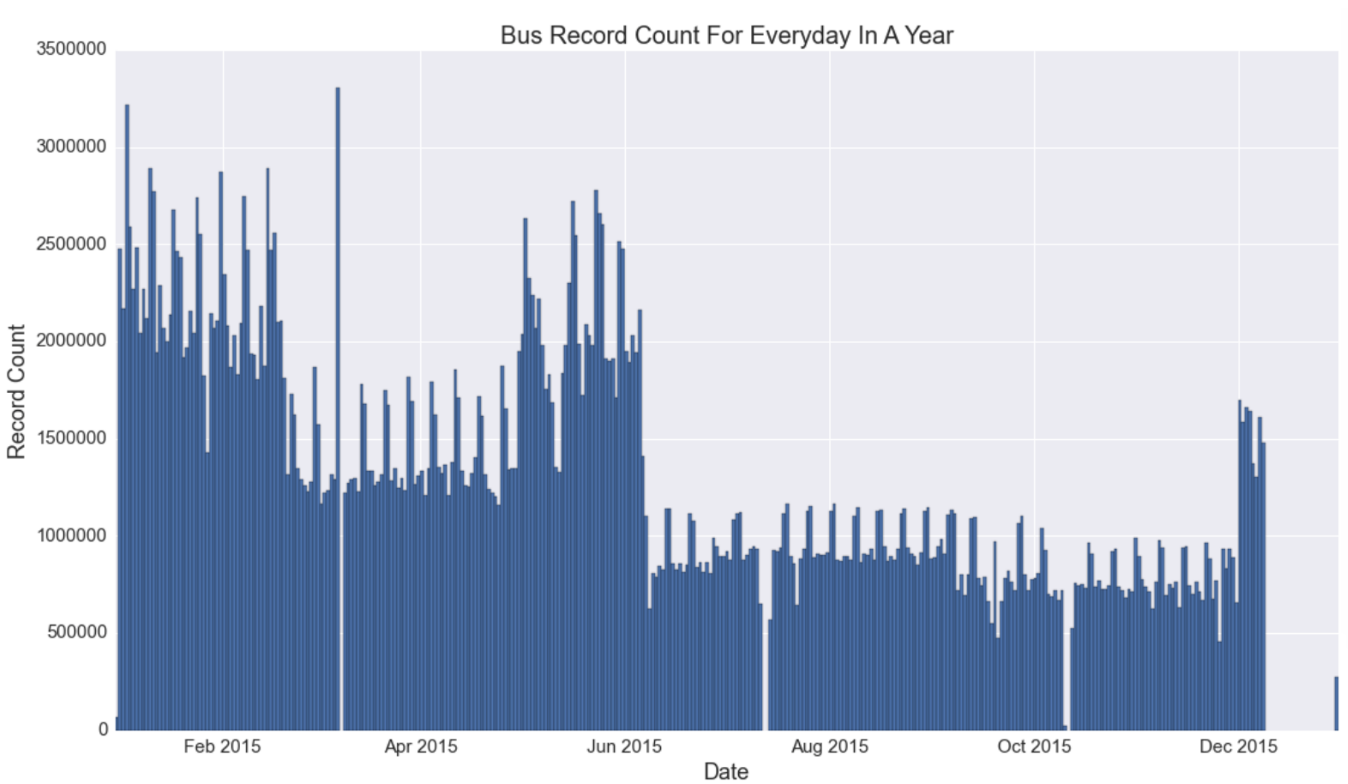

Figure 2: Visualization of records throughout the year.

Visualization of records throughout the year

The total number of bus records by date are shown in Figure **??**.
From the plot, some regularities are immediately apparent, including the seven-day cycle. However there are four obvious changes throughout the year. The first one is in February, second in May, third in June, and the last one in December. Further analysis is needed to find out these factors affecting the changes and could help with the bus schedule planning. Also, it can find that March 7th has an extremely high record but March 8th is a day without data which do not exist in other missing days. One can infer that data for March 8th were merged with March 7th.

Daily record counts by hour

From the plot, it can find that weekdays have the same trend and weekends have the same trend. On weekdays, the record count regularly decreases as the morning progresses before peaking twice in the middle of the day. In fact, one would expect the opposite based on typical characteristics of urban mobility: one peak during the morning rush-hour and one during the evening rush-hour.

Data density with respect to level of scheduled activity

As the bus is transmitting every 30 seconds, the interval between Bus Time records is expected to also be 30 seconds so long as the vehicle is operating with the AVL equipment activated. Figure 4 shows the actual intervals from the dataset collected. In fact the typical interval turns out to be 60 seconds, with a significant portion of even longer intervals. We examine these long intervals by comparing a measurement of total vehicle activity between the Bus Time data (in

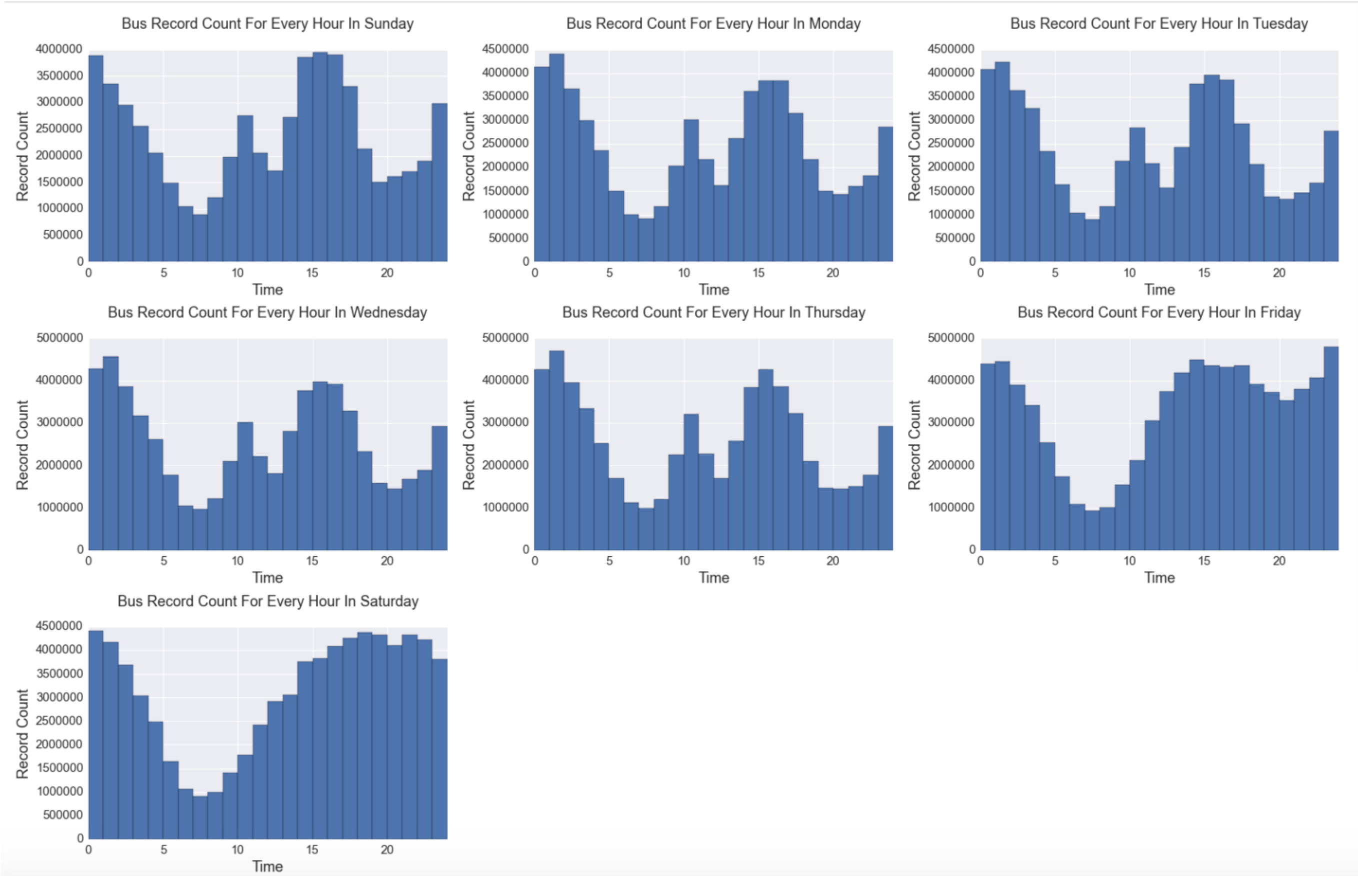

Figure 3: Bus records count for day of the week in 2015.

terms of record count) and the schedule data (in terms of aggregate running time of all scheduled trips). The measured activity from the two sources turn out to be strongly anticorrelated in the months analyzed (in the table below), whenever materially non-zero. In other words, generally when scheduled activity increases, the density of data available in our dataset decreases.

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct
-0.83	-0.49	-0.63	-0.72	-0.60	0.08	-0.61	-0.97	-0.52	0.02

To further diagnose the issue, the comparison is made on a finer temporal scale (6-minute timesteps) for a few random dates. Figure 5 shows very different patterns when overlaying the level of scheduled activity onto the level of reported activity derived from the Bus Time data. On one of the two weekdays examined (the first and third), two large gaps are clearly visible and span the entirety of the two daily rush-hour peaks in the schedule. On the weekend date examined, a short gap is visible in the late evening.

The time series analysis showing counter-cyclical data density, the strong anticorrelation, and the tactical examination of a few dates all lead to the conclusion that the dataset collected by NYU CUSP cannot be used for performance analysis of an entire year without major risks to accuracy, in the form of both bias (as discussed in section 0.3.5: Arrival time estimation techniques) and variance (due to the most typical interval being twice as long as specified by MTA). To continue with the demonstration of applying the data for performance metrics, we identified a week with the best density and limited gaps during rush hours (2015-12-01 to 2015-12-07).

Tactical validation of individual elements

Coordinates estimated using the Global Position System include some level of noise due to factors such as atmospheric conditions and multi-path effects. However the location elements of the Bus Time data appear to be transformed to adhere to shapelines of the reported bus route, eliminating any variation laterally across the street. Figure **??**

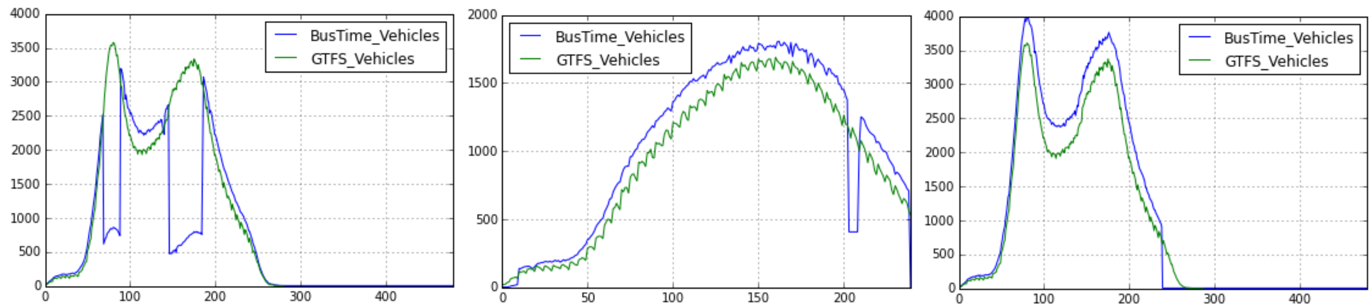

Figure 5: Comparison of active vehicle count for three example dates.

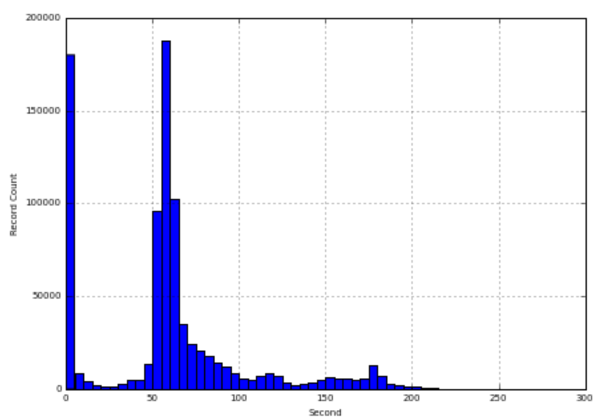

Figure 4: Distribution of intervals between SIRI response records.

shows an overlay of points from both Bus Time records and GTFS shape file data, for a short stretch of the B26 route in Fort Greene. While the coordinates of stops are on the sides of the street, no lateral variation is observable in the Bus Time coordinates, even though the vehicles are known to travel along all four lanes of that specific roadway. Some error may remain along the orthogonal plane (the basis for distance calculations), but GPS-assistive equipment installed after the initial launch of Bus Time uses gyroscopes and speedometer readings to reduce this error to an immaterial level.

The only elements reported by the vehicle itself are the time, location and headsign (the route and destination text displayed above the front of the vehicle). All of the remaining elements are inferred by the Bus Time server. Because those inferences are made real-time, the software has no opportunity to retrospectively validate and correct errors. This analysis validates two key features by checking for two metadata properties: that reported trip IDs are unique to each vehicle and that stop distances are unique for each shape. The first validation failed while the second property was confirmed. Figure 7 is a sample summary of records records grouped by both trip and vehicle, for a sample date and route; it shows the duplication of trip ID across multiple vehicles, the number of records associate with each duplicate, and the timespan of those records.

The conclusions are that the Bus Time server sometimes makes errors in inference of the vehicle's trip ID, the inferred trip ID is not persistent for a vehicle (that is, it "flip flops"), but the reported stop distances for the inferred shape are persistent. The variation in inferred trip ID is an important consideration in subsequent performance measurement because so much of the analysis requires grouping or sorting data (both input and output) by trip.

5.3 Demonstrations of the methods applied to the dataset

The practices of transportation planning and analysis rely heavily on vehicle arrival-time data. The first phase of this report explained the retrieval of real-time AVL data and application of methods for estimating those arrival times in actual operations. The following list demonstrates some typical analytic techniques and discusses their validity when applying these data, given the known limitations to its density and accuracy. Included in the input data are arrival times from the schedule, published in the widely-adopted General Transit Feed Specification, but discussion of its data generation process is not in scope. Generally, many high-level performance metrics are simple ratios expressing the proportion of

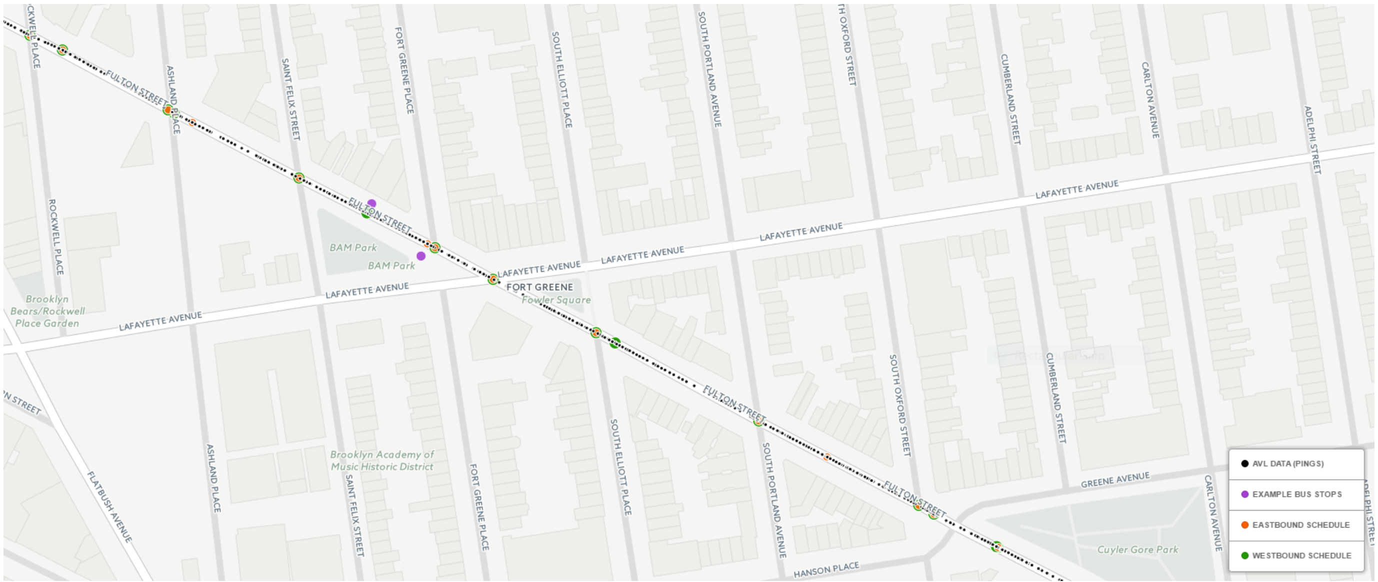

Figure 6: Sample of B25 route.

events (such as a vehicle arrival, or completed trip) that meet some criteria (such as arriving within 5 minutes of its scheduled time). The problem with these binary measurements is that the methods ignore the shape (referred to in mathematics as higher moments) of a distribution - for example, a "long tail," or if the distribution is multi-modal.

- **Distribution of headway**: In higher-frequency transit routes, arrivals are considered reliable when the headway is more consistent and closer to customers' expectation. (The ideal distribution is 100% density at the expected value, that is, no deviation. Our definition of Headway Adherence is binary and allows for some deviation). The distribution of headway for a less-than-reliable service is not always a normal bell-curve. This is because of the tendency for delayed buses to become even more delayed, as the larger number of waiting passengers increases dwell times, which in turn reduces overall travel speed. When bunching occurs, the headways of the "bunched" buses (i.e., those that closely follow a preceding delayed bus) approach zero, while there is no theoretical upper limit for the headway of the delayed bus.
- **Bunching rate**: Variations in dwell time and variations between-stop travel time related to traffic and operator behavior are the major causes of vehicle bunching along a route (Gellei 2010). Measurement of dwell time and travel time will be discussed later in this section. The resulting bunching condition can be measured. For this analysis, we consider the bunching condition to occur when headway is less than one minute. The bunching ratio is the percentage, at a certain stop, of arrivals under bunching condition compared to total vehicle arrivals. This is very similar to Headway Adherence except that it measures the tail of the distribution, not the middle. Bunching ratio can be calculated and compared for a variety of stops and routes, as shown in Figure **??**.
- **Distribution of running time adherence:** Best practice in schedule planning is to forecast running time using historical data, but exclude outliers. An outlier often represents an occurrence of some enroute incident (such as a police action or a parade) and should not influence planning a typical-day bus schedule. Outliers skew upward the distribution of actual running times, since there is a physical upper limit to the vehicle's speed, but no limit to the number and severity of enroute incidents. Because the schedules are created based on historical distributions excluding outliers, the resulting error, defined as running time variance, will tend to be positive. The error can be mitigated by artificially increasing the planned running time (for

			N	time_range
TRIP_ID	**trip_date**	**vehicle_id**		
FB_D5-Weekday-SDon-047200_B49_15	2015-12-03	MTA NYCT_5125	4	00:06:54
FB_D5-Weekday-SDon-051000_B49_15	2015-12-03	MTA NYCT_4855	1	00:00:00
		MTA NYCT_5125	13	00:24:51
		MTA NYCT_7146	2	00:04:14
FB_D5-Weekday-SDon-051200_B49_21	2015-12-03	MTA NYCT_7146	5	00:06:22
UP_D5-Weekday-SDon-006000_B1_1	2015-12-03	MTA NYCT_4877	40	00:40:23
UP_D5-Weekday-SDon-009800_B1_1	2015-12-03	MTA NYCT_4877	56	00:47:00
UP_D5-Weekday-SDon-010000_B1_2	2015-12-03	MTA NYCT_4893	52	00:45:10
UP_D5-Weekday-SDon-013800_B1_2	2015-12-03	MTA NYCT_4893	51	00:41:10
UP_D5-Weekday-SDon-014000_B1_1	2015-12-03	MTA NYCT_4877	42	00:33:08
UP_D5-Weekday-SDon-017800_B1_1	2015-12-03	MTA NYCT_4877	53	00:44:01
UP_D5-Weekday-SDon-018000_B1_2	2015-12-03	MTA NYCT_4893	43	00:33:46
UP_D5-Weekday-SDon-021800_B1_2	2015-12-03	MTA NYCT_4893	40	00:31:30
		MTA NYCT_7179	1	00:00:00
UP_D5-Weekday-SDon-022000_B1_1	2015-12-03	MTA NYCT_4877	50	00:41:21
UP_D5-Weekday-SDon-025500_B1_3	2015-12-03	MTA NYCT_4990	31	00:24:55
UP_D5-Weekday-SDon-025800_B1_1	2015-12-03	MTA NYCT_4877	39	00:31:26
		MTA NYCT_7179	7	00:05:17
		MTA NYCT_7191	1	00:00:00

Figure 7: Duplication of Trip ID across several vehicles

example, by including the outliers, or adding some arbitrary value), but it is generally not cost- effective to do so, or impacts the reliability of subsequent trips operated by the same vehicle or operator.

One example descriptive analysis is to plot the distribution of normalized running time variances. Visible in this example, Figure **??**, is both the central tendency – slightly positive – and the wide spread of the distribution, indicating very inconsistent running time adherence. Strategies to reduce the inconsistency are not in the scope of this project.

- **Performance with respect to vehicle dis-**

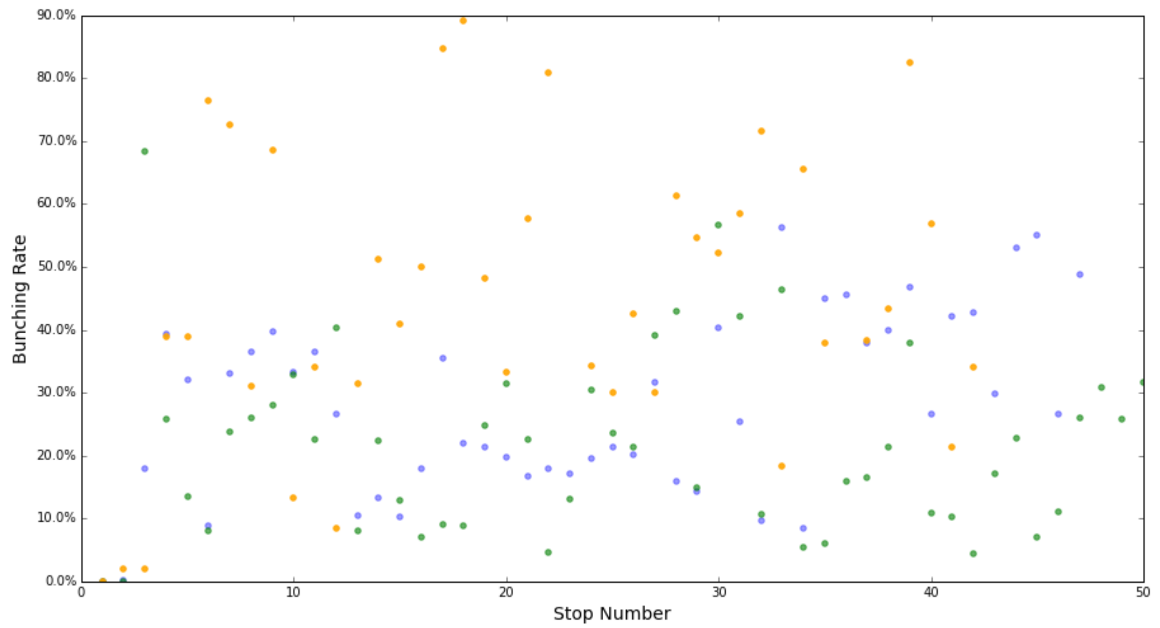

Figure 8: Headway Bunching Rate at Stops for Three Brooklyn Bus Routes.

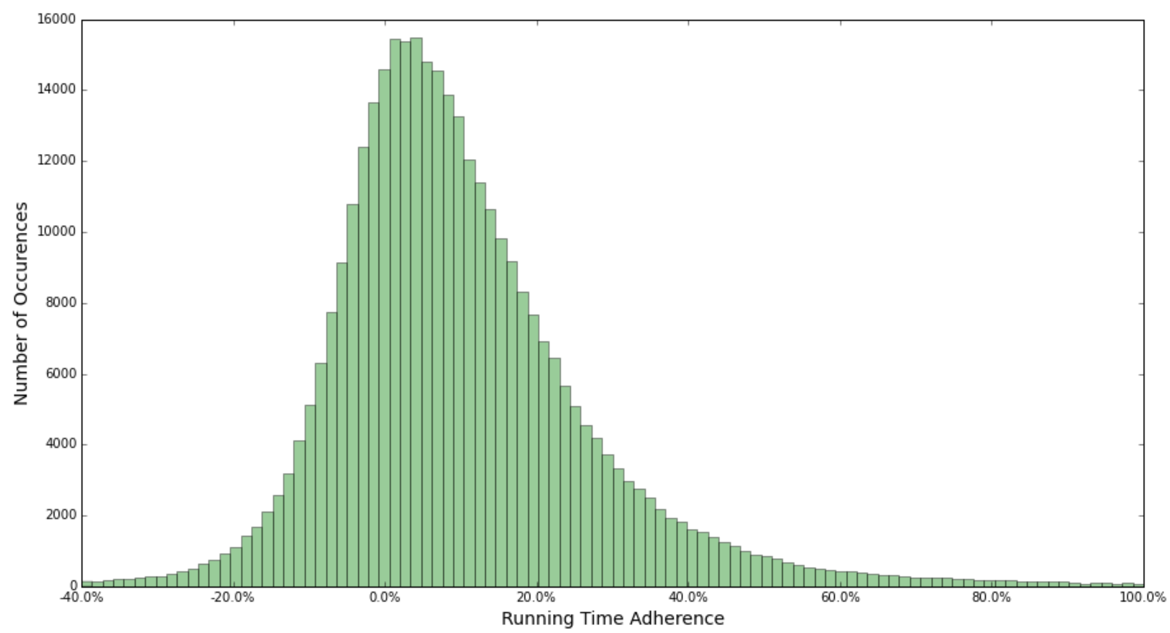

Figure 9: Running time adherence.

tance along route: It is generally accepted in the theory of urban public transportation that longer routes have worse reliability, as defined by the metrics discussed so far. Another analytic technique is to plot one of the metrics for a given route (or, more specifically - one shape variation of a route). Figure **??** is the summary of an ordinary

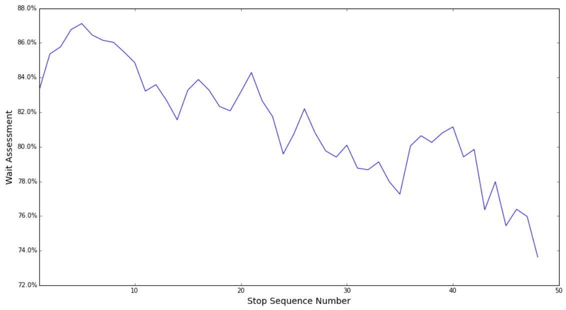

Figure 11: Average Wait Assessment by Stop Sequence Number, Route B41.

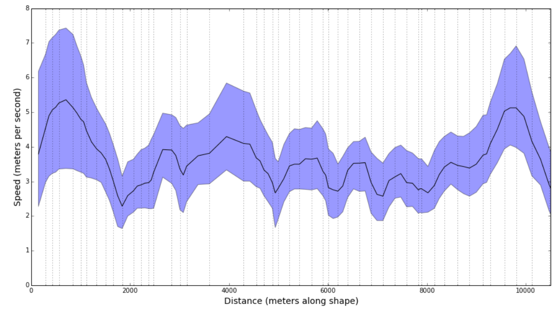

Figure 12: Moving speed distribution.

least squares (OLS) regression, taking Wait Assessment as the dependent variable and the vehicle's progression along the shape (in terms of number of stops made) as the independent variable. The resulting parameter value has strong statistical significance, rejecting the null hypothesis that route length has no relationship to performance. The example in Figure **??** both supports that conclusion and suggests which segments along the route contribute most to the decline.

- **Spatial distribution of travel speeds:** Because Bus Time records contain discrete time and location, speed calculations are difference-based averages, not instantaneous (or quasi-instantaneous) samples. The other challenge in descriptive statistics about speed at fixed location(s) is that the sequential observation points of multiple vehicles are not aligned; for example it is extremely unlikely that multiple vehicles record the "ping" from exactly 100 meters and again at exactly 200 meters along a route-shape. However re-sampling is possible if mean speeds are treated as continuous curves with respect to distance. The new distribution at a point is defined as the collection of mean speeds of all vehicles passing that point. **??** demonstrates the changing moments of the distribution over the length of a route-shape, along with grid lines indicating the stop locations.

- **Spatial distribution of slow/stopped condition** - A list of slow/stopped events can be created by identifying the beginning and end location and times for each occurrence of the condition. Naturally, many of these events will be at stop locations. That subset of the slow/stopped events may theoretically support analysis of dwell time, however even when using data having the maximum density possible according the data generation process (every 30 seconds), analysis of dwell time at individual stops is not possible; previous research suggests that the majority of stops have dwell time of less than 30 seconds (Pangilinan 2011). The remaining subset are slow/stopped events not occurring at stop locations. High spatial density of these events indicates a recurring problem with traffic flow, which in turn increases travel time and contributes to occurrence of bunching condition. (In order to exclude the normal interaction with traffic signals, the events can be filtered to only include those above some minimum, related to traffic signal cycle lengths).
Figure **??** shows the estimated percentage of vehicles in slow/stopped condition and many points along a route-shape. While some extreme values are apparent, they are generally at stop locations and the variance of the remaining points is minimal, even at other stop locations. This suggests that the density of the data may be insufficient to identify interruptions without a high rate of false negatives.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:        Wait Assessment   R-squared:                       0.290
Model:                            OLS   Adj. R-squared:                  0.290
Method:                 Least Squares   F-statistic:                     6701.
Date:                Fri, 22 Jul 2016   Prob (F-statistic):               0.00
Time:                        14:31:59   Log-Likelihood:                 25037.
No. Observations:               16421   AIC:                        -5.007e+04
Df Residuals:                   16419   BIC:                        -5.005e+04
Df Model:                           1
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [95.0% Conf. Int.]
---------------------------------------------------------------------------------
Intercept         0.8405      0.001   1141.083      0.000         0.839     0.842
stop_sequence    -0.0021   2.51e-05    -81.860      0.000        -0.002    -0.002
==============================================================================
Omnibus:                       95.810   Durbin-Watson:                   0.489
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               97.389
Skew:                          -0.189   Prob(JB):                     7.12e-22
Kurtosis:                       3.009   Cond. No.                         52.7
==============================================================================
```

Figure 10: Correlation of Wait Assessment and Stop Sequence Number.

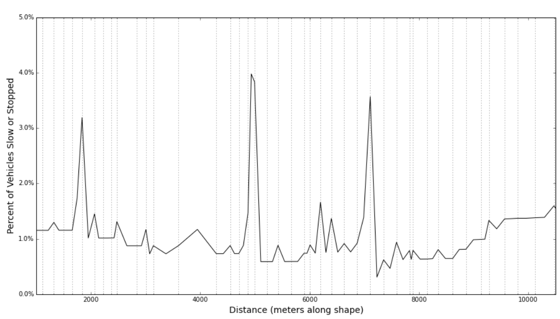

Figure 13: Stopped or Slow Vehicle Percentage.

6 LIMITATIONS OF THE ANALYSIS AND FUTURE WAYS TO ADDRESS THEM

6.1 Data

On the one hand, the huge size of data requires big data techniques. The biggest challenge for our approach to be implemented in practice by cities is that of processing large amounts of raw data into information because it requires big data techniques for anything beyond small samples (i.e. one line, one day, etc.). Even processed datasets being structured can demand more than what can be offered by single/dual core applications. Data supporting higher-level analysis techniques can be managed with any off-the-shelf database system, including *sqLite* (open source SQL) or even a series of CSVs. Furthermore, the large amounts of data require a special environment to deal with. Since the raw data could not be dealt with in local machines because of its volume, we used CUSP's cluster and Data Facility to process it.

It is worth noting that, while the MTA does not publish archived Bus Time data (except for a sample in 2014), we are providing the code for anybody to collect it, and this is an important step that must be kept into account when discussing about privacy. Besides this, the Vehicle ID field is not anonymized, so the contributions of our project are subject to have unlikely yet feasible implications regarding tracking of individuals, to the extent they can be identified in a relatable data source.

6.2 Accuracy

Accuracy of the location and time data, when reported, was not in question for the purposes

of this project. However, some of the elements are inferred and may cause errors if not handled properly. Limitations are generally the result of information gaps, specifically when no data is reported when or where expected. When no data is reported, it is unknown if there was actually an operating bus, but a problem arose with hardware or software. Also, if there was a bus (for example, present in previous and latter data points), it is possible that the bus did not follow the scheduled route.
There are also some limitations from the data that ends up being recorded:

- There are trips with data from more than one vehicle, which means that trips use a combination of reported trip_id and vehicle id.
- Uncertainty in dwell time.
- Lack of instantaneous speed information.
- Lack of schedule meta data.
- Since the system automatically infers trips and stations, there are signals showing disappearing and reappearing buses.

6.3 Frequency

The ideal interval for every continuous collected time stamp is 30 seconds. But there exists a large amount of data points intervals over 30 seconds or even over an hour. That may be due to delays in signal collection.
In order to analyse both the vehicle behaviour on individual street segments and intersections, and the dwell times, 60-90 second frequency is insufficient. Thirty seconds are also insufficient for the measurement of dwell time.
As the interval can not be change or fix in the dataset, this project reported this situation. Alos, as the dataset contain a whole year information, the trend and regularity can be obvious even though the interval is not within 30 seconds. In the future analysis, if GPS can get a more accurate data collecting, that will make the project analysis more accurate. Bus travelling data is necessary for future DOT planning decisions and bus behaviors analyzing.In order to improve the data collecting process, it could improve the sensibility of sensors.

6.4 Gaps

Frequent, long gaps in the data make unbiased performance analysis impossible over full days or date ranges, because the gaps tend to arise at the same times each day. Any statistical measurement involve time-of-day as an input variable will be biased. Some analysis may leave time-of-day as an unobserved component of the error term and still show significance in other variables, such as distance-along-trip or street features. Incomplete reporting of trips may cause bias in calculation of trip-based measurements, such as running time and On Time Performance, if not normalized. Just as mentioned in Accuracy sections, over-30-second interval may affect the accuracy of the analysis. Inferences of points beyond the range of data need to be flagged specially given the possibility that the vehicle did not operate the trip as planned. If better density is useful, DOT should request the additional elements in the static files provided by MTA

6.5 Tecnology

As the dataset for this project is extremely huge, the only way to make the process more faster is using big data technology. That would need the help of CUSP data facilities.So the main advantages of our approach rely on the fact that it is based on open source software (such as Python, HDFS and Spark) with a wide and increasing support community around them, but also on the fact that our code is simple to share and reproduce. Since Big Data software is relatively new, we kept most of the data manipulation in simple SQL or Python scripts, which is specially convenient for the client.

6.6 Reproducibility

The main risk identified by our approach is reproducibility in terms of data fetching and of framework dependency. Our data feed depends on a public API that is always subject to interruptions, and the Big Data processing requires the HDFS management system and Spark. It may take time and computational and technical resources to handle them, and deprecation is always a hazard. Despite of this, those potential sources of failure have been proven to be increasing in robustness and trustworthiness during the last years.

7 Conclusions and future research

7.1 Conclusions

This project studies the possibility of calculating performance metrics for bus performance in New York City based on the GPS data offered by MTA. The MTA bus GPS database collects the location of each bus every 30 seconds. After estimating the departure and arrival time for each bus at each stop, we use measurement like headways and wait assessment to assess bus performance and reliability. Taking one-year data as our object of study, the data is extremely huge so that big data techniques (mainly Spark) are widely used in this project. In order to make our work easy to share and reproduce, we choose to use the SQL API to decide on the parsing of the data, which enables the DOT to easily make changes by editing the SQL script. Positive aspects of our approach include the fact that we focused this analysis on the viewpoint of the DOT; the fact that we were able to successfully process large volumes of data; the fact that we identified pitfalls and challenges of using the MTA Bus Time data for the purposes of the sponsor agency and the fact that we enable a flexible implementation of different methods to estimate times and locations, as well as to measure schedule-reliability performance. Potential pitfalls and areas of improvement are the fact that we used formal metrics that were not tailored for the DOT nor New York City, and the abundance of assumptions that inevitably scale up, both with and without bias, in the form of measurement errors. Similar methodologies and concepts could be used in other cities and traffic models. The big data technology applied in not only traffic analysis but also city wide urban issue offers more reliable results compared with traditional sample studies.

7.2 Recommendations to DOT

Use of the public API for real-time archiving of Bus Time data mimics an ongoing process already performed internally at MTA, and can result in an incomplete view of the same data. Moreover, the static file currently used by DOT, presumably provided by MTA, contains both greater temporal density and some additional, potentially useful elements that are not available through the Bus Time API. The results of this project suggest that combining the advantages of both data acquisition processes would yield the ideal dataset needed for these analyses. The DOT may request that MTA include the additional, useful vehicle monitoring elements in the file provided directly to DOT and therefore eliminate temporal gaps that arise due to problems with the public API. The subsequent interpolation and performance analysis steps do not require the user to implement a so-called big data platform if Bus Time data is processed in smaller batches (for example, one full day) as it accumulates. Once headway, running time, and vehicle speed calculations are performed and output stored with proper indexing, the performance metrics can be dynamically analyzed using a typical personal computer.

7.3 Suggestions for future work

- Inference study of the effect of several traffic or design conditions on reliability.
- Autocorrelation to identify recurring patterns in reliability with respect to time (for example, every Monday morning)
- Integrate and Improve the current visualization tool developed by NYU CUSP, and adapt it to be responsive and interactive to different queries.
- Further optimize the algorithms hereby used for the Big Data portion of the work.
- Generalization of code for further applications:
  - Other transit system data feeds may use a different interface standard and as a result a different data structure- AVL systems in some cities collect stop departure and arrival time directly, instead of location at fixed time intervals).
  - Other transit modes, for example subway or bike-share, may contain different features specific to that mode.

Author contributions

All authors work contributed extensively to the work presented in this paper with discussions and communications with mentors and sponsor. Matthew P Urbanek extracted the data using MTA BusTime API and did the data quality analysis for the whole year data with Bonan Yuan. Sara Arango is responsible for the arrival time interpolation. Bonan Yuan parsed all the extracted data from the whole year json files and processed the analysis on the large dataset by Spark. Yuqiao Cen did the gap record analysis. Jiaxu Zhou developed the performance assessment metrics.

REFERENCES

[1] W. H. Lin and J. Zeng, "An experimental study of real-time bus arrival time prediction with GPS data," Transp. Res. Rec., no. 1666, pp. 101- 109, Jan. 1999.

[2] Rajat Rajbhandari, Steven I. Chien, and Janice R. Daniel, ”Estimation of bus dwell times with APC information”, Transportation Research Record 1841 Paper No. 03- 2675, 2003.

[3] Min-Tang Li, Fang Zhao, Lee-Fang Chow, Haitao Zhang, and Shi-Chiang Li, ”Simulation Model for Estimating Bus Dwell Time by Simultaneously Considering Numbers of Disembarking and Boarding Passengers”, Transportation Research Record 1971, 2006.

[4] LI Fazhi, Yang Dongyuan, and Ma Kai, ”Bus Rapid Transit (BRT) Bunching Analysis With Massive GPS Data,” National Science and Technology Support Program of China (NO. 2009BAG17B01), 2013.

[5] Brian Levine, Alex Lu, and Alla Reddy, ”Measuring Subway Service Performance at New York City Transit: A Case Study Using Automated Train Supervision (ATS) Track-Occupancy Data,” TRB 2013 Annual Meeting, 2013.

[6] Dan Wan, Camille Kamga, Jun Liu, Aaron Sugiura, and Eric B. Beaton, ”Rider perception of a âĂŸlight’ Bus Rapid Transit system - The New York City Select Bus Service,” Transport Policy 49 (2016) 41âĂŞ55, Apr. 2016.

[7] Yiming Bie, Xiaolin Gong, and Zhiyuan Liu, ”Time of day intervals partition for bus schedule using GPS data,” Transportation Research Part C 60, pp. 443âĂŞ456, Sep. 2015.

[8] Jeremy S. Safran, Eric B. Beaton, and Robert Thompson, ”Factors Contributing to Bus Lane Obstruction and Usage in NYC: Does Design Matter?” TRB 2014 Annual Meeting, 2014

[9] Christopher Pangilinan, Nigel Wilson, and Angela Moore, ”Bus Supervision Deployment Strategies and Use of Real-Time Automatic Vehicle Location for Improved Bus Service Reliability,” Transportation Research Record: Journal of the Transportation Research Board, No. 2063, Transportation Research Board of the National Academies, Washington, D.C., 2008, pp. 28âĂŞ33. DOI: 10.3141/2063-04.

[10] Yingxiang Yang, David Gerstle, Peter Widhalm, and Dietmar Bauer, ”The Potential of Low-Frequency AVL Data for the Monitoring and Control of Bus Performance,” TRB 2013 Annual Meeting, 2013.

[11] Shi An, Xinming Zhang and Jian Wang, ”Finding Causes of Irregular Headways Integrating Data Mining and AHP,” ISPRS Int. J. Geo-Inf. 2015, 4, 2604-2618; DOI:10.3390/ijgi4042604, Nov. 2015.

[12] Jinil Chang, Mohamad Tala, and Satya Muthuswamy, ”A Simple Methodology To Estimate Queue Lengths at Signalized Intersections Using Detector Data,” TRB 2013 Annual Meeting, 2013.

[13] Christopher Pangilinan and Kristen Carnarius, ”Traffic Signal Timing for Optimal Transit Progression in Downtown San Francisco,” San Francisco Municipal Transportation Agency, San Francisco, CA, 2011.

[14] Simon Reed, ”Transport for London Using Tools, Analytics and Data to Inform Passengers,” JOURNEYS, September 2013. https://www.lta.gov.sg/ltaacademy/doc/13Sep096-Reed_TfL-InformPassengers.pdf, accessed 20 Jul 2016.

Appendix A
Sample visualization

Figure 14: On time performance aggregated by stop over 2015.