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

0.3.4

0.4 Results and implications for practice

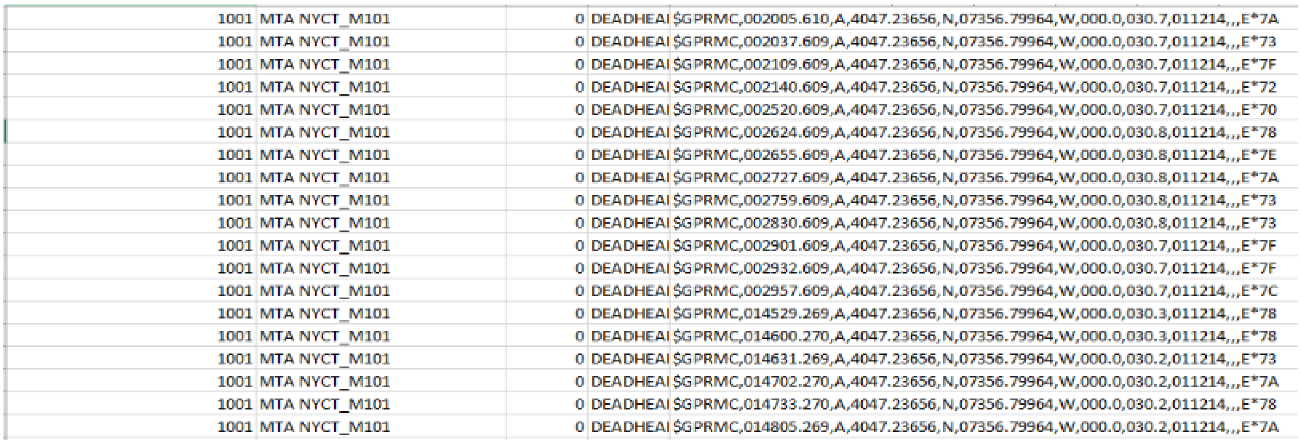
Table 1: Differences between DOT and Bus Time data.

	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

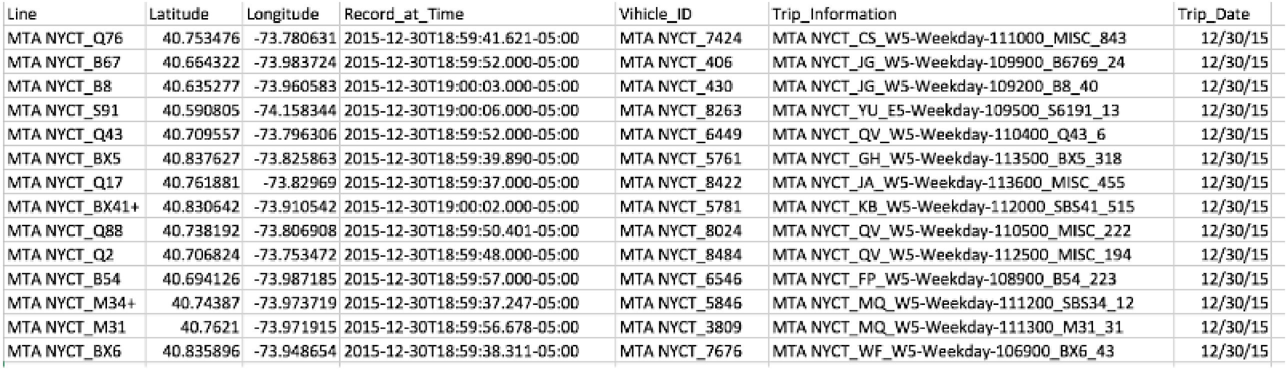
0.4.2 Discussion about the theory

Figure 1: Original sample extraction from DOT.



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

Figure 2: 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

0.4.3 Samples of the deliverables

0.4.4 Risks and advantages of the approach

0.4.5 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. Aimed at helping DOT improve the bus performance efficiently, the metrics focus on not only bus operation element, but also on other traffic aspects. Taking one-year data as our object of study, the data is extremely huge so that big data technology like Spark is widely used in this project. In order to make our work easy to share and reproduce, we choose to use SQL API to manipulate the data, which enable DOT to easily make change 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) Pros: 1. Focus on the viewpoint of DOT, the correlation elements we chose not only focus on the bus operation, but also other traffic elements. 2. With such large dataset and big data technique, the result of our analysis is more persuasive compared to traditional sample study. 3. Comparison among models(different models to estimate time as well as to measure performance) Cons: 1. The choice of elements based on formal studies, the objective of which are not all NYC. 2. Too much assumptions. 3. The departure and arrival time for each bus at each stop comes from indirect approximate methods, it is better for MTA to collect the accurate data to help the model works better. InfluenceëïijŽ 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.

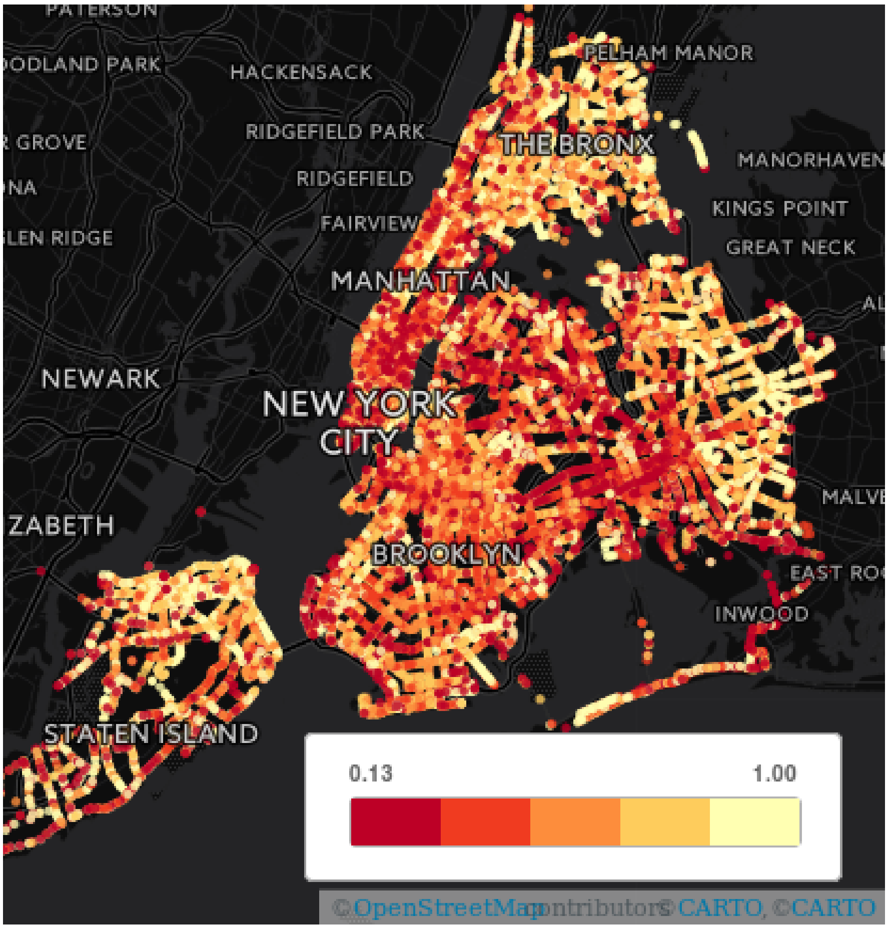
Future work

- 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 3: 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