TRB Paper Manuscript #15-0309

Develop New York City Surface Transit Boarding and Alighting Ridership Daily Production Application Using Big Data

Qifeng Zeng, Alla Reddy*, Alex Lu, and Brian Levine
** Corresponding author*

Qifeng Zeng
Computer Specialist
New York City Transit Authority
2 Broadway, Cubicle A17.092, New York, N.Y. 10004-2208
Tel: (646) 252-2400
Email: *Qifeng.Zeng2@nyct.com*

Alla Reddy
Senior Director, System Data & Research (SDR), Operations Planning
New York City Transit Authority
2 Broadway, Office A17.92, New York, N.Y. 10004-2208
Tel: (646) 252-5662
Email: *Alla.Reddy@nyct.com*

Alex Lu
Transit Analyst
P.O. Box 684, Ossining, N.Y. 10562-0684
Tel: (212) 340-2684
Email: *me@lexciestuff.net*

Brian Levine
Principal Transportation Planner, System Data & Research (SDR), Operations Planning
New York City Transit Authority
2 Broadway, Office A17.100, New York, N.Y. 10004-2208
Tel: (646) 252-5541
Email: *Brian.Levine@nyct.com*

Submitted for Consideration for Publication in
Transportation Research Records: Journal of the Transportation Research Board

Word Count: 274 (Abstract) + 6,273 (Text) + 4 * 250 (Figures) = 7,547 Words.
Submittal Date: November 15, 2014.

ABSTRACT

New York City Transit Authority (NYCTA) has put into production a ridership application that determines surface transit boarding and alighting locations for each one of approximately 2.8 million daily passenger trips on board 218 bus routes, to support bus service scheduling and planning. The application combines data from an Automated Vehicle Location (AVL) system, a multimodal entry-only non-geographic Automated Fare Collection (AFC) system, and General Transit Feed Specification (GTFS) schedule and shape file streams. To accomplish this, NYCTA developed a highly optimized network generation tool to estimate bus boarding/alighting locations and link loads, by creating a scaled-down custom network based on first-bus trajectory (from AVL boarding location data) and a few possible AFC/AVL-inferred second-leg pickup stops. Solving for the shortest walking path on this sub-network yields connection points and disembarkation points much more efficiently than solving for all 128 million potential O/D pairs citywide. The optimized program executes in less than three hours in an automated environment and is responsive to daily detours, special events, and weather-driven ridership, which can support operations management's needs for next-day reporting as well as the monitoring of fast-changing patterns and trends in ridership behavior. It also allows multiple days of route-level program output to be aggregated for schedule making purposes, providing a significantly more representative understanding of typical passenger loads that were historically obtained from a few labor-intensive on-board observations (ridechecks) collected over a multi-year period. Results were validated and found to be consistent with manual ridechecks, limited O/D surveys, and earlier estimation methodologies. Obtained accuracy was sufficient to achieve acceptance of AVL-AFC data in lieu of traditional on-board observations for NYCTA schedule making and vehicle assignment purposes.

INTRODUCTION

This paper presents New York City Transit Authority's (NYCTA's) production algorithm to determine surface transit boarding and alighting locations for each one of approximately 2.8 million daily passenger trips on 218 bus routes by combining data from:

- an Automated Vehicle Location (AVL) system with explicitly limited scope in providing analytics.
- a legacy regional multimodal entry-only Automated Fare Collection (AFC) system, designed over twenty years ago, that covers six distinct bus operators, seven bus rapid transit lines, two subway systems, an inter-state rail transit line, a tram, and an airport shuttle.
- a General Transit Feed Specification (GTFS) schedule-and-shape file stream generated by commercial-off-the-shelf (COTS) crew and vehicle scheduling system.

The goal of this project is to supply detailed ridership reports to operational management on daily basis. The reports, based on AVL and AFC data, will be compiled at the trip and stop level, with roll-up summaries. This will provide invaluable operational and analytical tool that can be used to assess the effects of planned service changes (General Orders, special events) and unplanned disruptions (weather, traffic, etc.) and will permit the monitoring of fast changing patterns and trends in ridership behavior. AFC data (over 9 million total daily transactions by over 3 million unique cards from an estimated 40 million outstanding farecards in active circulation) and AVL data (over 4.8 million time-and-location reports from 4,800 vehicles) are available as early morning bulk download for the previous 24-hour period. In order to process AFC/AVL data by the next day, the program must execute totally unattended within a maximum running time of five hours.

To account for frequent short-term route changes to surface transit network—publicized in Authority's GTFS feeds—the algorithm uses these changes as input and assigns passenger accordingly. It must also accept and make "best guess" decisions without human intervention for common error conditions, such as GTFS data not matching actual operations; wrong vehicle head signs; corrupt or missing AFC data; and busted or non-reporting Global Positioning System (GPS) devices.

To meet these goals and requirements, NYCTA took the following approach:

- For the most complex case, two consecutive bus trips, AFC data provides boarding timestamps to the nearest six minutes, allowing AVL to geocode to the nearest 0.5 miles (about 3-5 possible stops). A highly optimized network generation tool then produces a limited custom network based on first bus trajectory (from AVL boarding location) and the few possible second-bus pick-up stops; solving for shortest walking path on this small network thence yields transfer point and consequently disembarkation point from first bus.
- For the trips from bus to subway , determining the bus disembarkation stop requires solving for the shortest walk between route-specific possible alighting stops and actual subway entrance logged by AFC, which provides definitive geographic locations. Similarly, a small network is generated then solved for each transfer.
- For the trips from subway to bus, a daily lookup table is generated by finding the closest subway station to each bus stop within given distance. Among 3-5 possible bus boarding stops implied by AVL data, the one closest to subway is selected. An earlier methodology that solves schedule-based shortest path to determine exit stations was scrapped due to excessive computing time.

Relationship to Prior Work

Development of AFC origin-destination (O/D) modelling capability in New York City dates back to late 20th Century when Barry et al. *(2)* published an inference method assigning the destination of one trip based on next-trip origin. This approach had been known at NYCTA since 1997 when AFC system first began reporting card serial number data. Variation of this approach plus various extensions utilizing schedule data *(5)* was subsequently popularized and applied in Chicago *(3)*, Boston *(4)*, Beijing *(5)*, London *(6)*, Hong Kong *(7)* and no doubt other rail transit systems.

A decade later, Freimer et al. *(8)* reported that NYCTA had extended this approach into multimodal region-wide O/D models,—with bus boarding and disembarkation localization provided by schedule-based dead-reckoning creating a linked transit trip database from over 7 million daily AFC swipes. Combining New York City's famous traffic gridlock and lack of functioning AVL, bus O/D estimations and consequent vehicle-level passenger load determinations were insufficiently accurate for scheduling frequency determination and vehicle assignment. Few bus routes were validated against empirical data, and datasets required manual processing and "massaging" before becoming usable for O/D estimations. Only a few days' data from 2004 was processed, and model quickly became outdated. Stop localization was so problematic that NYCTA elected to exploit travel pattern symmetries in subsequent work *(20)* instead of guessing where bus boarding occurred using AFC data alone.

Meanwhile, as smartcard-based AFC, AVL, and even Automated Passenger Counters (APC) became popular elsewhere, algorithms for processing and combining AFC-AVL-APC data streams and generating O/D matrices thrived in both proprietary *(30)* and public space *(40)*, with research being carried out in locales including: Minneapolis *(9,1)*; Columbus, Ohio *(16)*; Chicago *(39)*; Montreal *(12)*; London *(11,12a)*; Taipei, Taiwan *(10)*; Jinan, China *(41)*; Shenzhen, China *(17)*; Santiago *(14)*; and São Paulo *(15)*. A recent TCRP study reviewed and catalogued many common analytical models utilized within the industry to process archived data *(19)*.

Despite the popularity of AFC-AVL-APC methodological research, few references exist in the literature for automatically generating surface transfer points. Relevant research focuses on grouping bus stops together using geographic proximity, text similarity, and land-use pattern to automatically provide transfer information *(1)*. Like many studies, databases were used for implementation, and the need for grouping highlights the challenges of creating and maintaining enormous lookup tables.

While O/D analysis of passively collected transit data is a crowded field, our work extends an operations-support, production-oriented philosophy of zero manual intervention (ZMI) generation of next-day "flash" reports—first accomplished in New York with AFC-based passenger-miles computation *(20)*, followed by track-circuit based performance monitoring *(21)*—to much more complex tasks of calibrating and generating full O/D matrices. Results were convincing and comparable to field observations at sufficiently high granularity and fidelity at individual vehicle load, passenger origins and destinations, peak-load points, and time-period levels—to be accepted by NYCTA scheduling dept. for use over labor-intensive onboard on/off counts (colloquially, 'ridecheck'). While newer or smaller transit properties could run SQL queries from composite smartcard-APC-GPS data, NYCTA accomplished this with legacy entrance-only AFC, simple low-cost AVL, and varying bus routes.

Summary of Innovations

This algorithm takes an unique approach and makes a major contribution as it amounts to an automatic, error-tolerant, and analyst-free daily re-calibration of a classic shortest-path passenger routing model utilizing daily GTFS file as base network and combined AFC-AVL data as constraint to model each individual farecard itinerary for every trip every day. It also demonstrates problem-specific optimization. Key innovations are:

- Large scale production implementation
- Zero manual intervention
- Each actual bus-bus and bus-subway consecutive passenger trips observed in AFC data (about 2.5 million daily transfers, 1.0 million unique customers) is solved discretely, rather than solving transfers between all 218 routes at over 16,000 stops system-wide.
- Dynamic transfer links between two bus routes are generated only from feasible disembarkation points for that passenger on that first bus to the few known possible second bus boarding locations (and not all bus routes in the vicinity).
- Tight on-the-fly integration of previously unconnected AVL, AFC, and GTFS data sources.

BIG DATA

Input data is taken from three major sources: MetroCard (NYCTA farecard) transaction data, MTA BusTime AVL data, and Authority GTFS feed. MetroCard's AFC data format was well-documented in prior Authority papers *(2,8,20,22,23)*. This discussion will focus on two new sources of data used by this algorithm.

Daily Schedule GTFS Feed

GTFS format, introduced in 2007, contains a wealth of information describing transit schedules, and routes in Geographic Information Systems (GIS) format. The Authority started publishing bus and subway GTFS data generated by industry-standard scheduling software on its website for third party developers in January 2010 *(31)*. Bus scheduling processes are described in detail by Holmes *(24)*. MTA GTFS feeds contain sufficient detail for systematically and accurately constructing traditional computer transit networks consisting of nodes and links elements and their properties such as running time and wait time.

Most transit systems publish system-wide schedule changes several times a year (colloquially, a 'Pick'). However, NYCTA routes sometimes undergo major route/schedule changes between Picks, which are very difficult for developers to keep track on daily basis. MTA GTFS files are constantly updated to include between Pick reroutes—providing impetus for building a robust application capturing everyday operational activities.

MTA BusTime AVL Data

Under MTA's BusTime project, NYCTA installed experimental AVL devices on B63 route in Brooklyn beginning April 2011. Almost all of over 4,800 buses running >300 NYCTA and MTA Bus local and express routes were equipped with AVL devices by March 2014. BusTime uses GPS hardware and wireless communications to track bus locations in real-time.

Due to legacy hardware and proprietary software issues onboard NYCTA buses, AVL computers actually "sniff" serial communication links between fareboxes and headsign units to determine basic trip

information like operator pass number (employee ID), route code, direction, and destination. GPS location is reported to central servers approximately every 30 seconds. Server software integrates this information with map, route, schedule data, previous real-time updates, headsign information, and real-time operational data like operators' timekeeping system—linking vehicle fleet numbers with a particular trip within scheduled runs. Sophisticated inferential algorithms then provide a "best guess" in real-time of (a) whether bus is in-or-out of service, (b) what route variant—'path' in TA parlance—is being served, (c) route direction, and (d) if off-route—i.e. detour in effect. Based on inferred information it predicts what stops bus will make and distance from each stop. BusTime server software utilizes the open source OneBusAway package *(27)*; output data is publicly accessible to developers as a feed *(28)*.

However, AVL archive data requires significant post-processing and scrubbing to make usable. Because BusTime 'path' is a real-time "best guess", it can be inconsistent from stop to stop. Only after trip is complete can path be conclusively determined— path is not traced out until bus reaches its terminus. Another issue is dealing with buses with head signs set to "Not In Service" or "Next Bus Please" but yet remain in service. Also, about 10% of GPS devices are not functioning on an average basis. This process and associated AVL issues are described in detail in Levine et al. *(25)* and Glikin et al. *(26)*.

METHODOLOGY

Dynamic Transfer Link Network

Most bus origin-destination estimation studies *(12a, 38,39)* use pre-generated bus-bus transfer lookup-table to infer alighting bus stop(s) or area. Transfer lookup-tables consist of permissible transfers systemwide. Each study delicately defines policies and crafts methodologies to create lookup tables. It's the most important factor influencing quality of results.

Bus-bus lookup approaches are not appropriate for New York City surface transit daily production application. The method generally involves great deal of human judgment and tedious time-consuming manual operation to generate transfer tables. Even if systematic approaches is found for table generation, plenty of resulting exceptions still requires manual editing. Therefore, it is most suitable for research studies where dimensions and complexities of transport networks can be contained. New York City transit network has over 16,000 bus stops and 426 subway stations with various transfer exceptions that can change frequently, making it impossible, due to size and complexities, to create and maintain a production oriented lookup table.

Fundamental network computing method are explored for this application. Shortest path methods are used for deriving bus alighting locations. Network computing methods usually prepare static transit networks for shortest-path *(35a)*. Surface transit networks include bus routes links, and walking links connecting bus routes via bus stops. Walking links are conceptually same as transfer tables except for each link having different impedances. Underlying structure of walking links is street network; the street network, in theory, can connect any two bus stops. Even with moderate surface networks and strict rules in governing walking link creation, size of walking links permuted by bus stops could easily grow out of control.

In addition, classic Dijkstra's shortest-path algorithm has order of growth proportional to vertices squared, i.e. $O(|V|^2)$ *(29)*. The entire New York City surface network couldn't be solved in reasonable

time (Figure 1(a)) with 128 million unique potential O/D pairs on 24 hours bus network with classical method of transit network generation and shortest path.

A highly optimized network generation tool is developed, restricting shortest-path problem sizes when solving for a specific farecard's exact trajectory given fuzzy information about boarding locations. A novel approach, called "Dynamic Transfer Link" method, is developed for this study. Instead of creating walking links statically, the method takes advantage of known next bus trip's fuzzy boarding locations (due to six-minute farebox timing error margin) or subway stations, and creates walking links dynamically during shortest path computing process. Only possible walking links between two bus routes or bus route and subway system meeting predefined criteria (usually "distance") are generated. Number of walking links created dynamically during shortest path process is dramatically reduced to a fraction of static network size. Onerous transfer exceptions constraints are automatically and implicitly considered by exploring only paths traced by an actual farecard. It takes merely 1½ hours to compute over 2 million daily bus passengers using "Dynamic Transfer Link" method with 1.9 miles as maximum walking distance allowed.

The illustrated example of "Dynamic Transfer Link" is shown in Figure 1(b). Four consecutive MetroCard transactions for a subway-bus-bus-subway itinerary (Station 0, Route 1, Route 2, and Station 4):

- From Station 0 to Boarding Stop 2, a transfer link is created by lookups based on the geographically closest subway station to bus stops on Route 1.
- From Routes 1 to 2, dynamic transfer links are created by linking all bus stops downstream of Boarding Stop 2 to all possible Boarding Stops 3, so long as transfer links are less than 1.9 miles—all possible walking transfers.
- From Route 2 to Station 4, stops downstream of Stop 3 are linked with subway stations to create the bus-subway transfers.

With dynamic transfer links, a partial transit network specific to this farecard itinerary is completed and fed into Dijkstra's algorithm for path finding using shortest path assumption. Output results include boarding and alighting stops on each individual bus trip.

Implementation Flow Chart

This application takes five general major steps (Figure 1(c)):

1. Generate subway-bus lookup table and bus route network from GTFS.
2. Determine the arrival time at all stops for all trips from BusTime data.
3. Attach each AFC transaction to bus stops groups.
4. Build and solve custom shortest-path network for each bus-bus AFC transaction sequence. (Individual passenger travel path is the basic output)
5. Results are factored to account for O/D indeterminate AFC transactions, cash passengers, and fare evaders.
6. For longer-period summaries, data from multiple days are blended.

Numerous proof-of-concept methods to generate O/D matrices have been tested both by NYCTA and outside research groups *(16)* using familiar synthetic methodologies such as iterative proportional fitting (IPF) that can generate a full O/D matrix from marginal on/offs obtained from APC or AFC, but to date no other groups have reported an automated data merge of entry-only AFC and AVL data by treating each

farecard discretely, thereby preserving as much of the actual O/D information as possible without analyst manual intervention and model calibration. Complete O/D matrix is generated from all 2.8 million daily swipes.

Step 1: Construct Network

Computer transit networks have links (bus, subway, walk for transfer, O/D access and egress) and nodes (subway stations, bus stops, transfer points, etc.). Transit routes are usually static and can be automatically generated from GTFS. However, coding of transfer, access, and egress links are more challenging. Transfers between subway stations are usually straightforward, as physical connections exist between stations without incurring additional fare. Bus-bus transfers occur on the street, which presents numerous ways of connecting. Developers usually have to decide which transfers are practical feasible links. Generally, certain criteria are set for feasible connections, normally based on threshold distance, and links are generated systematically. With this approach, no transfer links are created at this stage.

Subway to bus transfer matrices are built from GTFS for that specific day based on geographic proximity. GIS analysis was conducted using the shortest distance between geocodes of subway stations and bus stops. Separate pre-processing generates a system wide bus network that describes nodes and links from GTFS schedule data and the distance between nodes using GTFS bus shape files and GIS linear referencing feature.

Step 2: Deriving Bus Stop Arrival Time

GIS-based bus route tracking was developed to identify and validate route variant for each individual bus trip. Each trip is compared to system wide bus network and GTFS route shape. After correct path is assigned, entire bus trip is re-processed to establish arrival times at each stop. If no valid path was found, the trip is removed.

Step 3: Merging AFC and AVL Data

Prior modelling efforts *(8)* were done without AVL data and had tremendous challenges in determining bus location. AFC data provides the farebox number when passengers board the bus, but not exact geographical location due to the constant movement of the bus. The AFC transaction time is linked to bus stop arrival time, developed in Step 2 to determine boarding stops. Since MetroCard is entry-only, only boarding points can be determined at this stage. The limitation of bus farebox memory when MetroCard was introduced constrained timestamp resolution to one-tenth of an hour (six minutes), during which bus might have covered 3-5 stops. Thus, a number of possible boarding stops are selected.

Step 4: Generating Dynamic Bus Transfer Links and Running Shortest Path

With numerous known boarding points for individual journeys, this step applies the “Dynamic Transfer Links” and Dijkstra’s shortest-path algorithm method developed for the application to determine bus boarding and alighting points as well as walking links, and consequently obtaining the paths of the journey. As heart of the application, the “Dynamic Transfer Links” method is detailed in “Dynamic Transfer Link” section appearing earlier.

Step 5: Non-Farecard Passenger Trip Estimation

Almost all the transit system has non-farecard passengers. In New York City transit, about 9% (May 2014) of bus passengers still use cash fare. Cash to AFC passenger ratio by bus route and direction are calculated and corresponding O/D patterns are used for allocation. Fare evasion ratio *(20)* at bus route level is used to further adjust passenger trips. Using the same approach, individual fare evader trips are synthesized based on AFC O/D patterns. Figure 2(a) shows NYCT borough-level adjustment factor summary.

Step 6: Blending/Trending Data

For schedule production purposes, medium-term ridership trends are more helpful than daily results. When combining data from multiple days, as bus assignments differ each day, schedule run number and scheduled trip start time are used as keys to merge multiple days' data in a database. Since AVL malfunctions tend to be randomly distributed throughout the fleet, multiple-day averages provide reliable data on all trips specifically even if some trips ran with busted GPS on some dates. Medium-term averages also smooth out effects of non-recurring noise like street congestion, special events, weather conditions, or simply non-routine customer travel.

NEW YORK CITY SURFACE TRANSIT APPLICATION

Several key assumptions were needed to customize general methodology described to New York City surface transit daily production application. Most are typical reassurances of observations and policies widespread in AFC analysis practice. Length of maximum transfer is a key parameter affecting results and outcome. Sensitivity analysis was conducted to determine appropriate transfer length in New York City's environment.

One-to-One Passenger-Farecard Relationship

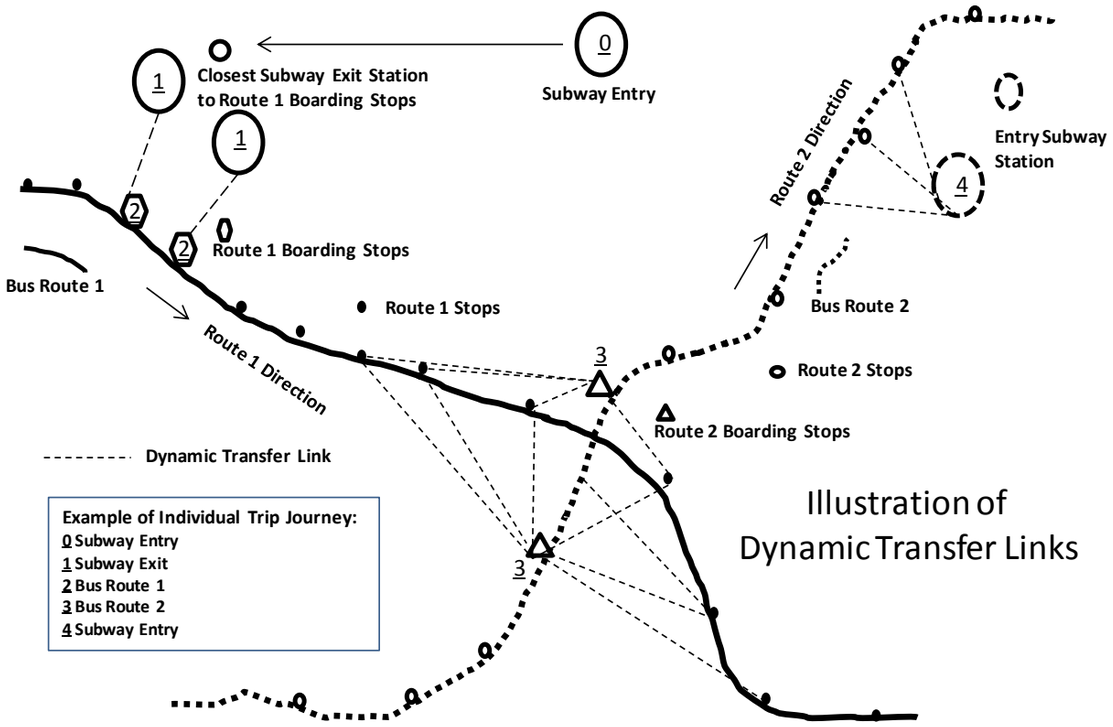
Individual passengers are identified by their MetroCard, assuming that a passenger uses one card throughout the day. However, this assumption is incorrect in some scenarios. Most common and troublesome scenario is when passengers purchase and begin to use new MetroCards midday. Connection between old and new MetroCards then cannot be established. In March 2013, Authority Tariff was modified to levy $1.00 surcharge for each new farecard issued—reducing of plastic waste, decreasing required MetroCard Vending Machine (MVM) servicing frequencies (to re-fill farecards), and reducing trash in stations. This fare policy was effective in encouraging customers to add value or time to existing MetroCards and thus significantly improving passenger-farecard assumption's validity. By May 2014, 11% of total in-system sales transactions were new cards, compared to 39% before March 2013.

1

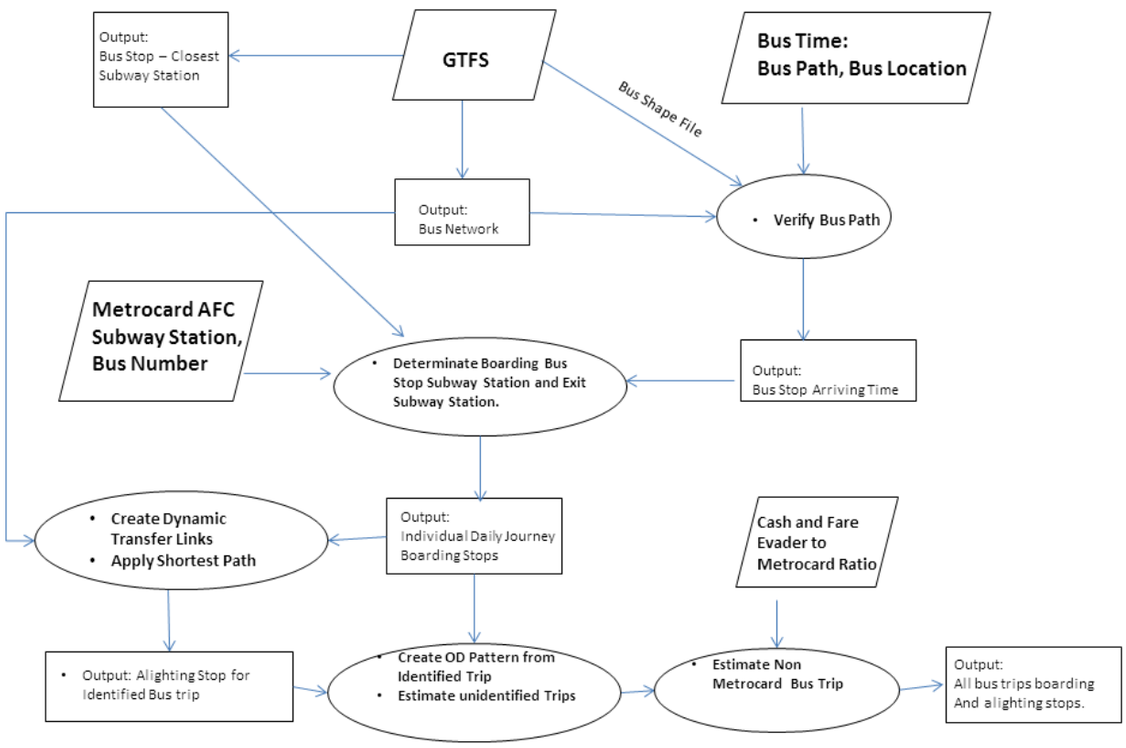
(a)

O/Ds solved	**All O/Ds systemwide**	**AFC observed O/Ds only**	**AFC observed O/Ds only**	**AFC observed O/Ds only**	**AFC observed O/Ds only**
Network Generation Approach	**Classical lookup**	**Route-based stop masking**	**Distance-based stop masking**	**AVL-based origin masking**	**Directional dest. masking**
Potential Origins	~16,000	~90	~32	3-5	3-5
Potential Destinations	~16,000	~90	~32	~90	~23
O/D permutations (vertices) solved	~16,000² /2 = ~128 million	~90² /2 = ~4,050	~32² /2 = ~512	4*90 /2 = ~180	4*23 /2 = ~46
Number of sub-problems solved	1 (to populate lookup table)	<2.8 millon (worst case)	<2.8 millon (worst case)	<2.8 millon (worst case)	<2.8 millon (worst case)
Total operations	$(128m)^2$ = $16{,}384 * 10^{12}$	$4{,}050^2 * 2.8m = 45.9 * 10^{12}$	$512^2 * 2.8m$= 734 bn	$180^2 * 2.8m$= 90.7 bn	$46^2 * 2.8m$= 5.9 bn
Est. execution time	317 years	324 days	5.2 days	15 hours	1 hour

(b)



(c)



2 **Figure 1** Algorithm Methodology Detail: (a) Theoretical Permutation Estimation for Various Algorithmic
3 Approaches; (b) Illustrative Example of Dynamic Transfer Link Generation; (c) Final Algorithm Processing
4 Flowchart.

Passenger May Travel Maximum of 1.9 Miles While Connecting

To determine connection validity between consecutive waypoints, maximum thresholds for time/distance are required. In New York City, people are more likely to use longer "walk link" for mid-distance travel than other cities. Besides walking, customers have rich selections of transportation alternatives including bicycle, taxi, skateboarding, black car (gypsy cab), third-party bus shuttle, ethnic jitney, horse-drawn carriage, pedicab, commuter rail, and ferry. With high density commercial and residential development throughout the City, higher probabilities of having one or more activities in between consecutive trips (i.e. trip-chaining) occurs. With timespans between consecutive system entries varying from 6 minutes to 24 hours, connections between consecutive trips (i.e. between two waypoints), may not necessarily be short walks made solely for transfer purposes, as is typical for other cities.

Rather than setting thresholds from theoretical walking distance tolerance, sensitivity analysis of farecard trajectory observations was carried out to determine threshold from percentage of passenger trips identifiable. If distance between inferred alighting and known boarding stops is greater than maximum transfer length, it is assumed that an error condition has occurred and passenger's alighting stop must be determined without reference to next trip.

Figure 2(b) shows percent of total passenger trips identified as function of maximum transfer threshold from ¼ to 3 miles for selected May 2014 weekdays. At one-quarter mile, 70% of trips are identified. The percentage increases rapidly to 79% as transfer threshold increases to 10,000ft (1.9 miles). However, from 1.9 to 3 miles, additional trips identified rapidly approach point of diminishing return. Thus, 1.9 miles, in the upper ranges of reasonable walking distances in New York, is selected as transfer threshold.

Every MetroCard Returns to the Same Location

This assumption is necessary when using 24-hour blocks of MetroCard data. It's assumed last trip of the day terminates where first trip originated *(2)*. Given nighttime activity dynamics, Barry argued 3am is when days start in New York City. In this study, we used midnight as start-of-day for data processing convenience with most daily input files. The 3am schedule was originally proposed for subway-only trips, where 1.3% of daily ridership occurs between midnight and 3am—much higher than equivalent 0.6% figure for buses. Furthermore, a recent study *(37)* shows that subway O-D estimating improve only ~1% if starting time moves from midnight to 3am in New York City.

Dealing with Non-Contiguous Metrocard Itineraries

Consecutive recorded locations of system entries ('swipes') do not make logical sense in some situations. Quite common in AFC data are consecutive swipes on two bus routes miles apart—perhaps in different boroughs. Such transfers are either rare or geographically impossible. Possible explanations for this occurrence are:

1. Most frequently, passenger used another transport mode between two consecutive swipes.
2. Passenger may have used different MetroCards mid-journey. When passengers change cards, inferred destinations of both old and new cards are incorrect—resulting in irrational connections between swipes.
3. More than one Pay-per-Ride passengers can share one card, resulting in non-sequential system entry transactions and causing non-contiguous itineraries.

4. Loss of transaction data occurs due to farebox failures; sequence of swipes for MetroCards having lost transactions is affected.
5. AVL/GPS devices may not function as expected, resulting in bus trips failing to provide passenger boarding locations. When geographic fixes are not available, travel paths for customers having unknown boarding locations cannot be built.

In general, when connection between two consecutive swipes is irrational, passenger is treated as having paid cash or evaded the fare. If boarding stop can be determined by AVL, boarding stop is assigned; otherwise both boarding and alighting stops are factored from O-D patterns generated from other identifiable trips.

MetroCard and Non-AFC Passengers Have Similar Trip Patterns

Since inception in 1997, MetroCard market share has steadily increased. MetroCard accounted for over 95% of NYCTA subway/bus usage in May 2014. Average weekday bus ridership in May (excluding evaders) is 2.3 million, yet only 8.5% (190,000 passengers) paid cash or used paper transfers. Non-MetroCard boardings like cash, single ride tickets (essentially tokens), senior return tickets and others *(32)* do not generate transaction information. While non-MetroCard passengers might have different O/D patterns, since MetroCard accounts for majority of travel (and generated O/D matches land-use patterns well), it's reasonable to assume differences are *deminimis*. Non-MetroCard passengers are thus assumed to have similar trip pattern as AFC passengers.

Special AFC Processing

NYCTA is a complicated system; one general modelling methodology could not provide the entire solution. Also, due to differing implementations of MetroCard AFC on transport carriers or modes post-dating early 1990s AFC design, some exception processing is required:

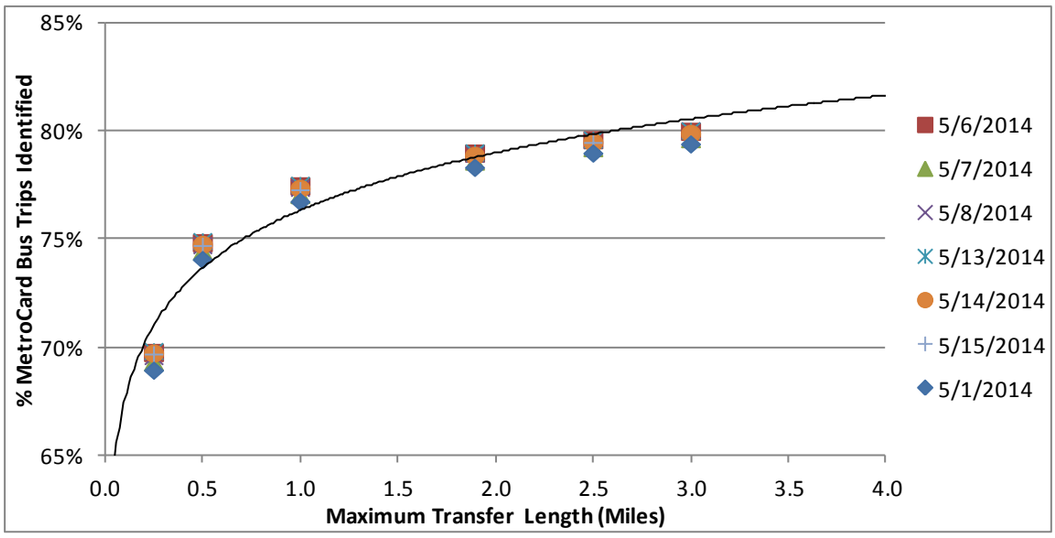
Bus Rapid Transit—Since 2008, seven new rapid bus routes (**+select**bus service, SBS) were implemented in New York. Except for S79 on Hylan Boulevard, Staten Island *(33)*, all remaining six routes including Bx12 *(36)*, M15 *(35)*, M34, B44, Bx41 *(34)*, and M60 utilize roadside proof-of-payment fare collection. Passengers use MetroCard to register at wayside MetroCard Fare Collectors (MFCs) and receive receipts retained for random on-board inspections. SBS physical buses no longer have passenger information, and thus special AFC data processing is required. Assuming SBS passenger always boards next arriving bus, onboard passenger loads can be determined. Since roadside farebox locations are static, specific boarding stops are determined. Dynamic transfer links are created similar to the regular bus transfers.

Connections with Regional Transit Operators—Numerous non-NYCTA transit operators in the region also accept MetroCard as form of payment. AFC data gathered using their terminals also require special processing, detailed in Figure 2(c).

(a)

Category	Manhattan	Bronx	Brooklyn	Queens	Staten Is.	Systemwide
AFC Passengers	85.85%	77.38%	82.35%	87.42%	75.37%	82.26%
Non-AFC Passengers						
Cash	3.82%	5.86%	5.78%	5.37%	7.61%	5.40%
Partial Fare	1.90%	3.32%	2.30%	1.92%	4.01%	2.50%
Front Door	2.12%	5.22%	3.17%	1.76%	6.67%	3.42%
Rear Door	0.32%	1.46%	0.44%	0.14%	0.29%	0.63%
Child Over 44"	1.61%	2.27%	1.81%	0.87%	1.87%	1.75%
Child Under 44"	3.06%	3.23%	2.71%	1.32%	2.38%	2.70%
Flash Pass/Wheelchair	1.01%	0.98%	1.15%	0.89%	1.50%	1.05%
Broken Farebox	0.30%	0.30%	0.30%	0.30%	0.30%	0.30%
Total	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**

(b)



(c)

Operator	Required Special Processing
Staten Island Railway (SIR)	Staten Island Railway (SIR) is a separate rapid transit line that doesn't connect to NYC subways. It has 22 stations, two of which (St. George and Tompkinsville, near ferry terminal) have fare control facilities. Fares are charged for both entries and exits, similar to downtown terminal ticket barriers in typical commuter rail practice. SIR entry AFC transaction is no different from subway entries. Since most boardings from suburban stations have destinations of either St. George or Tompkinsville, exit transactions are considered as boardings at any other non-fare control stations. For exit transactions, dynamic transfer links from buses include all possible transfers from any downstream bus stops within 1.9 miles of suburban SIR stations.
Staten Island Ferry (SIF)	Staten Island Ferry is a free shuttle operated by NYC Dept. of Transportation between South Ferry, Manhattan and St. George, Staten Island (SI). Named after such famous SI politicians as Marchi and Molinari, SIF serves mainly Staten Island residents' commuting needs to Downtown Manhattan. SIF route is five miles long. Many SI residents make daily intermodal trips of subway-ferry-bus or bus-ferry-SIR daily. Although physical distance between South Ferry and St George is >1.9 miles, transfers via ferry are definitely legitimate. Therefore, a pseudo St. George bus stop is created with South Ferry geocodes. When dynamic transfer links are created from St. George to Downtown transit services, pseudo locations are used, representing SIF.
Port Authority Trans-Hudson (PATH) Train	PATH is a rail transit service between New York and New Jersey, with six stations in Manhattan. MetroCard can be used to pay PATH fares. Some individuals' itineraries include travel on PATH. To model their activities, the six PATH stations in Manhattan are replaced with closest TA subway stations. All N.J. PATH stations are replaced with either World Trade Center or Herald Square subway stations, which are likely Manhattan destinations of PATH passengers.
Nassau Inter County Express (NICE) Bus	NICE (Nassau Inter County Express) bus mainly serves Nassau County, to the east of Queens. NICE adopted MetroCard as primary fare media and have service closely coordinated with TA facilities and routes. Over 8,000 transfers were made between NICE and TA/MTA buses in October 2013. NICE bus makes GTFS and GIS data available, but does not provide similar AVL data. Based on NICE route provided in AFC data, we consider all stops on that route possible boarding stops. When transfers are made between TA/MTA and NICE bus, dynamic transfer links are generated between TA/MTA route's all possible alighting stops within 1.9 miles of NICE route's every possible entry stop.
Bee-Line (Westchester County Bus System)	Bee-Line is Westchester County bus, to the north of Bronx, with over 7,000 monthly transfers to/from TA/MTA buses. Bee-Line bus does not provide GTFS data; we simply use Bx16, which operates along the Bronx-Westchester county line, as approximate geographic representation of all Bee-line routes.

Figure 2 Algorithm Application Detail based on New York City empirical data: (a) Non-AFC Passenger Adjustment Factors; (b) Maximum Transfer Length Sensitivity Analysis; (c) Special AFC Processing due to regional integration of farecard systems.

RESULTS VALIDATION

Three different approaches were used to rigorously validate model outputs:

1. Route-level average trip length
2. Bus passenger load at each stop:
    a. versus classic manual ridecheck, time-period level average
    b. versus experimental APC data, trip level
3. Aggregate origin-destination matrix

Unlinked Trip Length Validation

Validation process starts with unlinked passenger trip length. This is a route characteristic that depends on the land-use and activity patterns at the stops—provided routing does not change, average passenger lengths-of-haul by route should remain relatively constant. Conventionally, this is considered a key model quality measurement. Since model output is individual bus passenger boarding and alighting stops, lengths-of-haul is easily determined.

NYCTA had developed an AFC-based prior model *(20)* that directly measures unlinked trip length but makes a constant-speed bus operations assumption. That model was accepted by Federal Transit Administration (FTA) for official passenger-miles computation. Figure 3(a) and 3(b) show comparisons between AFC-only and improved AFC-AVL models for local and express bus routes derived from March 6, 2013 data.

Average trip length difference for local routes is <2.3%, with errors appearing randomly distributed and approximately equal number of overestimated routes compared to underestimated ones, largely validating new model for local bus. However, validation results show that AFC-AVL model estimates express bus trip lengths on average 34% higher than AFC-only model. To understand large differences in express bus results, recall that:

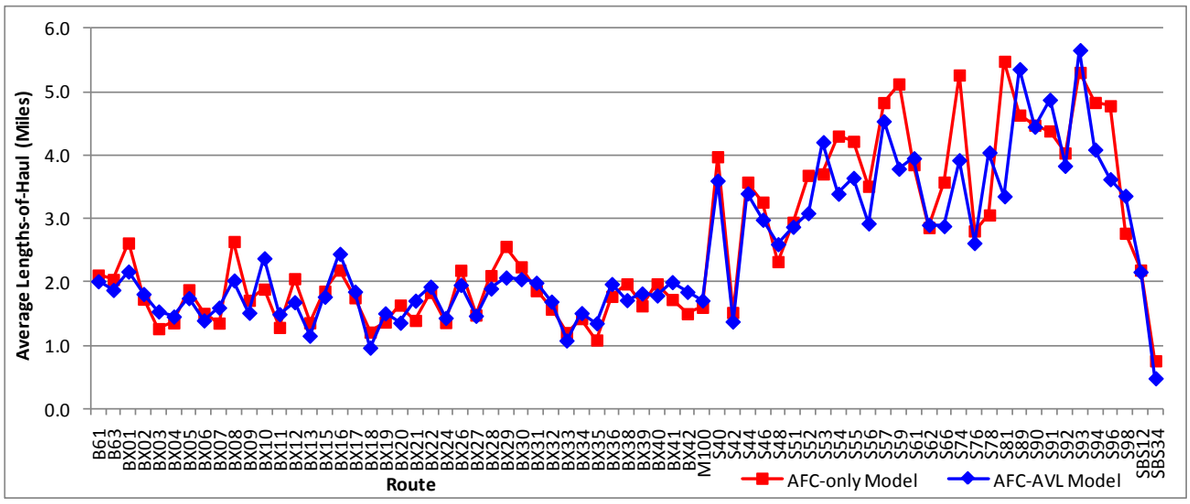
> Constant average speed assumptions implicit in distance-time conversions are a little more problematic, as NYCTA has routes where average speeds vary considerably over the whole route. Express Bus average speeds might be 15 mph at the residential end, 30~45 mph in non-stop express portions, and <10 mph in downtown—all within one trip. [...] Using average speeds leads to biases in these cases. *(20)*

Figure 3(c) (from *(20)*) clearly shows the authors anticipated that for inbound express buses, passenger-mileage underestimation would occur for all boardings. As no ridechecks are performed on express buses, that assumption could not be validated then. The new AFC-AVL model supports that assertion.

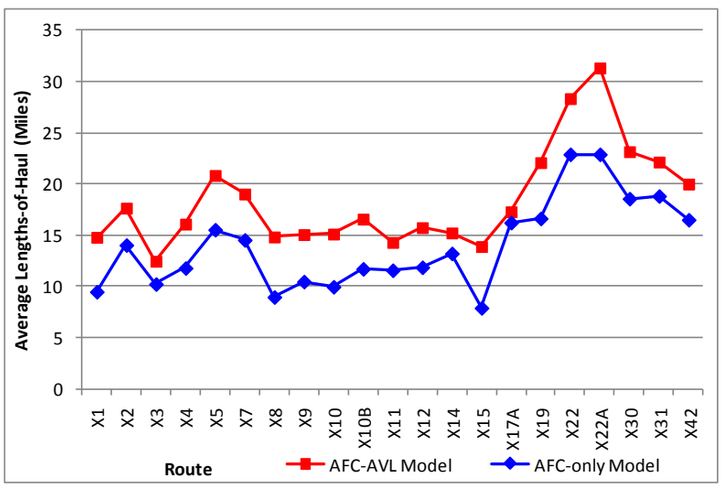
Ridecheck Validation—Ridership Profile

Main validation effort is focused on comparison against ridecheck data. One goal of this study is to provide an all-electronic alternative for bus scheduling ridecheck. Comparing it with ridecheck, hithertofore considered standard, would essentially determine this model's practical worthiness.

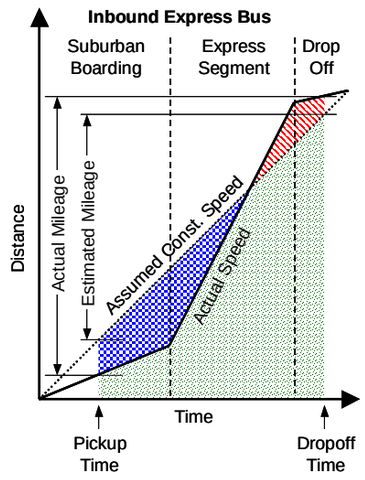
(a)


Average Lengths-of-Haul (Miles)
Route
AFC-only Model
AFC-AVL Model

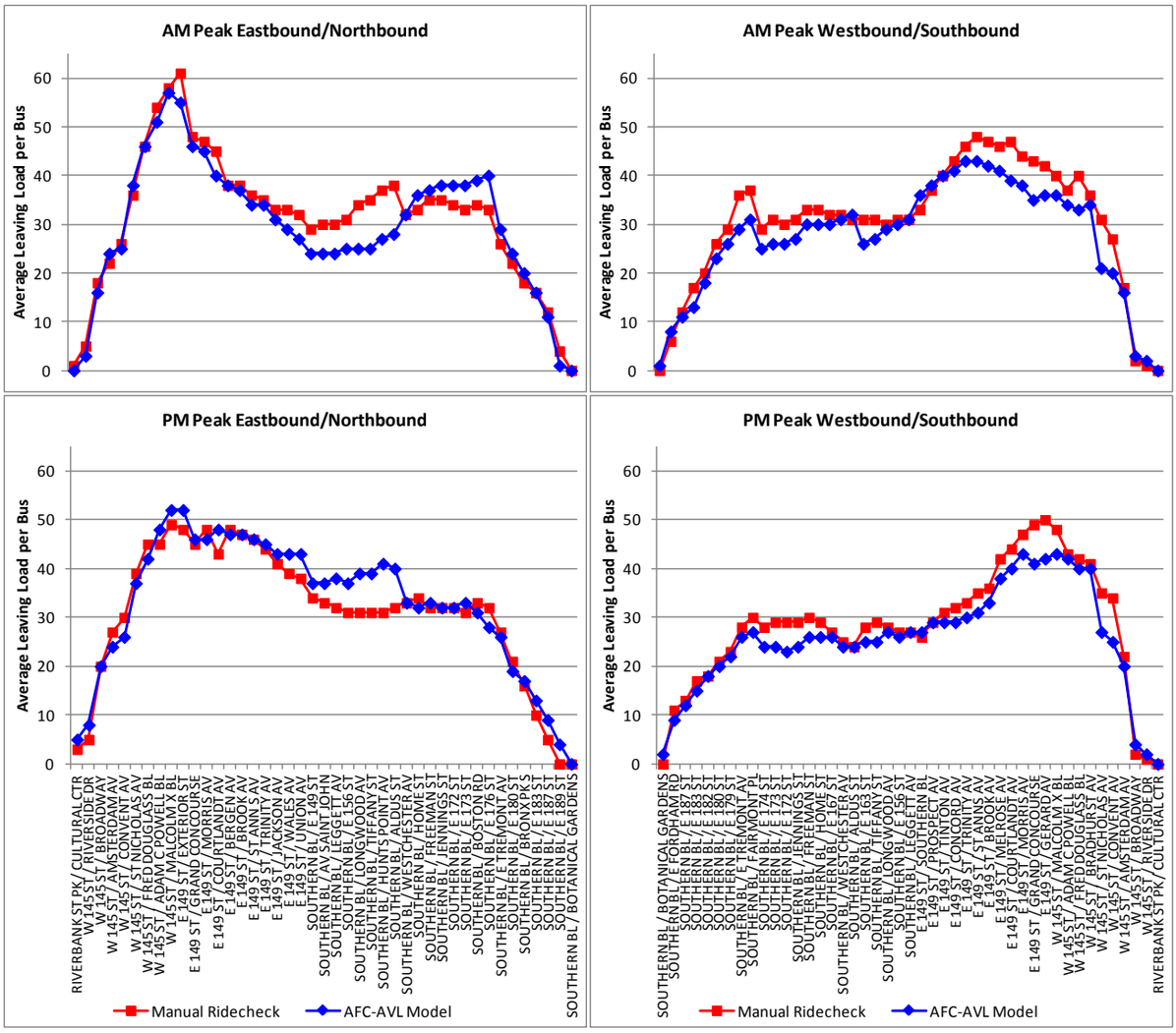
(b)

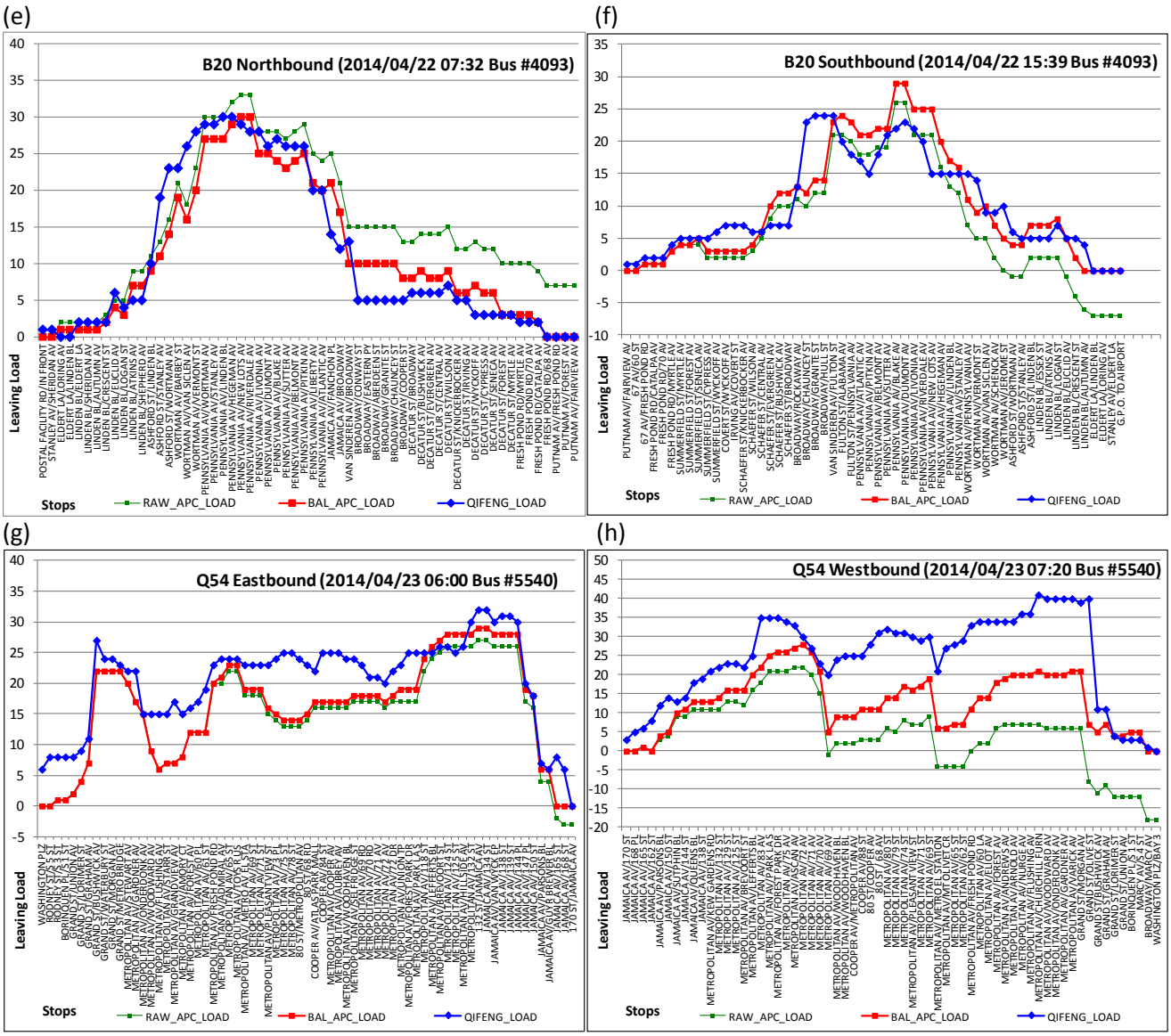

Average Lengths-of-Haul (Miles)
Route
AFC-AVL Model
AFC-only Model

(c)


Inbound Express Bus
Suburban Boarding
Express Segment
Drop Off
Distance
Actual Mileage
Estimated Mileage
Assumed Const. Speed
Actual Speed
Time
Pickup Time
Dropoff Time

(d)


AM Peak Eastbound/Northbound
AM Peak Westbound/Southbound
PM Peak Eastbound/Northbound
PM Peak Westbound/Southbound
Average Leaving Load per Bus
Manual Ridecheck
AFC-AVL Model



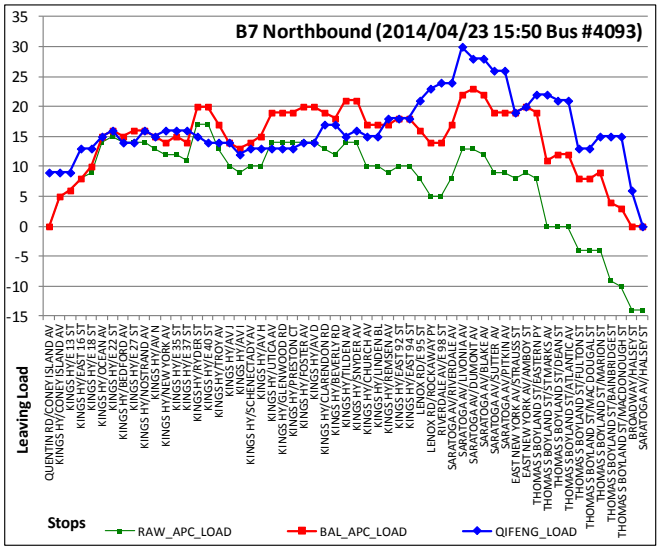


Figure 3 Basic Algorithm Validations: (a) Average Length-of-Haul Comparison—Local Bus; (b) Average Length-of-Haul Comparison—Express Bus; (c) Length-of-Haul Underestimation in Earlier Algorithm *(30)*; (d) Ridecheck Link-Load Comparisons on Bx19; (e) APC-AFC Link-Load Comparison on B20 Northbound ; (f) APC-AFC Link-Load Comparison on B20 Southbound; (g) APC-AFC Comparison on Q54 Eastbound; (h) APC-AFC Comparison on Q54 Westbound; (j) APC-AFC Comparison on B7 Northbound.

Leaving load from each bus stop, averaged over time periods (i.e. AM Peak, Midday, etc.) is used for quality control comparison. Very few true parallel tests were possible because the Authority had already cut traffic checker forces in anticipation of system wide AVL deployment, even before AVL devices were reporting any data. However, Bx19 (Riverbank Park, Manhattan to Botanic Gardens, Bronx via 145 St. and Southern Blvd.) had ridechecks conducted on November 19, 2012 when most buses at West Farms Depot were reporting AVL data. A same-day comparison at time period- and stop-level (Figure 3(d)) allows for easy visual examination showing excellent load profile and absolute value correspondence between manually-collected ridecheck results and automatically-generated AFC-AVL model. More importantly, peak load point moves only by one or two stops no matter which data source is used, meaning model deployment would not produce any sudden shifts in frequency determination and scheduling considerations.

Ridecheck Validation—Correlation Studies

To understand model usefulness in finding spot overcrowding or bunching (uneven loading) situations, ridecheck is matched to model data by unique bus number, run, and stop time (using manual fuzzy match allowing for inevitable timing calibration issues and/or typos in vehicle/run numbers.) Figure 4(a,b,c) shows Bx19 comparison at individual trip-by-trip level using matched trips. AFC-AVL model tended to provide slightly lower boarding and disembarkation estimates than ridecheck figures (by up to 5% on average, although there were plenty of cases where AFC estimated a higher number), but slightly higher total passenger-stops (by about 8%; a passenger-stop is one passenger travelling one stop.) Overall, R-square for on/off is 95% and for passenger-stops is 90%, demonstrating direct correlation between two data sources.

Comparison to Experimental APC Data

In parallel with AFC model development, NYCTA installed experimental APC equipment in selected buses to gauge fleet wide deployment cost and feasibility. A comparison with APC data provided additional model validation opportunities but also highlighted needs for post-processing APC data.

Figure 3(g,j) reveals situations where APC software didn't correctly track beginning-of-trip, resulting in boardings at origin being missed and consequent negative loading near the terminus. Balancing algorithms fixed negative loading issues but had no basis for synthesizing missing boardings. Figure 3(e) shows classic passenger-remaining-onboard issues when disembarkations are not correctly counted. Figure 3(h) shows a very crowded trip where blocked front door sensors or other issues made it difficult track boardings correctly, whereas disembarkation was accurately counted leading to negative loads.

Figure 4(d) shows stop-level correlation analyses showing differences between AFC and APC data of about one-third. Likely reasons include situations shown in Figure 3(g,j) which might be corrected by keeping APC unit powered while bus engine is switched off and/or headsign not set in period prior to departure from origin.

While APC data streams, once stabilized, can be helpful, time-tested method of measuring farebox takings (i.e. AFC counts) continues to be an important data source for service adjustments. APC data requires substantial post-processing prior to use *(30)*. APC hardware is not reliable compared to AFC equipment. APC dedicated maintenance costs are relatively high for data collection, whereas AFC costs are shared with fare collection business needs. Additionally, O/D matrices or individual travel itinerary are not available directly from APC. Future work could focus on limited fleet deployment of APC, integrating sample APC data streams as calibration factors in an AFC-AVL model.

Aggregate Origin-Destination Matrix Validation

In late 2013, NYCTA studied transit service in Co-op City, a community of high-rise apartments in northern Bronx with over 40,000 residents. In conjunction with limited ridecheck and origin-destination market surveys, AFC-AVL model was used and generated satisfactory results. One unanticipated benefit was incidental availability of limited O/D survey data, used for additional model validation. Conducted throughout Co-Op city during weekdays, 1,363 customers were interviewed. Distribution of destinations are shown in Figure 4(e). Intermodal linked trips derived from AFC-AVL model for over 10,000 daily Co-op City originating trips yielded very similar results.

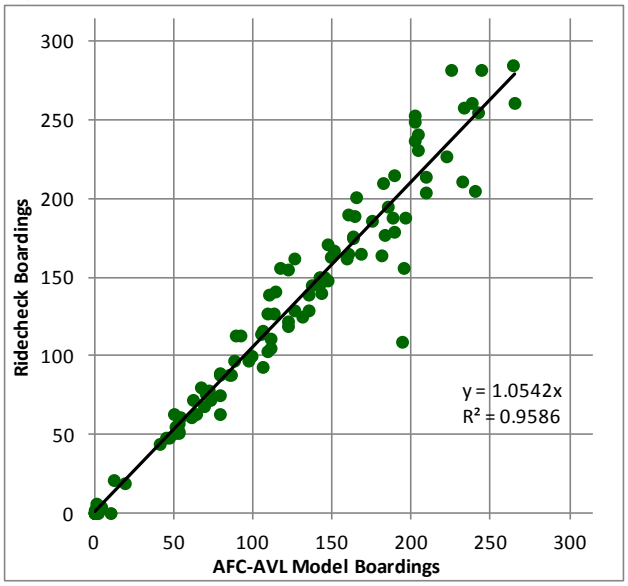
General Observations

Some bus routes have undergone intensive stop-by-stop comparison with ridecheck data. Few ridechecks were available for same-date comparison, thus most validation was completed by comparing monthly average data to ridecheck from earlier dates. Generally speaking:

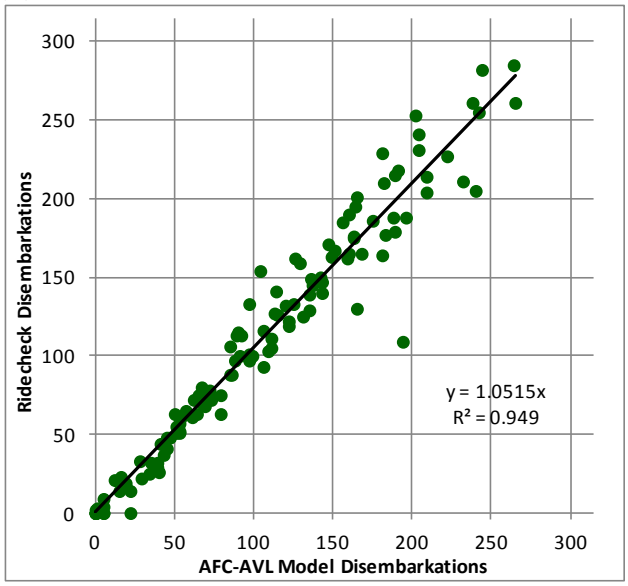
- Most comparisons yield very satisfying results in terms of load profile shape and values.
- Monthly averages usually generate better results than individual dates. AFC transaction time is rounded to nearest six-minutes. For an individual date, model may not accurately pin-point every stop. As monthly average, correct bus stops have higher probabilities of occurrence.
- Some routes (e.g. M116) were found to have less total ridership than ridecheck, which suggests that model underestimated fare evader numbers and/or students failing to swipe their MetroCard when boarding.
- Weekday results are usually more consistent than weekend. On weekends, services are more irregular due to events and G.O.s. Passengers are also more likely to combine bus with other transportation modes.

So far, 90% of express bus routes and over 40% of Bronx and Manhattan local routes have been validated against historic ridecheck data. Express bus was given priority due to ongoing NYCTA efforts to restructure and rationalize routes. Most routes generated satisfactory results; special exception processing continues to be added where necessary as model is fine-tuned and GTFS data is cleaned up.

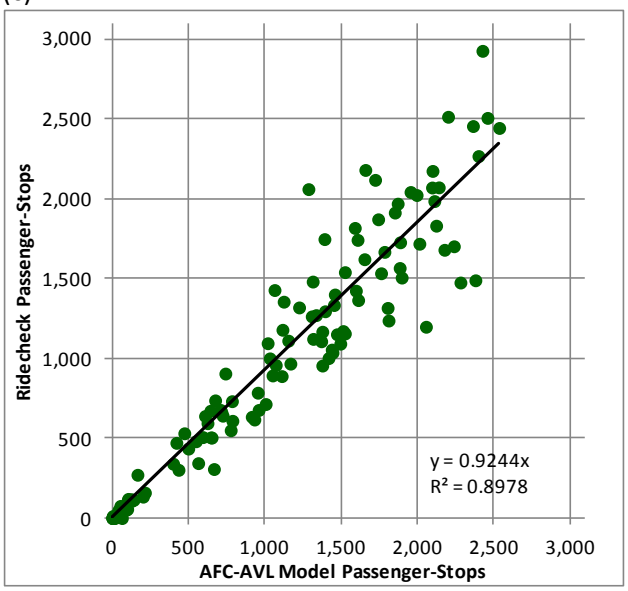
(a)



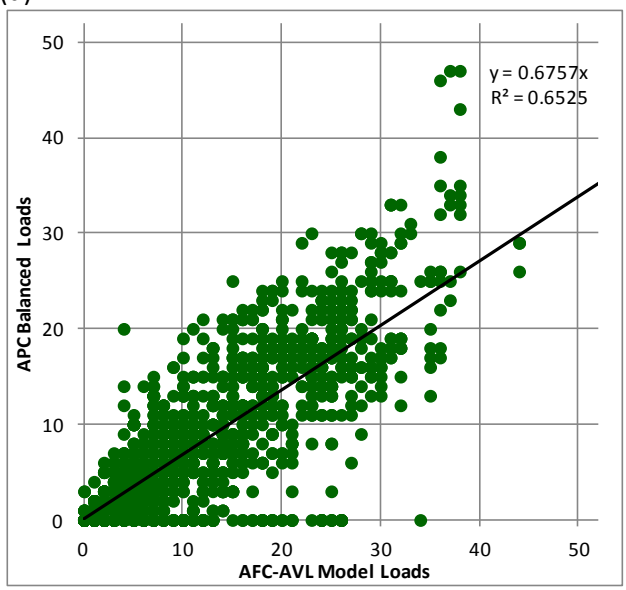
(b)



(c)



(d)



(e)

Origin	Destination	AFC-AVL Model	Market Survey
Co-op City	Co-op City	8%	12%
Co-op City	Bronx (outside Co-op City)	55%	59%
Co-op City	Manhattan	27%	20%
Co-op City	Brooklyn	3%	3%
Co-op City	Queens	5%	4%
Co-op City	Elsewhere	2%	2%

Figure 4 Advanced Algorithm Validation: (a)Time-Period Level Same-Day Model-Ridecheck Correlation—Ons; (b) Time-Period Level Same-Day Model-Ridecheck Correlation—Offs; (c) Time-Period Level Same-Day Model-Ridecheck Correlation—Sum of Link Loads; (d) Trip Level Same-Day Model-APC Correlation—Link Loads; (e) Model-Survey Comparison in Co-op City O/D Study.

CONCLUSIONS

New York City surface transit boarding and alighting production application's success with big data shows promise of estimating bus boarding and alighting counts with AVL data. Once limited to academic research, a daily production application has been implemented with high quality of results. This being a well-researched area, great benefits are leveraged when transit authorities discover production applications.

This application validates results with ridecheck, APC system, and market research surveys, with encouraging results. Critical passenger information like origin-destination matrix and travel patterns needed for surface planning and scheduling, traditionally obtained through labor-intensive, time and budget constrained survey, can now be acquired through applications taking advantage of big transit data. Besides underlying significant economic benefits, availability of data in term of quantity, timeliness, type detail, and location precision makes it a clearly better option than surveys.

The heart of this innovation, the "dynamic transfer link" network generation method in shortest path, proved to be effective and feasible in solving large scale and highly complex realistic transit networks. Even though most effort spent in customizing methodologies to complicated New York City surface transit system with various subsystems and exceptions, basic methodologies developed is easily adapted to other transit systems.

Future of Production Algorithms

As faster computational tools become available to work with even larger volumes of data, real-time analytics have become the norm in transportation. Specialized solutions like massively-parallel graphics co-processors *(42)* have been adapted for transportation analytics in research settings. Thus, needs for pure algorithmically-based optimization may decline in future. Nonetheless, core ideas demonstrated in this paper are useful for practitioners in the following areas:

1. **Exploratory research**—taking advantage of new data streams and to produce sample analyses (on commodity desktops) that could justify larger expenditures in specialized analytical equipment
2. **Short-term production**—public authorities have long procurement cycles for technological equipment; lean algorithms can be used in production on existing equipment (e.g. general purpose virtual servers), allowing analytics to be more responsive to operational needs
3. **Scope & capability expansion**—even with specialized equipment, explicit trade-offs to minimize execution time, problem simplification, and task-specific optimization would allow same equipment to handle larger problems.

With algorithmic innovations to optimize computation time, highly complex tasks can be manageably adapted for daily production. This successful effort shows it is feasible to introduce automation in transportation planning—taking advantage of rapid advancement of technology in separate data sources, and integrates them into a powerful information resource. With efficient algorithm designs, modern computer hardware can handle large-scale transportation modelling at detail levels, even on standard desktop PCs.

Further Applications

With production-driven generation of extensive O/D matrices on a daily basis, new concept of "micro transportation planning" emerges, where planning occurs at individual passenger level. Traditionally, transportation planning utilizes scarce and highly aggregated sample data. With daily processing of individual trips, farecard histories can be built over longer periods and more useful information could be inferred.

Compared to subway stations, bus stops are much closer to passenger actual residence and work locations. Over 80% of MetroCard trips start at identical stations each weekday *(37)*. This methodology was used in Co-op City study to identify residents. Most frequent stops other than residence might be work location. It could be possible to synthesize journey-to-work data. Based on MetroCard attributes (student, senior, and other status), together with Census socioeconomic data associated with home stops, valuable market research could be generated without unreliable and costly manual surveys.

Compared to traditional ridecheck, AFC-AVL model has many advantages:

- Ridecheck has limited time frame due to cost considerations, hence effect of day-to-day variation is unknown. Using average loads from AFC-AVL model can avoid over- or under-design caused by perturbations on sample day.
- Ridecheck is done few routes at a time. AFC model provides system wide information everyday, which assists planners in optimizing entire system by resource reallocation.
- Daily data can help planner to identify repeating patterns of operations or ridership disturbance.

Potentially, similar applications taking advantage of new data sources, algorithms, and hardware could contribute to a new era in micro-level transportation planning:

- Farecards could be linked to demographic (income, race, etc.) groups in aggregate, to observe travel patterns and economic geography.
- Observe impacts of non-recurring versus recurring passenger congestion.
- With suitable assumptions, trip purpose can be derived, and customer activity habits can be studied and marketing efforts could be coordinated based on this data.

ACKNOWLEDGEMENTS

Authors gratefully acknowledge support and assistance of Wan Chen, Stella Levin, and Rob Hickey during algorithm development, and Cheri Tan and Tuan Huynh for assisting with validation and development of the production system. Responsibility for errors or omissions remains with the authors. Opinions expressed or implied are the authors' and do not necessarily reflect official policies of any organization.

REFERENCES

(1) Lee, Sang Gu, M. Hickman, D.Q. Tong. Stop Aggregation Model: Development and Application. *Transportation Research Record: Journal of the Transportation Research Board 2276* (2012), pp. 38–47. TRB #12-1287.

(2) Barry, James Jerome, R. Newhouser, A. Rahbee, and S. Sayeda. Origin and Destination Estimation in New York City with Automated Fare System Data. *Transportation Research Record 1817* (2002), pp. 183-187. TRB #02-1045.

(3) Zhao, Jin Hwa, A. Rahbee, and N.H.M. Wilson. Estimating a Rail Passenger Trip Origin-Destination Matrix Using Automatic Data Collection Systems. In *Computer-Aided Civil and Infrastructure Engineering*, No. 22, pp. 376-387 (2007).

(4) Guptill, Robert. Data from MBTA's Automated Fare Collection (AFC). Presented at *TRB National Transportation Planning Applications Conference*, May 2009. Retrieved from http://www.trb-appcon.org/TRB2009presentations/s19/07_impact.ppt on June 27, 2010.

(5) Sun, Yanshuo and P.M. Schonfeld. Schedule-Based Route Choice Estimation with Automatic Fare Collection Data for Rail Transit Passengers. Presented at *Transportation Research Board 93rd Annual Meeting*. Paper #14-0834.

(6) Frumin, Michael, J.H. Zhao, N.H.M. Wilson, and Z. Zhao. Automatic Data for Applied Railway Management: Case Study on the London Overground. In *Transportation Research Record: Journal of the Transportation Research Board 2353*, pp. 47-56 (2013). TRB #13-2987.

(7) Wong, S.C., and C.O. Tong. Estimation of Time-Dependent Origin-Destination Matrices for Transit Networks. In *Transportation Research Part B: Methodological*, Volume 32, Issue 1, pp. 35-48, January, 1998.

(8) Barry, James Jerome, R. Freimer, and H.L. Slavin. Use of Entry-Only Automatic Fare Collection Data to Estimate Linked Transit Trips in New York City. In *Transportation Research Record: Journal of the Transportation Research Board 2112*, pp. 53-61 (2009). TRB #09-2812.

(9) Liao, Chen-Fu, and H. Liu. Mining Bus Location, Passenger Count and Fare Collection Database for Intelligent Transit Applications. Presented at the *21st Annual Transportation Research Conference*, April 27-28, 2010, St. Paul, Minn. Retrieved from http://www.cts.umn.edu/Events/ResearchConf/2010/presentations/24-liao.pdf on June 27, 2010.

(10) Ro, Wei-Yuan (羅惟元). Using Taipei Easycard Transaction Data to Explore the O-D Table of Bus Passengers. Masters Thesis, Graduate Institute of Transportation Management, Tamkang University, Damshui, Taiwan, June 2008. Retrieved from http://tkuir.lib.tku.edu.tw:8080/dspace/handle/987654321/33825 on June 27, 2010.

(11) Seaborn, Catherine, J. Attanucci, and N.H.M. Wilson. Analyzing Multimodal Public Transport Journeys in London with Smart Card Fare Payment Data. In *Transportation Research Record: Journal of the Transportation Research Board 2121*, pp. 55-62 (2009). TRB #09-3419.

(12) Morency, Catherine, M. Trepanier, and B. Agard. Measuring Transit Use Variability with Smartcard Data. In *Transport Policy* **14**:3, pp. 193-203.

(12a) Wang, Wei, J.P. Attanucci, and N.H.M. Wilson. Bus passenger O-D estimation and related analyses using automated data collection systems. In *Journal of Public Transport* **14**:4, pp. 132-150 (2011).

(14) Munizaga, Marcela, C. Palma, and D. Fischer. Estimation of a Disaggregate Public Transport OD Matrix from Passive SmartCard Data from Santiago, Chile. Presented at *Transportation Research Board 90th Annual Meeting*. TRB #11-0430.

(15) Farzin, Janine M. Constructing an Automated Bus Origin-Destination Matrix Using Farecard and Global Positioning System Data in São Paulo, Brazil. In Transportation Research Record 2072, Transportation Research Board of the National Academies, 2008.

(16) McCord, Mark R., R.G. Mishalani, P. Goel, and B. Strohl. Iterative Proportional Fitting Procedure to Determine Bus Route Passenger Origin–Destination Flows. In *Transportation Research Record: Journal of the Transportation Research Board 2145*, pp. 59-65 (2010). TRB #10-3425.

(17) Shi, Xin Yi and H.F. Lin. The Analysis of Bus Commuters' Travel Characteristics Using Smartcard Data: the Case of Shenzhen, China. Presented at *Transportation Research Board 93rd Annual Meeting*. TRB #14-2571.

(18) Zúñiga, Felipe G., J.C.A. Muñoz, R.E. Giesen. Real-Time Prediction and Update of Dynamic Origin-Destination Matrices on a Transit Corridor. Presented at *TransLog Transportation and Logistics Workshop*, Hamilton, Ont., Canada (2009).

(19) Furth, Peter G., B. Hemily, T.H.J. Muller, and J.G. Strathman. *Using Archived AVL-APC Data to Improve Transit Performance and Management*. Transit Cooperative Research Program, Report 113 (2006).

(20) Lu, Alex, and A.V. Reddy. Algorithm to Measure Daily Bus Passenger Miles Using Electronic Farebox Data for National Transit Database Section 15 Reporting. In *Transportation Research Record: Journal of the Transportation Research Board 2216*, pp. 19-32 (2011). TRB #11-0368.

(21) Levine, Brian, A. Lu, and A.V. Reddy. Measuring Subway Service Performance at New York City Transit: A Case Study Using Automated Train Supervision (ATS) Track-Occupancy Data. In *Transportation Research Record: Journal of the Transportation Research Board 2353*, pp. 57-68 (2013). TRB #13-2997.

(22) Reddy, Alla V., A. Lu, S. Kumar, V. Bashmakov, and S. Rudenko. Entry-Only Automated Fare-Collection System Data Used to Infer Ridership, Rider Destinations, Unlinked Trips, and Passenger Miles. In *Transportation Research Record: Journal of the Transportation Research Board 2110*, pp. 128-136 (2009). TRB #09-0809.

(23) Hickey, Robert L., A. Lu, and A.V. Reddy. Using Quantitative Methods in Equity and Demographic Analysis to Inform Transit Fare Restructuring Decisions. In *Transportation Research Record: Journal of the Transportation Research Board 2144*, pp. 80-92 (2010). TRB #10-0280.

(24) Holmes, Mark. HASTUS Implementation at MTA Bus Company. Presented at *American Public Transit Association Multimodal Operations Planning Workshop*, Seattle, Wash. (2011). Retrieved from http://www.apta.com/mc/multimodal/previous/2011/Presentations/Session-10-MTA-Bus-Company's-Transition-into-HASTUS-M-Holmes.pdf on July 5, 2014.

(25) Levine, Brian, S. Iyer, and A.V. Reddy. Development of Automated Vehicle Location (AVL) Data Based System to Improve Bus Service at New York City Transit. Presented at *Transportation Research Board 93rd Annual Meeting*. TRB #14-1711.

(26) Glikin, Michael, and A.V. Reddy. MTA Bus Time Implementation & New Applications. Presented at *American Public Transit Association Multimodal Operations Planning Workshop*, San Francisco, Calif. (2013). Retrieved from http://www.apta.com/mc/multimodal/previous/2013/presentations/Presentations/Glikin%20Reddy%20MOPW%202013.pdf on July 5, 2014.

(27) Barbeau, Sean J., A. Borning, and K. Watkins. OneBusAway Multi-region: Rapidly Expanding Mobile Transit Apps to New Cities. Presented at *Transportation Research Board 93rd Annual Meeting*. TRB #14-4285.

(28) Metropolitan Transportation Authority. Archived B63 Bus Time Data. Retrieved from http://bustime.mta.info/wiki/Developers/ArchiveData on July 5, 2014.

(29) Thorup, Mikkel (1999). Undirected Single-Source Shortest Paths with Positive Integer Weights in Linear Time. In *Journal of the Association for Computing Machinery* **46**(3):pp.362–394. doi:10.1145/316542.316548.

(30) RSM Services. Ridecheck Plus—Denver, Colo. Case Study. Retrieved from http://www.rsmservices.com/rsm-projects/Denver-CO/30.html on July 12, 2014.

(31) Tollerson, Ernest. MTA's Open-Data Opportunities in 2013 and Beyond—Developing All-Agencies Policies, Projects & Public Private Partnerships (P3s). Presented at *APTA TransiTech Conference*, Phoenix, Ariz., March 20, 2013.

(32) Reddy, Alla V., J.A. Kuhls, and A. Lu. Measuring and Controlling Subway Fare Evasion: Improving Safety and Security at New York City Transit. In *Transportation Research Record: Journal of the Transportation Research Board 2216*, pp. 85-99 (2011).

(33) Beaton, Eric B., T.V. Orosz, R. Ravit, R. Thompson, and D. Tyson. A Limited for the 21st Century: Applying BRT Principles to Create Select Bus Service on Hylan Boulevard. In *TRB 93rd Annual Meeting Compendium of Papers*. TRB #14-2807.

(34) Beaton, Eric B., E. Bialostozky, O. Ernhofer, T.V. Orosz, T. Reiss, and D. Yuratovac. Designing Bus Rapid Transit Facilities for Constrained Urban Arterials: Case Study of the Selection Process for the Webster Avenue Bus Rapid Transit Running Way Design in New York City. In *Transportation Research Record: Journal of the Transportation Research Board 2352*, pp. 50-60.

(35a) Sheffi, Yosef. Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods. 1984. Prentice-Hall, Inc., Englewood Cliffs, New Jersey 07632.

(35) Beaton, Eric B., J.E. Barr, J.V. Chiarmonte, T.V. Orosz, D. McAfee, and A. Sugiura. Select Bus Service on M15 in New York City: Bus Rapid Transit Partnership. In *Transportation Research Record: Journal of the Transportation Research Board 2277*, pp.1-10 (2012). TRB #12-1809.

(36) Barr, Joseph E., E.B. Beaton, J.V. Chiarmonte, and T.V. Orosz. Select Bus Service on Bx12 in New York City: Bus Rapid Transit Partnership of New York City DOT and Metropolitan Transit Authority New York City Transit. In *Transportation Research Record: Journal of the Transportation Research Board 2145*, pp. 40-48 (2010). TRB #10-1693.

(37) Tempest Ludlow, Pete, and T. Stasko. Validate O-D Assumption Using Entry Only Swipe Data. Working Paper, System Data Research, Operations Planning, New York City Transit (2014).

(38) Gordon, J. B., H. Koutsopoulos, N. Wilson, and J. Attanucci. Automated Inference of Linked Transit Journeys in London using Fare-Transaction and Vehicle Location Data. In *Transportation Research Record: Journal of the Transportation Research Board 2343*, pp.17-24 (2009).

(39) Cui, A. Bus Passenger Origin-Destination Matrix Estimation Using Automated Data Collection Systems. MS Thesis. Massachusetts Institute and Technology, Cambridge, 2006. Retrieved from http://dspace.mit.edu/handle/1721.1/37970 on July 15, 2014.

(40) Furth, Peter G., J.G. Strathman, and B. Hemily. Making Automatic Passenger Counts Mainstream: Accuracy, Balancing Algorithms, and Data Structures. In *Transportation Research Record: Journal of the Transportation Research Board 1927*, pp.207-216 (2005).

(41) Li, Da Ming, Y.J. Lin, X.L. Zhao, H.J. Song, and N. Zou. Estimating a transit passenger trip origin-destination matrix using automatic fare collection system. In *Proceedings of the 16th International Conference on Database Systems for Advanced Applications*, pp. 502-513 (2011).

(42) Zhang, Jianting, S.M. You, and L. Greenwald. High-Performance Spatial Query Processing on Big Taxi Trip Data using GPGPUs. Retrieved from http://www-cs.ccny.cuny.edu/~jzhang/papers/taxi_p2pdis.pdf on July 18, 2014.