CAPSTONE PROJECT FOR THE MASTER OF SCIENCE IN APPLIED URBAN SCIENCE AND INFORMATICS

BUS RELIABILITY METRICS USING PUBLIC MTA BUS TIME DATA

July 22, 2016

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

Advisors:
Dr. Kaan Ozbay
Dr. Huy T. Vo

Sponsor Agency:
Department of Transportation, City of New York

Center for Urban Science and Progress
New York University

Contents

BRIEF

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 customer perceptions of dependability. The *New York City Department of Transportation* (NYC DOT) wished to engage the *New York University Center for Urban Science and Progress* (CUSP) to explore the use of public vehicle location data from the *Metropolitan Transit Authority* (MTA) Bus Time to generate operational data relevant to the DOT's planning decisions. This information is provided in the form of reliability metrics for bus service.

Based on the MTA *Automated Vehicle Location* (AVL) data and its public Bus Time API, the team performed a data assessment analysis for the data generating process and the data collection process. We also deliver methods for estimating bus travel and stop times, measuring reliability with different metrics, and settles the ground for the DOT to identify the distribution of reliability measurements as a function of factors regarded as relevant to their practice.

AKNOWLEDGEMENTS

We would like to express our sincere appreciation to all who have lent us hands during this time. First of all, we would like to show our sincere gratitude to our sponsor agency, the New York City Department of Transportation, for giving us an opportunity to focus on this important city issue. We would specially like to to thank Jeremy Safran, who gave us plenty of insight into the way the DOT wants to treat and understand bus data analytics. Secondly, we would like to express our gratitude to our advisors, Dr. Kaan Ozbay and Dr. Huy T. Vo., not only for their technical expertise but for their generosity and willingness to lead us through this process. Last but not least, we would like to thank Kai Zhao, Abdullah Kurkcu, Ender Morgul, and our classmates in CUSP who gave us assistance and advice during this process. We will always be thankful.

0.1 OVERVIEW AND BACKGROUND

0.1.1 Statement and Description

Although demand for public transportation in New York City is growing, bus ridership levels are declining. Many reasons can explain this, including drops in vehicle speeds and customer perceptions of dependability. This project focuses on identifying these two factors along the city's MTA operated bus routes, in the form of reliability metrics that are relevant to the DOT as the municipal traffic authority.

The DOT keeps records of the MTA bus system, but the agency lacks a formal process for compiling and analysing it according to their decision making capabilities (such as those related with road design and traffic management), which substantially differ from the MTA's purposes and reach. Despite the MTA (which is the agency in charge of operations) has internally defined metrics used for scheduling, planning and analysis of the bus service, this agency pays more attention to the bus level instead of the whole system level, which is more of the interest of the DOT.

To help the DOT improve their planning efficiency, this project aims to measure the dependent variable metrics related to bus performance and reliability, enabling the agency to, in a further analysis, understand the effect of independent variables that affect service quality. CUSP developed methods for estimating bus travel times and measuring reliability, while performing data quality analysis of the MTA Bus Time API data (from which the data was collected to this exercise) and the MTA schedule data (referred to as *GTFS* in this document for its format, *General Transit Feed Specification*).

0.1.2 Goals

The goals of the project can be summarized as follows:

1. Perform a thorough data quality assessment on the MTA Bus Time data.
2. Implement metrics related to the performance of the buses with respect to their planned schedule.
3. Document the entire process and deliver flexible code, so both data quality assessment and reliability measurements are more clearly evaluated by the agency.

The project has been developed following the milestones below:

- Bibliography review;

- Data extraction;
- Identification of pitfalls and irregularities in the data generating process;
- Estimation of the departure and arrival times at bus stops and other locations (in other words, extrapolation);
- Measurement of bus performance and reliability metrics, with a flexible implementation;

0.2 PROJECT OFFERINGS

The contributions from our work can be summarized as follows:

1. The estimation of departure and arrival times based on AVL data for the specific case of the Bus Time API records.
2. Flexible and potentially novel bus performance metrics for the data set in question.
3. Quality assessment for the Bus Time API dataset.
4. Flexible and reproducible code to allow further implementations and variations of our analysis.

From the SOW, include the deliverables here.

0.3 APPROACH

In this section we discuss the data parsing, processing and reliability measurement techniques.

0.3.1 Data techniques

MTA's AVL data is acquired using a "get" request to a web service offered by the MTA to the public. The service uses a data standard called *Service Interface for Real-Time Information* (SIRI).

[Discuss parameters, API response time, and dependency constant internet connection. Also discuss accumulated size (each is 3MB, so an entire year can be >3TB).]

Per MTA's recommendation, the response data is requested in JSON format (*Javascript Object Notation*). JSON offers a flexible structure, for example allowing elements to be stored in hierarchies, or a combination of named and unnamed elements. JSON does not transform directly to a tabular layout and as a result cannot be imported by typical data analysis tools like Microsoft Access or Excel without first parsing it.

Parsing can be performed using a variety of approaches. A small program that can be written on almost any personal computer (a macro, or command-line script) may be acceptable for parsing one JSON response, requiring in the order-of-magnitude one second for each JSON. However this is likely unacceptable for reading and aggregating any meaningful amount of archived JSON response files (entire days, or entire lines over multiple days). Advanced techniques requiring additional software or hardware (to be discussed further on in this report) can significantly improve processing time by distributing the data across multipler processing units. Regardless of the technique, the extracted data has a straightforward structure, containing text and numerical elements (which can be stored as text), appropriate for storage in a comma-separated values (CSV) file.

A risk that remains to be investigated is incomplete (blank) vehicle data elements. Faster techniques for parsing of JSON into tabular format require each observation to contain data according to each element extracted. This results in rectangle-shape data set with no need for validation of each row.

]Discuss GTFS. Bus schedules for a given trip include a reference date]

The SIRI standard calls for date-time representation according to the ISO 8601 standard. While this can be read directly and used for many typical calculations (for example, elapsed time), it poses a problem when performing any analysis requiring date or time data from the planned schedule for the buses. Those analysis require both the âĂIJtrue" date-time

element as well as the trip reference date. Here are two example comparisons of a bus trip's estimated departure from its first stop and the corresponding scheduled departure time. Such a comparison may be used to explain whether bus reliability can be attributed to operational issues such as late departures from the depot. Note the conversion.

trip_id	Time stamp	Converted time stamp	Scheduled time
OH_B6WeekdaySDon077600_M101_100	20160613T13:09:32.00004:00	0d13:09:32	13:08:55
MV_B6WeekdaySDon041500_M5_206	20160614T01:20:42.00004:00	1d01:20:42	25:12:15

Note that time zone information is dropped in this conversion. While this may become problematic if these approaches are applied outside of New York City, the only related risk in this application is to the validity of data from bus trips that occur during Daylight Savings Time shifts. We elect to deal with those issues tactically rather than add complexity across the entire data set.

0.3.2 Big Data Techniques

HDFS HDFS 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 The total dataset size for one year is over 2 terabytes and stored in nested json format, which is impossible to analyse by any traditional way. Apache Spark is an open source cluster computing framework that performs parallelized stream computing using multiple CPU cores. Apache Spark 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.

JSON ELEMENT(schema)	Column NAME	Explanation
LineRef	ROUTE_ID	Name of bus line(B42)
VehicleLocation.Latitude	latitude	latitude of record
VehicleLocation.Longitude	longitude	longitude of record
RecordedAtTime	recorded_time	What time it gets recorded
VehicleRef	vehicle_id	ID of vehicle
FramedVehicleJourneyRef.DatedVehicleJourneyRef	TRIP_ID	Same as trip_id in GTFS*
FramedVehicleJourneyRef.DataFrameRef	trip_date	Date of the trip
JourneyPatternRef	SHAPE_ID	Same as shape_id in GTFS*
StopPointRef	STOP_ID	Id of next stop,Same as stop_id in GTFS*
Extensions.Distances.DistanceFromCall	distance_stop	Distance to next stop
Extensions. Distances.CallDistanceAlongRoute	distance_shape	Stop_s total distance along the shape
Extensions. Distances.PresentableDistance	status	Report the current status of bus to next stop [1]
DestinationRef	destination	Headsign of bus

0.3.3 Measurement techniques

Three methods are investigated for estimating a vehicleâĂŹs arrival time at a certain location, such as a bus stop. Arrival time estimates are the basis for most measurement and identification related to bus reliability, such as vehicle speeds, block times, and headway (which in turn is required for wait assessment).

The first method is use to use a spatial algorithm to identify any data points within a certain radius of the location, and make an estimation from within that subset. Possible calculations on that subset are to take earliest or latest time recorded, the median time recorded, or apply an interpolation to generate a point-estimate at the exact location. The second method is to apply an interpolation algorithm to all observations from a given vehicle on each trip. The tradeoff is that this may be more computation than required if not many along the trip are to be estimated.

The third method is to apply the interpolation only for stops that are reported in Bus Time as a âĂIJMonitored Call,âĂİ that is, the upcoming stop. The advantage of only interpolating the reported stops is that it generates less bias for the stop times by ignoring the empty records. Filling the gaps for missing records will assume the bus making uniform motions while in reality the bus speed changes a lot due the complicated traffic condition and traffic lights especially in New York. The tradeoff for the third method is that since many trips are not completely recorded, we may lose a lot of information for the missing part.

Average vehicle speed is calculated as the difference between estimated arrival times at two points divided by the distance traveled. A key assumption is that vehicle motion is monotonic (never reversing direction) along any given axis. Approaches to account for variations in vehicle heading remain to be explored. The size of the post-processed average

speed data set depends on how many point-pairs within each trip are to be calculated. A dataset reporting the total time (âĂIJblock timeâĂİ) and average speed, using only the beginning and end points, would be manageable for most typical desktop application; an entire year would be less than 1GB and could be organized by month, route etc. for easier access.

Calculation of headway also requires estimates of all vehiclesâĂŹ arrival times at a certain location. Missing estimates (whether due to the data generating process or the estimation algorithm) must be identified, as skipping them may result in an overstatement, based on the time between two or more vehiclesâĂŹ arrivals. A compiled, processed dataset supporting headway measurement – that is, containing arrival times of each trip at each stop – is approximately 2M rows per day of operations. While the calculations on selected data are simple enough for any typical desktop application, the total size of the database (50-100GB) would require additional hardware and a database management system.

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. The wait assessment for bus is only measured on weekdays. 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. âĂŞ 9 a.m., 4 p.m. âĂŞ 7 p.m.) and plus 5 during off-peak (12 a.m âĂŞ 7 a.m., 9 a.m. âĂŞ 4 p.m, 7 p.m. âĂŞ 12 a.m.) For example the standard for transit lines operating at XXX scheduled headway is +3 mins during peak hours and +5 mins during off-peak hours. Peak hours are defined as 6am to 9am and 4pm to 7pm. Wait Assessment is a simple calculation that can be performed after all headway calculations have been performed for a given location.

OTP(On Time Performance) is defined as the absolute value of decrement between actual arrival time and schedule arrival time. It reflects how close the actual arrival time is compared with schedule arrival time directly. Different from the measurement of on-time performance percentage by MTA, which is the percentage reflects the number of buses that arrive within a certain time before or after the published schedule, we use the distribution of a group of data to describe the on-time performance for a single trip. The main reason is the criteria is not provided by MTA for buses. However, During low-frequency period, on-time performance is more important while during the high frequency period, the headways matter more. OTP = |Actual Arrival Time - Schedule Time|

Running Time Adherence (measured in

Similarly, headway regularity (measured in If two consecutive buses are further from (or closer to) each other than the scheduled headway, the difference is called a longer (or shorter)

headway difference. Bus bunching is an extreme example of short headway. The definition equations for headway regularity metrics are shown in the following equations:

0.3.4 Big Data techniques

Due to the large volume of data, which is about 3 Terabyte, we decided to apply big data techniques for the data. We apply Apache Spark along with some Spark SQL techniques for manipulating the data in large scale. It takes around 30 - 45 minutes for the processing the entire year data depending on the performance of the server.

0.4 RESULTS AND IMPLICATIONS FOR PRACTICE

Table 1: 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

0.4.1 Data quality assessment

In this part, it is trying to find out whether the data are reliable or not. It is a overview of the whole dataset. By visualizing the date, it can find out that whether it is reasonable or not.

Missing data

The first step to figure out the data quality is to know the information that the data contain. By summarizing the data get from all the jsons files by using Spark, it can find that the data cover a total number of 318 different bus lines and 340 days in a year. So though this project tried to focus on the whole year data (from 2015-1-1 to 2015-12-31), there still existed missing data for some days. Here shows all the missing days in Figure XXX.

Figure 1: Missing days in the dataset.

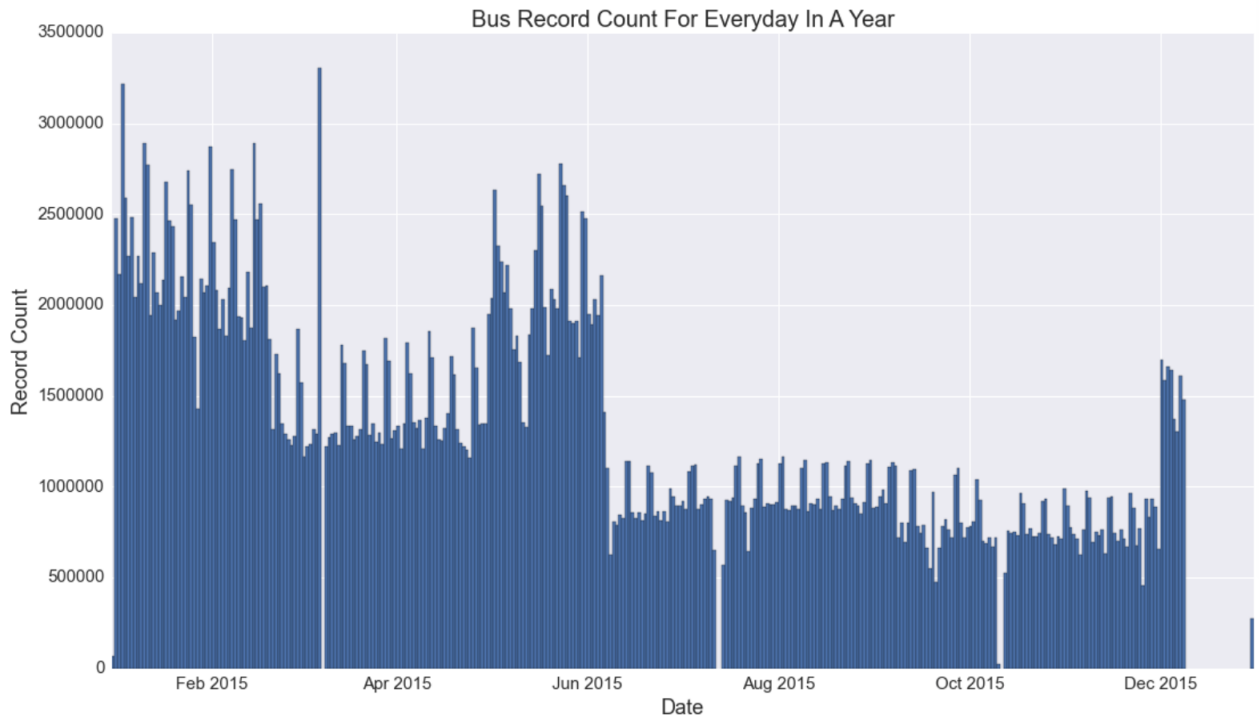
```
['3/8/15', '7/12/15', '7/13/15', '10/11/15', '12/9/15', '12/10/15', '12/11/15', '12/12/15', '
12/13/15', '12/14/15', '12/15/15', '12/16/15', '12/17/15', '12/18/15', '12/19/15', '12/20/15'
, '12/21/15', '12/22/15', '12/23/15', '12/24/15', '12/25/15', '12/26/15', '12/27/15', '12/28/
15', '12/29/15',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

From the list of missing days , it can find that December is a month with the most missing data, which contains 21 days without data. So 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 could come out. But if it is caused by some human factors of systems factors, it should be avoided in the future.

Visualization of records throughout the year

Here shows the whole year respond record in Figure XXX.
From the plot, it can find that there are some regularities existing. Normally, seven days is a cycle, and weekdays have less record count than weekends. But January, February and

Figure 2: Bus records count for all days in 2015.

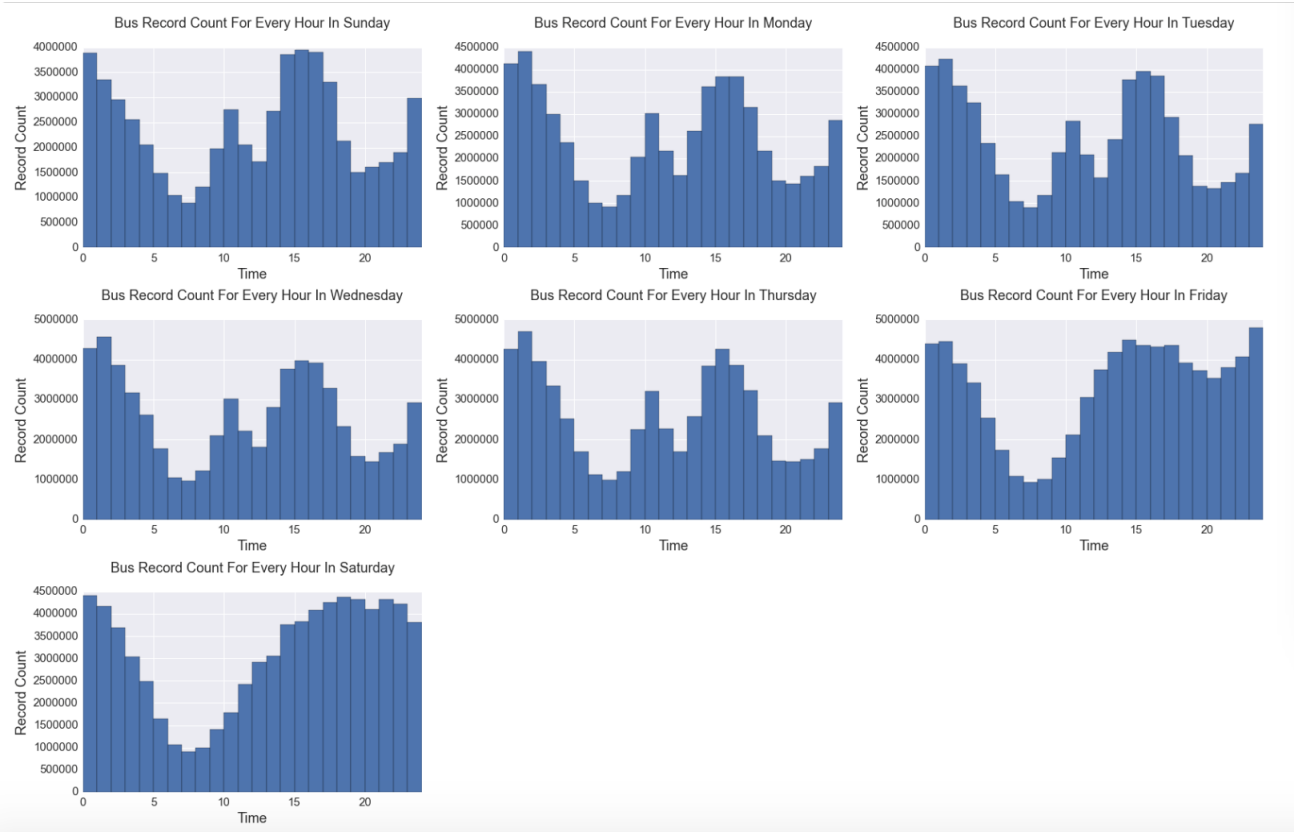


May are not that obvious. And as there are some missing data in December, the regularity in December is not obvious. There exist four obvious change in this plot. The first one is in February, second in May, third in June, and the last one in December. It may exist some factors that affect these change. Further analysis is needed to find out these factors 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. It can infer that there exist some possibility that data for March 8th may merge with March 7th.

Daily response records

Here shows the everyday respond record in Figure XXX.
From the plot, it can find that weekdays have the same trend and weekends has the same trend. In weekdays, 9am to 11am would exist another peak in a day. And after 5pm, bus records declines regularly. It is reasonable because 9am to 11am and 5pm are the rush hours in a day. And in weekends, after 8am, bus records increase regularly. Based on the visualization of the bus records for everyday and for every hour in a day, it can find that the data are quite reasonable and the further analysis based on these data would be reliable.

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



Gaps (time dimension)

As the bus is transmitting every 30 seconds, for the first part of gaps analysis, it find out which hour in a day has the least 30-second interval. Here shows the 30-second interval distribution for a day in Figure XXX.

From the plots, it can find that 9am is the time that has the least bus respond record interval within 30 seconds. That may exist some gap in 9am. So for further analysis, it will compare the data with GTFS data. Here shows the focuses of future gap analysis and some example about compare between AVL data and GTFS data. Missing observations occur XXXX often XX% of gaps occur when there is a gap in the API response time, although we cannot rule out an error in the AVL signal Missing observations ARE/ARE NOT correlated toâĂę. Charts summarizing number of vehicles with data, compared to what is expected based on GTFS

LOTS OF PLOTS

Figure 4: Bus records count for day of the week in 2015.

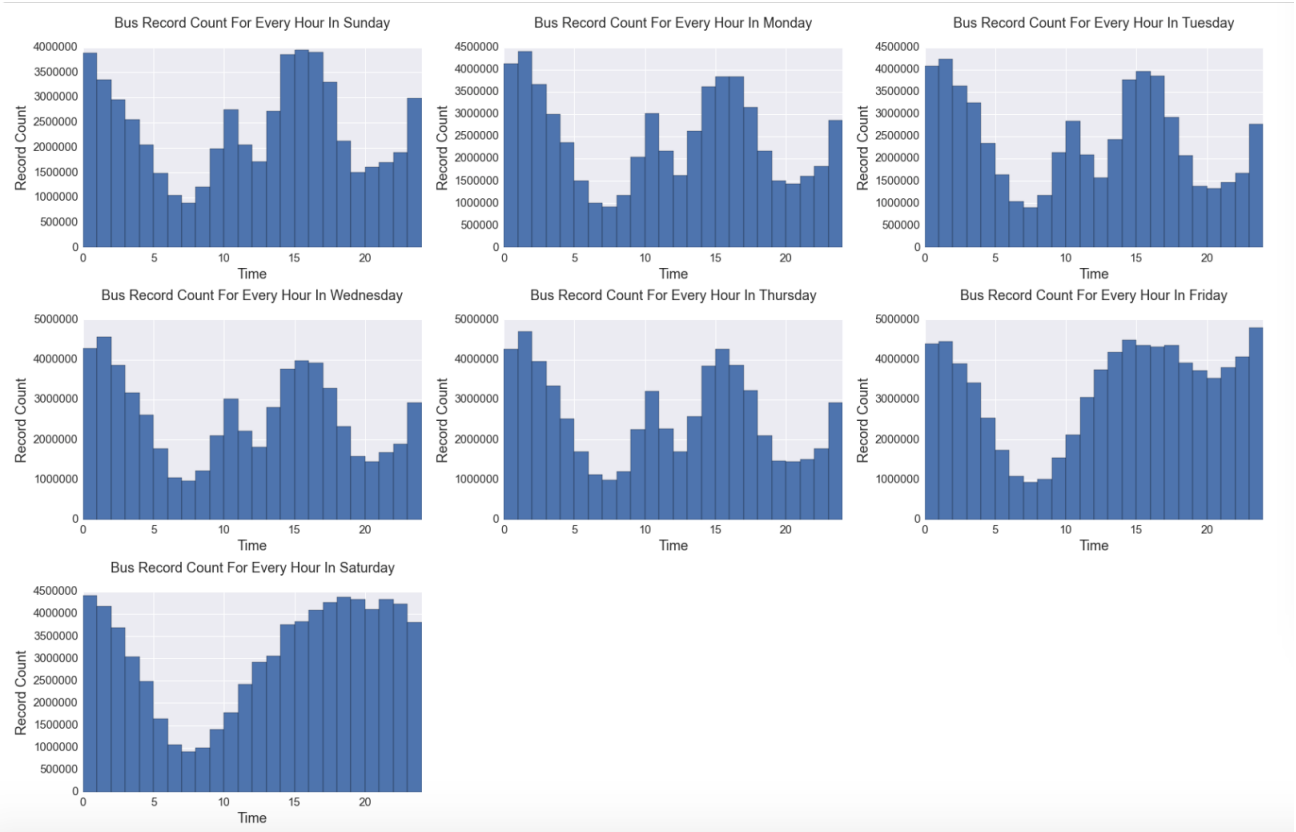


Figure 5: Original sample extraction from DOT.

1001	MTA NYCT_M101	0	DEADHEA|$GPRMC,002005.610,A,4047.23656,N,07356.79964,W,000.0,030.7,011214,,,E*7A
1001	MTA NYCT_M101	0	DEADHEA|$GPRMC,002037.609,A,4047.23656,N,07356.79964,W,000.0,030.7,011214,,,E*73
1001	MTA NYCT_M101	0	DEADHEA|$GPRMC,002109.609,A,4047.23656,N,07356.79964,W,000.0,030.7,011214,,,E*7F
1001	MTA NYCT_M101	0	DEADHEA|$GPRMC,002140.609,A,4047.23656,N,07356.79964,W,000.0,030.7,011214,,,E*72
1001	MTA NYCT_M101	0	DEADHEA|$GPRMC,002520.609,A,4047.23656,N,07356.79964,W,000.0,030.7,011214,,,E*70
1001	MTA NYCT_M101	0	DEADHEA|$GPRMC,002624.609,A,4047.23656,N,07356.79964,W,000.0,030.8,011214,,,E*78
1001	MTA NYCT_M101	0	DEADHEA|$GPRMC,002655.609,A,4047.23656,N,07356.79964,W,000.0,030.8,011214,,,E*7E
1001	MTA NYCT_M101	0	DEADHEA|$GPRMC,002727.609,A,4047.23656,N,07356.79964,W,000.0,030.8,011214,,,E*7A
1001	MTA NYCT_M101	0	DEADHEA|$GPRMC,002759.609,A,4047.23656,N,07356.79964,W,000.0,030.8,011214,,,E*73
1001	MTA NYCT_M101	0	DEADHEA|$GPRMC,002830.609,A,4047.23656,N,07356.79964,W,000.0,030.8,011214,,,E*73
1001	MTA NYCT_M101	0	DEADHEA|$GPRMC,002901.609,A,4047.23656,N,07356.79964,W,000.0,030.7,011214,,,E*7F
1001	MTA NYCT_M101	0	DEADHEA|$GPRMC,002932.609,A,4047.23656,N,07356.79964,W,000.0,030.7,011214,,,E*7F
1001	MTA NYCT_M101	0	DEADHEA|$GPRMC,002957.609,A,4047.23656,N,07356.79964,W,000.0,030.7,011214,,,E*7C
1001	MTA NYCT_M101	0	DEADHEA|$GPRMC,014529.269,A,4047.23656,N,07356.79964,W,000.0,030.3,011214,,,E*78
1001	MTA NYCT_M101	0	DEADHEA|$GPRMC,014600.270,A,4047.23656,N,07356.79964,W,000.0,030.3,011214,,,E*78
1001	MTA NYCT_M101	0	DEADHEA|$GPRMC,014631.269,A,4047.23656,N,07356.79964,W,000.0,030.2,011214,,,E*73
1001	MTA NYCT_M101	0	DEADHEA|$GPRMC,014702.270,A,4047.23656,N,07356.79964,W,000.0,030.2,011214,,,E*7A
1001	MTA NYCT_M101	0	DEADHEA|$GPRMC,014733.270,A,4047.23656,N,07356.79964,W,000.0,030.2,011214,,,E*78
1001	MTA NYCT_M101	0	DEADHEA|$GPRMC,014805.269,A,4047.23656,N,07356.79964,W,000.0,030.2,011214,,,E*7A

0.4.2 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 Automated Vehicle Location data and application of methods for estimating those arrival times in actual operations. The

Figure 6: Our data extraction using Spark.

Line	Latitude	Longitude	Record_at_Time	Vihicle_ID	Trip_Information	Trip_Date
MTA NYCT_Q76	40.753476	-73.780631	2015-12-30T18:59:41.621-05:00	MTA NYCT_7424	MTA NYCT_CS_W5-Weekday-111000_MISC_843	12/30/15
MTA NYCT_B67	40.664322	-73.983724	2015-12-30T18:59:52.000-05:00	MTA NYCT_406	MTA NYCT_JG_W5-Weekday-109900_B6769_24	12/30/15
MTA NYCT_B8	40.635277	-73.960583	2015-12-30T19:00:03.000-05:00	MTA NYCT_430	MTA NYCT_JG_W5-Weekday-109200_B8_40	12/30/15
MTA NYCT_S91	40.590805	-74.158344	2015-12-30T19:00:06.000-05:00	MTA NYCT_8263	MTA NYCT_YU_E5-Weekday-109500_S6191_13	12/30/15
MTA NYCT_Q43	40.709557	-73.796306	2015-12-30T18:59:52.000-05:00	MTA NYCT_6449	MTA NYCT_QV_W5-Weekday-110400_Q43_6	12/30/15
MTA NYCT_BX5	40.837627	-73.825863	2015-12-30T18:59:39.890-05:00	MTA NYCT_5761	MTA NYCT_GH_W5-Weekday-113500_BX5_318	12/30/15
MTA NYCT_Q17	40.761881	-73.82969	2015-12-30T18:59:37.000-05:00	MTA NYCT_8422	MTA NYCT_JA_W5-Weekday-113600_MISC_455	12/30/15
MTA NYCT_BX41+	40.830642	-73.910542	2015-12-30T19:00:02.000-05:00	MTA NYCT_5781	MTA NYCT_KB_W5-Weekday-112000_SBS41_515	12/30/15
MTA NYCT_Q88	40.738192	-73.806908	2015-12-30T18:59:50.401-05:00	MTA NYCT_8024	MTA NYCT_QV_W5-Weekday-110500_MISC_222	12/30/15
MTA NYCT_Q2	40.706824	-73.753472	2015-12-30T18:59:48.000-05:00	MTA NYCT_8484	MTA NYCT_QV_W5-Weekday-112500_MISC_194	12/30/15
MTA NYCT_B54	40.694126	-73.987185	2015-12-30T18:59:57.000-05:00	MTA NYCT_6546	MTA NYCT_FP_W5-Weekday-108900_B54_223	12/30/15
MTA NYCT_M34+	40.74387	-73.973719	2015-12-30T18:59:37.247-05:00	MTA NYCT_5846	MTA NYCT_MQ_W5-Weekday-111200_SBS34_12	12/30/15
MTA NYCT_M31	40.7621	-73.971915	2015-12-30T18:59:56.678-05:00	MTA NYCT_3809	MTA NYCT_MQ_W5-Weekday-111300_M31_31	12/30/15
MTA NYCT_BX6	40.835896	-73.948654	2015-12-30T18:59:38.311-05:00	MTA NYCT_7676	MTA NYCT_WF_W5-Weekday-106900_BX6_43	12/30/15

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 are data of 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 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 (defined 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 XXX.

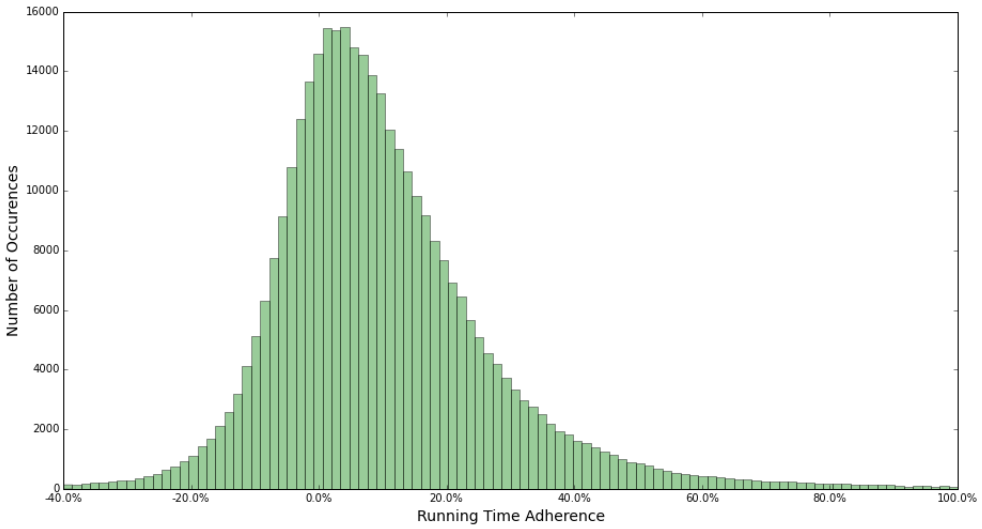
- **Distribution of running time adherence:** Best practice in schedule planning is to forecast running time using historical data, but exclude outliers. An outlier often represents an occurrence of some enroute incident (such as a police action or a parade) and should not be factored into planning a 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 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, 7, 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 distance 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 XXX is the summary of an ordinary least squares (OLS) regression, taking Wait Assessment as the dependent variable and the vehicleâĂŹs 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 XXX 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

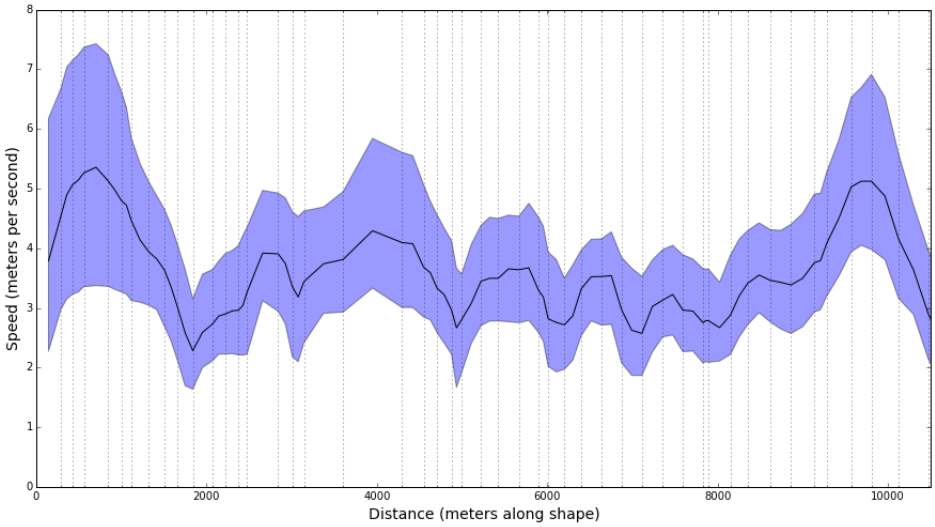
Figure 7: Running time adherence.



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. **Figure XXX** 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

Figure 8: Moving speed distribution.



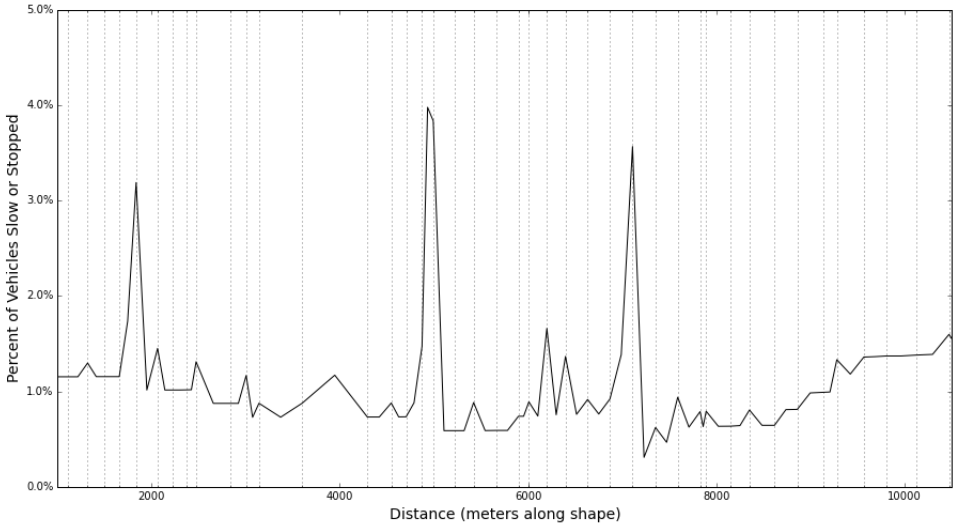
bunching condition. (In order to exclude the normal interaction with traffic signals, the events can be filtered to only include those above some minimum, related to traffic signal cycle lengths).

Figure XXX 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 .

0.4.3 More discussion on the approach

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.
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

Figure 9: Moving speed distribution.



software is relatively new, we kept most of the data manipulation in simple SQL or Python scripts, which is specially convenient for the client.
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
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 Privacy.

0.4.4 Conclusions and future work

Conclusions

This project studies the bus performance in New York City and its correlation metrics 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 figure out the 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. Conclusion for the model we use. (Introduction, Pros and Cons, Bias) Model for time estimation Measurement for bus performance Method for feature selection Conclusion for the results we get. (How the result influence the future operation) Conclusion for the whole analysis process. (Pros and Cons, Bias, Influence)
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 agency (as it was requested from us) 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 in the form of measurement errors.
InfluenceïijŽ
The main goal for us is to figure out the other potential influence factor besides operation to help DOT control and improve bus performance on planning level. Similar methodology and concept could be used in other cities and traffic mode. The big data technology applied in not only traffic analysis but also city wide urban issue offers more reliable result compared with traditional sample studies.

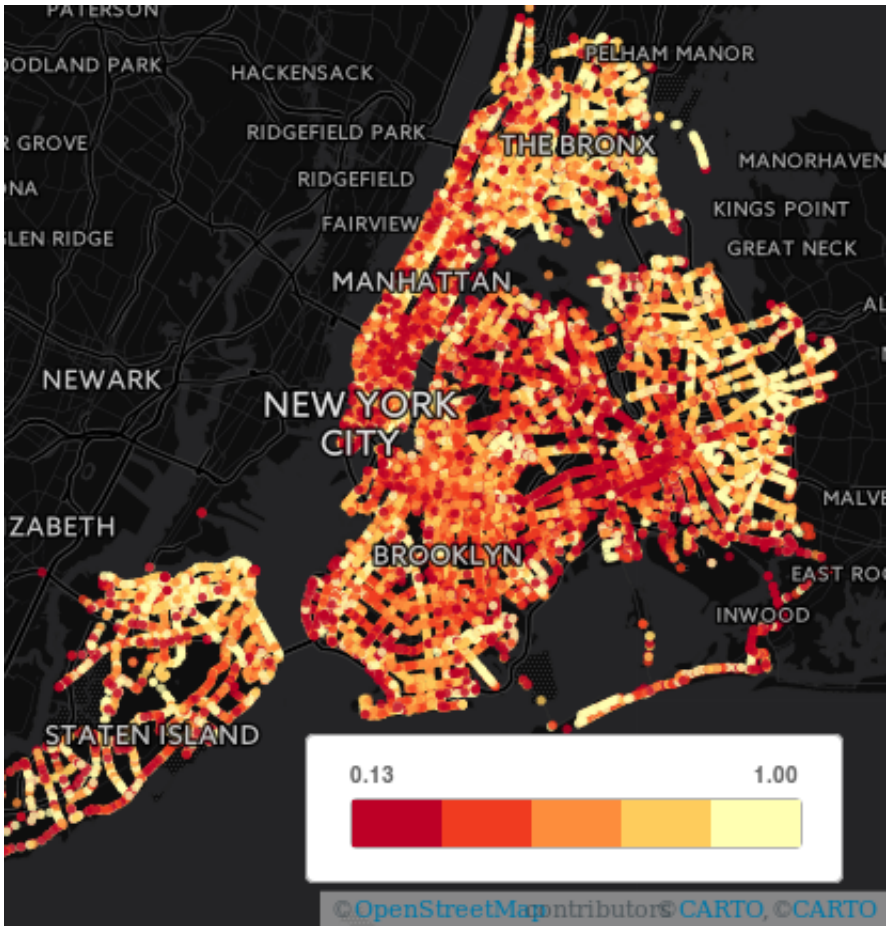
Future work

- Inference study of the effect of several traffic or design conditions on reliability.
- Improve the current visualization tool, and adapt it to be responsive and interactive to different queries.
- Have a more detailed work flow including our code, to be easily implemented into every day tasks in the DOT.
- Further optimize the algorithms hereby used for the Big Data portion of the work.
- Further Applications

- Other cities (need to pay attention to data structure, bus GPS data in some cities collect departure and arrival time instead of location).
- Other traffic mode (eg. Subway, city bike. Need to pay attention to the different features of each kinds of traffic mode).

SAMPLE VISUALIZATIONS

Figure 10: On time performance aggregated by stop over 2015.



SAMPLE RESULTS

Table 2: Sample result metrics file (headshot example).

route_id	stop_id	date	OnTimePerformance	Peak_WaitAss	OffPeak_WaitAss
Q07	350232	1/2/2015	0.831395	0.891566	1
Q25	550032	1/2/2015	0.720737	0.72388	0.760368
Q34	501132	1/30/2015	0.666667	0.910714	0.918699

Bibliography

[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, âĂIJEstimation of bus dwell times with APC informationâĂİ, 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 DAT, National Science and Technology Support Program of China (NO. 2009BAG17B01), 2013

[5] Brian Levine, Alex Lu,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, Eric B. Beaton, Rider perception of a âĂIJlightâĂİ Bus Rapid Transit system - The New York City Select Bus Service, Transport Policy 49 (2016) 41âĂŞ55, Apr. 2016.

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

[8] Jeremy S. Safran, Eric B. Beaton, 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, 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, Satya Muthuswamy, A SIMPLE METHODOLOGY TO ESTIMATE QUEUE LENGTHS AT SIGNALIZED INTERSECTIONS USING DETECTOR DATA, TRB 2013 Annual Meeting, 2013