TRB Paper Manuscript #11-0368

An Algorithm to Measure Daily Bus Passenger Miles Using Electronic Farebox Data for National Transit Database (NTD) Section 15 Reporting

*Alex Lu, Alla Reddy**
** Corresponding author*

Alex Lu
Principal Transportation Planner
New York City Transit
2 Broadway, Cubicle A17.111, New York, N.Y. 10004-2208
Tel: (646) 252-5664
Email: *Alex.Lu@nyct.com*

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

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

Word Count: 232 (Abstract) + 5,692 (Text) + 6 * 250 (Figures) = 7,424 Words.
Submittal Date: February 28, 2011.

Note to Editors: Certain figures embedded in this document are not publication-resolution. Higher resolution artwork will be supplied separately to Phyllis Barber at <*pbarber@nas.edu*>.

ABSTRACT

New York City Transit (NYCT) implemented an automated algorithm to estimate daily bus unlinked trips, infer passenger-miles, and compute average trip lengths by route using transaction data from an entry-only Automated Fare Collection (AFC) system. Total onboard miles are inferred by taking advantage of symmetries in bus passengers' daily activity patterns. NYCT's algorithm utilizes rigourously-tested engineering assumptions to detect common data errors from mechanical failures, imperfect driver-farebox interactions, and operational reality, applying statistically measured adjustment factors to correct or interpolate for missing passengers from non-AFC boardings and malfunctions. Surveys revealed that under typical operating conditions, non-AFC passengers and farebox data transmission errors accounted for 12% and 5½% of missing ridership, respectively. The fault-tolerant algorithm uses non-geographic transaction data from an AFC system without Automated Vehicle Locator (AVL) functionality, directly computing aggregate passenger-miles by inferring origin locations from transaction timestamps using scheduled average speed assumptions, and without assigning each passenger's precise destination. NYCT focused on fully automatic, production-ready algorithms by rejecting alternatives requiring excessive coding effort, processor time, difficult-to-obtain data, or manual intervention in favour of logical inference, statistical estimation, and symmetry. Meticulous parallel testing demonstrated that resulting average trip lengths are stable across days and correlate well with manually collected stop-by-stop ridership data. Annual passenger-miles are within –1% to 4% of the National Transit Database (NTD) ±10% sample data and were approved by Federal Transit Administration (FTA) for NTD Section 15 submission.

INTRODUCTION

This paper presents New York City Transit's (NYCT) algorithm to directly estimate route-level daily bus passenger-miles and average trip lengths for National Transit Database (NTD) Section 15 reporting from transaction data streams generated by entry-swipe fareboxes not equipped with Automatic Vehicle Locator (AVL) systems. At this algorithm's core are engineering assumptions designed to simplify data processing and minimize manual exception-handling requirements, permitting a high degree of automation while tolerating common data errors from mechanical failures, imperfect passenger-farebox or driver-farebox interactions, and operational reality.

The algorithm's program implementation reads daily Automated Fare Collection (AFC) system transaction data and outputs route-level unlinked trips and passenger-miles. Bus trips are inferred from transaction data by observing sequence of headsign changes in each vehicle's data. Passenger boarding locations en-route is determined using transaction timestamps relative to times when bus trips started (converting time to distance using route-specific time-of-day-dependent scheduled running time profiles). Aggregate passenger-miles travelled are estimated by taking advantage of statistical symmetry between opposing-direction boarding and alighting activities within a 24-hour period, without inferring each passenger's precise destination. Correction factors developed from one-time surveys plus routine reports adjust for "missing" passengers – for reasons ranging from fare evasion to AFC equipment malfunction.

The resulting annual passenger mile estimate is comparable to or better than traditional Federal Transit Administration (FTA) prescribed 700 trips-per-year (*1*) stratified ridecheck sample method (95% ±10% error margin) and was approved by FTA for NTD reporting in lieu of sampling.

Relationship to Prior Work

Transportation planners routinely use magnetic and contactless farecard data for off-line planning and modelling purposes in cities varied as Chicago (*2*,*3*), New York (*4*,*5*), Boston (*6*), Minneapolis (*7*), London (*8*,*9*,*10*), Taipei (*11*), Hong Kong (*12*), China (*14*), and elsewhere (*15*,*16*,*17*), and farebox receipts on an aggregate basis have long been used to monitor ridership trends and adjust service levels (*18*,*19*). However, most analysis methods are complex and often required analysts to "massage" raw AFC data, reducing their effectiveness for routine daily reporting.

Early-generation AFCs (like one Scotland's Lothian Buses used in the 1990s) required drivers to punch at every timepoint, for verification of zone fares, potentially allowing construction of origin-destination (OD) matrices. In Taipei, route-by-route OD matrices have been estimated from the non-AVL Taipei Youyoka (EasyCard) data using transaction time differential and average travel speed methods, using transactions on the same farecard over multiple days to estimate running times (*11*). Extensions taking advantage of transferring passengers' farecard traces (providing more accurate localizations by estimating when specific buses served transfer points) were proposed (*20*), but these approaches are too complex for fully-automated production environments where zero manual intervention is a goal.

Many smartcard systems now use physical location (although generally not AVL data) to charge zone fares. In rural Formosa, the TaiwanTong (Taiwan Easy Go) smartcard uses AVL-equipped fareboxes and a tap-on, tap-off system to charge mileage-based fares on intercity local buses on a production basis since 2007 (*21*). NJ Transit uses a Global Positioning System (GPS) based system for reporting boarding stops, but not zone fare computation (*22*). However, the authors are not aware of any bus operators inferring geographic information on a systemwide routine daily basis from non-AVL fareboxes.

Although NYCT developed this algorithm independently, subsequent literature review revealed a key symmetry assumption was previously demonstrated on two Pittsburgh light rail lines (*23*), five Los Angeles bus routes using "location-stamped" farebox data, and verified by comparison with APC (*24*). This work's contribution is, therefore: (a) an additional sixteen-route New York dataset generally supporting the symmetry assumption, but also demonstrates circumstances where it may not hold; (b) although not immediately transferable to other AFC systems with different error patterns, assumptions devised to handle faulty farebox data is helpful as a case study; (c) methods and simplifying assumptions allowing fully automated data analysis without manual intervention, required for 100% data reporting at large agencies; (d) together with classic code-optimization approaches, simple geometric transformations allowing sequential processing of each bus's farebox data without explicitly enumerating centroids or route load profile histograms, producing reasonable execution times (~2.5 million daily transactions in ~3 minutes), another production deployment prerequisite.

Another contribution is the large-scale and thorough surveys and electronic data analyses conducted in a fairly comprehensive estimation of factors describing numerous reasons for AFC data losses under NYCT's typical operating conditions. Although these factors are not directly applicable to other metropolitan areas, it likely provides the first published survey in several years that examines the fraction of bus ridership that isn't captured by fare collection systems – the "AFC unaccountable" riders.

Brief History of Automated Fare Collection

First faregates in United States were installed experimentally in 1964 at Forest Hills and Kew Gardens Long Island Rail Road stations in Queens (*25*); first systemwide installation was on Illinois Central Railroad (IC) in 1965 for its busy Chicago commuter service (today's Metra Electric.) Financed entirely from private funds, AFC was expected to reduce operating costs by decreasing on-board crew sizes and eliminating station agents at all but busiest stations. Cubic's IC system featured entry-exit swipes (NX) to enforce zonal fare structures, checks against fraud, used ticket collection, and ridership/revenue data collection capabilities (*26*). It served as prototype for the San Francisco Bay Area Rapid Transit (BART) (*27*), Washington Metropolitan Area Transit Authority (WMATA) (*28*), and Philadelphia's Port Authority Transit Corporation (PATCO) Lindenwold Line NX-zonal AFC systems (*29*). These railroad-style systems required complex computer data processing on faregates or remotely on a central computer, and thus weren't suitable for buses. Similar systems are still in use on Japan's and Taiwan's commuter railroads, and London Underground (*30*).

Metropolitan Atlanta Rapid Transit Authority (MARTA)'s desire for simpler AFC systems resulted in Duncan (traditionally a parking meter vendor) developing turnstile machines for entry-only subway fare collection. Chicago Transit Authority (CTA)'s ChicagoCard, Boston Massachusetts Bay Transportation Authority (MBTA)'s previous generation "T-Pass", and NYCT's MetroCard systems could all be considered MARTA's 1977 system's conceptual descendents.

Bus fareboxes had hitherto been much simpler devices, mechanically registering coins deposited on accumulating registration counters. Duncan's 1973 "Faretronic" farebox was first to electronically count coins and collect revenue/ridership data by fare class. Keene quickly followed suit, introducing a design meeting Urban Mass Transit Administration (UMTA) Section 15 reporting requirements, also collecting fuel consumption and bus mileage data (*31*). In New York, mechanical fareboxes were preferred for ease of maintenance until widespread deployment of Cubic's MetroCard for buses in 1997. Venerable GFI fareboxes featuring magnetic pass readers requiring cash single fares lasted in Boston until Scheidt-Bachmann's CharlieCard was introduced in 2006.

Purpose and Need

Prior to development of GPS, Automated Passenger Counters (APC), and AVL systems, fareboxes could be instrumented to record revenue trips, bus mileage, fuel consumption, and even engine maintenance related data, but geographical information couldn't be recorded. However, Section 15 has required revenue passenger-miles data since at least 1978 (*32*). FTA's reporting manual advises transit agencies without APC/AVL to use ridecheck sampling (*1*) – assigning surveyors to ride buses from origins to destinations, a time-consuming process.

More recently, FTA recommended conversion to 100% electronic data reporting. With FTA support, NYCT developed this algorithm to leverage daily AFC data streams, uniquely tailored to MetroCard system's data recording methods. While NYCT's primary motivation was to simplify auditing, improve data quality, obtain reliable monthly data unavailable from annual sampling, and avoid manual data collection pitfalls, this implementation produced savings in both survey and analytical resources, although one-time investment in algorithm development and programming was required.

Issues with Surveyor Data Collection

Several issues are inherently problematic with surveyor data collection: high costs, difficulty of processing large passenger volumes, missed assignments, data interpretation issues, data entry and analysis costs, and potential data collection inconsistencies (*5*). Specifically for NTD bus passenger-miles data, several other considerations make sample methodologies challenging:

1. NTD requires monthly ridership reporting, but sample design calls for 95% ±10% error margin annually. Monthly results therefore vary widely (implied error margin is about ±30%), making results difficult to explain and practically useless.
2. Missed assignments are particularly problematic for NTD data, because specific bus trips are randomly scheduled per NTD methodology requirements. If surveyors miss that trip for any reason (e.g. travel delay), substitutions by subsequent trips are unacceptable. Extra random samples must be taken, increasing survey costs.

3. Unlike NYCT's other surveys, NTD ridechecks are conducted for entire span of service – 24 hours daily, 7 days per week. Data collection resource costs during "off-hours" cannot be shared with routine surveys, requiring dedicated overnight and weekend surveyors.

ELECTRONIC FARE MEDIA DATA

NYCT's farebox data is captured in two files, Transaction (EU65) file and Trip file. EU65 is generated daily and contains one record for each MetroCard point-of-entry (POE) transaction (subway/bus). Each record (Figure 1(a)) contains farecard ID, date, time, transaction type, fare media class, bus number, carrier, farebox number, value deducted, and POE "location" code. Subway POEs identify turnstile location, but bus codes identify only route/direction (from destination rollsign). Without AVL, boarding locations cannot be known exactly. Transaction time is rounded to nearest one-tenth hours (i.e. six minutes).

Each Trip file record contains information about *partial* trips. New records are generated when:

1. New driver logs on (at mid-route relief points);
2. Destination rollsign is changed (including mid-route, when authorized to bypass stops with 'Next Bus Please' sign);
3. When predetermined times are reached (e.g. 09:00, 09:30, 15:30 – when tariff changes for certain fares).

Separate records must be combined to determine trip-level information. Trip data mostly relates to drivers (Run No., Employee No., etc.) and fareboxes' cash register functions (cash received by fare class, reduced fares paid, paper transfers issued, etc.)

Designed for farebox and fare media auditing, both Transaction and Trip file data models are not completely normalized, and not ideal for data mining. Available computer technology at reasonable costs during MetroCard's design phase (early 1990s) influenced these decisions. Transaction data was constrained by available storage on MetroCard's magnetic stripes (16 bytes per transaction). Farebox "probing" time (downloading daily data during farebox emptying) was also a likely consideration.

Select Bus Service (SBS) MetroCard Fare Collectors (MFCs)

NYC Department of Transportation (NYCDOT)'s Bus Rapid Transit (BRT) service is branded **+selectbus**service and utilizes off-board fare collection with Proof-of-Payment (POP). MFC fare validation machines and Parkeon's Coin Fare Collector (CFC) meters are installed at each stop. Riders pay prior to boarding using either MetroCards or cash. Each transaction generates a receipt, which may be examined by inspectors any time while travelling (*33*). Receipt-issuing transactions are logged. NYCT's Office of Management and Budget (OMB) analyzes numerous data streams manually to produce monthly summary revenue reports by fare media, location, and direction.

Common Farebox Data Errors and Problems in Interpretation

AFC has built-in validation features to ensure internal consistency. Problems in interpretation arise because of operating procedures and practices that AFC wasn't designed to capture, resulting in missing data or data that incorrectly describe field operations (*19*). Difficulties fall within four broad categories:

1. **Drivers Don't Correctly Change Headsigns:** AFC relies on destination sign information to determine route/direction. Headsign units occasionally malfunction and show garbled destinations; some drivers then don't properly enter signcodes. Fareboxes thus encode entire day's transactions to one direction.
2. **Farebox Doesn't Correctly Register Cash Fares:** Cash transactions are recorded as segment totals by fare class, in Trip file. Ordinary fare is $2.25, but special passengers pay half fare ($1.10). Normally, concessionary passengers show appropriate identification to drivers, who push a button to register half-fares. However, many don't wait for drivers to setup fareboxes, often resulting in incorrectly registered half-fares, like two half-fares becoming a full fare. Totals thus couldn't be used to ascertain transaction counts.
3. **Passengers Fail to Pay Correct Fares:** From AFC's perspective, passengers not paying fares shouldn't exist in transaction files. While fareboxes provide an "evader" button (Key 5), this isn't consistently utilized by operators. For passenger counting purposes, evaders occupy seats and therefore are passengers. Thus, transaction counts don't necessarily translate directly into passengers. Furthermore, some passengers pay partial fares ('short drop'), using coins, partially-loaded farecards, or combinations of both. The not-quite-one-to-one relationship means that special care is required when interpreting transactions.
4. **Actual Trips Operated Don't Match Published Schedules:** Fareboxes record data about field operations, including actual Run No. and departure time. Matching AFC data to scheduled trips is very difficult, because of dispatchers' ad-hoc changes in response to traffic conditions. No electronic record is generally kept of these adjustments. Paper forms are voluminous and stored separately at 19 bus depots, making them difficult to corral and key. Automated matching using departure time, Bus No., and Run No. produced approximately 70%~80% matches. Essentially, AFC data cannot be matched to schedules for extensive daily analyses.

VALIDATING NECESSARY ALGORITHM ASSUMPTIONS

Limitations in farebox data streams meant engineering assumptions were necessary for passenger-miles inference. Due to difficulties in matching trips, mileage estimation uses transaction data alone, supplemented with scheduled running times.

Localization Relative to Trip Origins

To determine locations from transaction times, buses are assumed to enter service at route origin, and must reach final destination before changing directions or deadheading to depot. The speed-distance-time formula is assumed to convert time elapsed since trip begin (called 'relative time') into approximate boarding locations along the designated route – like "dead-reckoning" navigational devices, which tracks distance from known points accurately to derive current location. Obviously, traffic conditions and schedule accuracy affect localizations.

Average NYCT bus speeds of 6~12 mph and transaction times rounded to nearest six minutes imply geographical discriminating power of approximately 0.6~1.2 miles (groups of 2~4 bus stops). While this range seems broad for individual boardings, distance travelled can be thought of as the *difference* between two uniformly distributed random variables, making it an unbiased estimator of actual distance:

1. Because estimated boarding and alighting times are subject to, on average, same uniform rounding aberrations, errors can cancel each other out (Figure 1(b)).
2. Distribution of rounding errors is symmetrical and triangular; passengers are equally likely to board near six-minute period beginnings (mileage underestimation) as near the ends (mileage overestimation). The expected error is zero.

Thus, in aggregate, transaction time resolution doesn't pose undue difficulty for reporting total passenger miles and average trip lengths, although this property makes AFC data unhelpful for accurately monitoring *individual* passenger ODs.

Constant Average Speed Assumption

Constant average speed assumptions implicit in distance-time conversions are a little more problematic, as NYCT 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. Certain Crosstown buses (e.g. M72, M86) operate non-stop on Central Park's parkways at higher speeds than city streets at both ends. Other Crosstowns (e.g. BX12, BX40/42) have "suburban ends" in Eastern Bronx that operates at somewhat higher speeds than Western "urban ends". Using average speeds leads to biases in these cases.

However, impacts of "asymmetrical" routes generally cancel out. Typical local routes (Figure 1(c)) generate equal overestimation for boarding passengers (slash pattern, cumulative speed > route average) as underestimation (checkboard pattern). Express routes have lower-than-average speeds in suburban pickup zones, leading to minor overall mileage underestimations (Figure 1(d)). NYCT's scheduling software can actually provide average speeds between timepoint locations (major enroute stops). With additional development work, routes could be split up into zones with different scheduled average speeds, to improve accuracy, particularly on express routes.

These assumptions are difficult to validate directly. Validating inferred locations entail equipping test vehicles with AVL and comparing AVL data with inferred AFC locations over many different routes. As long as localization errors cancel out on average and without significant bias, estimated passenger miles are accurate.

Symmetric Daily Activity Pattern Assumption

AFC reports fare payment transactions, which occurs upon boarding on NYCT's buses, making it impossible to ascertain alighting times or locations directly. Two assumptions allow relative times (and thus locations) of disembarkations to be inferred on an aggregate basis:

1. **Conservation of Passengers:** One passenger boarding a bus at one location sometime during the day is balanced by another passenger deboarding another bus at that location sometime that day. Essentially, this describes bus passengers' typical behaviours. Passengers leave their home stop in the morning and arrive at that same stop at the day's end; passengers disembarking at work are picked up from the same stop at lunch to run errands; those alighting to run errands are collected moments later from there, and so on.

2. **Equal and Opposite Passenger Activities in Opposing Directions:** Passengers alighting for any reason boards another bus from the same location, on the same route or group of routes, in the opposite direction. Since 70% of NYCT stops are served by only one route, passengers arriving and leaving by bus generally don't have options besides making return trips on an opposing number. Even where multiple routes serve one stop, shared routes (e.g. BX40/BX42, BX1/BX2) often have equal trunk service frequencies, thus passengers arriving on one route and returning on another is compensated by passengers arriving on the latter and returning via the former. This assumption is violated when passengers switch to different modes or "triangulate" by travelling to non-origin locations on their second trip. However, this is rare, and where it happens, effects likely occur equally in both directions, e.g. likely as many passengers alight from eastbound BX12 at Fordham Plaza to run errands then continue east – as passengers who alight from westbound BX12 and continue west.

To validate these assumptions, Surface Ridecheck (SR) data, collected for schedule-making purposes, is used. SR provides intensive record of all boarding and alighting locations on all bus trips for one particular route on survey day (except for small fraction of missed trips). This data, when summarized by stop for the whole day, demonstrates close correlation between daily boarding and alighting activities at all stops, and that boardings in one direction closely correlates with alightings in the other. BX55 is a north-south limited-stop feeder that replaced Bronx's Third Avenue el. BX55's On-Off correlation R-squared of close to one demonstrates that opposing direction daily boardings is a good predictor for daily alightings (Figure 2(a),(b)).

However, proving that correlation holds for BX55 doesn't mean it holds systemwide. A (non-scientific) sample of different routes types was selected and R-squared computed for cumulative boardings (Ons) and reverse-direction alightings (Offs). Although R-squared remained high for many route types (Figure 2(e)), several low values require explanation. S60 (Grymes Hill) is a circulator with low ridership (210 weekday riders) whose purpose is basically to drive passengers "up the hill" from Victory Blvd, resulting in an asymmetrical dogbone-shaped route-path (Figure 2(c)) that serves more stops and is longer uphill. Low ridership also contributes insufficient riders to show symmetry. B51 (Manhattan Bridge) has an asymmetrical baseball-cap shape (Figure 2(d)) with limited pick-up stops outbound but makes many drop-off stops inbound, serving as distributor in Manhattan's Chinatown for bus passengers transferring from other buses in Downtown Brooklyn, but not the reverse. Essentially, these routes have asymmetrical shapes, violating symmetry assumptions. Both "oddball" routes were eliminated in the 2010 Route Rationalizations (*34*).

Unbalanced Route Assumption

NYCT has "unbalanced" bus routes, i.e. significantly different daily inbound and outbound ridership. General reasons include: (a) different late-night travel patterns, particularly for shift workers; (b) morning rush carpooling and drop-offs when the reverse isn't feasible in afternoons; (c) trip triangulation from different afternoon modal preferences and/or additional activities like after-work errands.

Figure 2(e) actually shows several routes with substantial unbalance: M15 (1/2 Av), busier southbound (predominantly morning rush direction) because of substantial afternoon traffic congestion, diverting riders to parallel Lexington Av subway; B42 (Canarsie Beach), busier southbound (afternoon) because of timed outbound connections with Canarsie Line at Rockaway Parkway; B31 (Mill Basin), busier southbound (afternoon) because of substantial competition with MTA Bus's BM4, like all express buses, is busier (and more frequent) in the morning; Q46 (Union Tpk), Q83 (Liberty Av), busier eastbound (to Long Island) because of morning carpooling. Two assumptions allow unbalanced routes not to be treated differently:

1. **Symmetry for Unbalanced Routes:** Symmetric daily activity pattern is assumed to hold, as substantiated by high R-squared on these routes.
2. **Small Convexity Differences:** Convexity differences between true disembarkation cumulative ridership curves and opposing-direction boarding curves are small, such that algorithm's curve-swapping implementation (Figure 4(d)) doesn't introduce unacceptable passenger-mile estimation errors by not explicitly scaling curves to match unbalanced opposing-direction boardings.

End-of-Trip Assumption

If maximum trip time (plus reasonable recovery time) is reached and no changes in headsign occurred, it is assumed that driver forgot to do so. The bus is assumed to turn around and begin return trip in the opposite direction. Swipes subsequent to this point are assigned to the subsequent trip. This process is repeated until next headsign change is detected (occurs when subsequent relief drivers 'remember' to change it), or until bus returns to depot for the day.

Correction Factors Assumption

Only transaction data is used in this algorithm, making correction factors necessary for non-AFC customers (cash, evaders, flash passes, etc.) Non-AFC passenger boarding locations and counts (due to reduced-fares, short-drops) cannot be reliably ascertained, thus no mileages can be calculated. Correction factors for counts and miles implicitly assume non-AFC customers (25% of ridership) have substantially similar trip patterns to AFC customers – i.e. average trip length for non-AFC is same as AFC customers. Available evidence suggests that cash riders are not equally distributed throughout Chicago's subway system (*3*), and both cash fares and evaders are likely concentrated in lower-income areas in New York (*35*). Due to NYCT's substantial interlining and split-depot bus operations, cash data isn't available by route. With additional research and/or data collection, correction factors can be calibrated at a depot or route level, potentially capturing between-neighbourhood variations.

THE AFC PASSENGER MILEAGE ALGORITHM

This algorithm's input is NYCT's daily AFC file, described above. Output is a list of partial mileage integrals (representing each six-minute period on every trip), when summarized, produces total passengers and passenger-miles traveled, by route, for that day. With this algorithm, passenger miles cannot be inferred for each trip or each direction because passenger disembarkations don't generate swipe records, requiring the symmetric daily activity pattern assumption (Figure 3(a)). Average trip length is computed from passenger miles and counts.

With assumptions in place, algorithm derivation essentially reduces to several fairly basic transformations on a numerical integral of cumulative On-Off (passenger) and Location (miles) curves (Figure 3(d)). To identify cumulative activity curves using AFC data:

1. Filter out only bus records from bus and subway combined transaction file;
2. Sort bus transactions by bus number, then date/time, to put specific vehicles' daily activity in temporal sequence;
3. Cut transaction records into trips, using headsign (location code) changes and end-of-trip assumptions to identify new trips;
4. Convert swipe date/time into time elapsed since last headsign change ('relative time'), translate to miles along the route using scheduled average speeds.

Lookup tables (Figure 4(b)) relates location codes, day types, and time-of-day (by hour) to route, direction, route length (in miles), and running time. Because both route length and running time is time-of-day dependent, transaction time (to nearest hour) is used to lookup distance estimation (from trip origin). Once computed, cumulative On-Off curves are fed into the algorithm's integration stage (Figure 3(b)). As the flowchart (Figure 4(a)) demonstrate, most program control logic deals with manipulating transaction file, detecting swipes, six-minute periods, and trips. Actual numerical integral for computing passenger miles constitutes merely a few lines of code, due to efforts expended in seeking geometric transformations that simplify calculations and reduce program execution time. The production program processes one weekday in ~3 minutes on typical computer workstations.

Lookup tables are updated four times a year when schedules change. Queries extract this data directly from NYCT's computerized crew scheduling and run-cutting system.

DERIVING CORRECTION FACTORS

Despite widespread adoption of MetroCard for subway fare payment (>95%), AFC achieves only 75% market share amongst paid bus riders, because only 93 of the 2,265 MetroCard Vending/Express Machines (MVM/MEMs) are located at bus stops. The wide retail MetroCard distributor network isn't always immediately convenient to every bus patron. Bus fareboxes labour in extremely harsh conditions, subject to constant vibration, vandalism, and foreign materials introduced by wet or sticky farecards. Transactions are sometimes lost even after fares are correctly deducted from valid farecards. While New York has strong criminal laws that permit up to seven years' imprisonment for persons assaulting public transit employees (*36*,*37*), NYCT operates in tough urban neighbourhoods where some customers blatantly refuse to pay fare or even attack drivers (*38*).

For all these reasons, correction factors must be applied to AFC data to correct for these "AFC Unaccountable" ridership – actual trips consumed that aren't recorded in transaction file. Generally speaking, unaccountable ridership falls within three broad categories:

- **Cash Passengers**, whose transactions cannot be individually ascertained. NYCT's farebox reports cash received by trip and fare class, not individual transactions.
- **Non-Farebox Passengers**, who don't interact with fareboxes due to reasons like broken fareboxes, fare evasion, paper tickets, and flash passes.
- **Farebox Data Transmission Errors** result in missing transaction data even though AFC fares were paid normally.

Cash Passenger Adjustment

NYCT produces monthly summary actual revenue reports by route and fare media. Bus trips are broken out into percentages by fare media. Cash revenues registered divided by full cash fare (currently \$2.25 for local buses) yields an approximate (lower bound) cash passenger count. Based on the correction factor assumption, cash adjustment factor (updated monthly) is the ratio of cash passengers (non-AFC fares) to farecard transaction counts (AFC fares), before adjustments for non-farebox passengers.

Non-Farebox Passenger Adjustment

Non-farebox adjustment is the most complicated factor to estimate, requiring special surveys that measure fraction of fare evaders, pass riders, paper tickets, and fraction of trips operated with malfunctioning fareboxes (resulting in operator permitting passengers to board without fare).

FTA requested, and NYCT conducted such a detailed and extensive study in 2008, gathering data at 10,635 bus stop events, witnessing 22,980 boardings (Figure 5(a)). Specially trained surveyors were assigned to ride specific trips using a stratified random sample covering 24/7 operations. Surveyors, seated near the bus's front door, classified each boarding as one of 13 mutually exclusive categories. 421 discrete trips were observed between May 13 and November 29, 2008. Ridership where transaction data was lost (farebox malfunction and/or memory failure) was separately estimated using AFC data.

Non-farebox passengers are expressed as a fraction of AFC counts (to provide an expansion adjustment factor), computed as $2{,}477 \div 20{,}503 = 0.121$ (alternatively $10.8\% \div 89.2\% = 12.1\%$). This factor can be updated periodically with surveys as deemed necessary. Recent research by the San Francisco Municipal Railway (*39*) supports use of this "evasion" factor.

Farebox Data Transmission Error Adjustment

NYCT's AFC group produces bi-monthly reports estimating farebox data transmission error impacts. Ratio of farebox-registered cash to coins physically received and counted at the Consolidated Revenue Facility (colloquially, "*Money Room number*") is reported by depot. When revenues received exceed passenger registrations, reasons generally are farebox undercounting or improper data upload; passenger overpayments are very rare. Money Room number is the adjustment factor to compensate for data transmission errors. Systemwide weighted average (by each depot's monthly trips operated) is used.

SBS Correction Factors

NYCT obtains **+selectbus**service ridership directly by "probing" MFCs. Transaction counts from MFCs and CFCs are available from June 29, 2008, when POP was launched. Monthly raw count must be corrected for non-receipt passengers and data transmission errors.

NYCT conducted a BX12 payment study in May-June 2009, observing 1,881 fare payment transactions for 2,278 boardings, leaving 397 boardings (or 17.4%) unaccounted for (Figure 5(b)). However, 3.4% of non-receipt boardings were by Children Under 44" (requiring no receipts when accompanied by paying adults), and about 1% were boardings by railroad universal farecard "UniTickets" holders (estimated from sales and usage data). This 4.4% combined total was actually valid boardings that legitimately did not require receipts. Thus, "unaccountable" boardings account for 17.4% – 4.4% = 13.0% of ridership (*33*). Non-receipt adjustment expansion factor is thus 13.0% ÷ (1 – 13.0%) = 14.9%.

Using outlier analysis and manual data correction (substituting averages where values are missing), NYCT's ridership unit produced MFC transmission error estimations. The ratio of monthly registered to estimated swipes (based on MFC reporting failure patterns) is used.

Total Adjustments

Adjustment categories are mutually exclusive, thus factors are simply added together to obtain an overall adjustment factor (Figure 5(c)). Factors are expressed as percentages of raw transaction file counts. SBS counts are separately added manually in spreadsheets. Same adjustment factors are applied to both unlinked-trip and passenger-mile statistics.

PARALLEL TESTING

Although each assumption was verified using statistical analyses of historical data, direct comparison of results to related data is required to test overall algorithm performance. NYCT chose a three-pronged approach for this validation.

Comparison with Route Average Trip Lengths from Ridecheck Surveys

"Average miles travelled" by route (program output) is compared to known values computed from recent SR surveys. Although not directly comparable because survey and AFC dates are different, if route and population patterns haven't changed, trip lengths should be a route characteristic that varies relatively little. Even though ridechecks are not 100% accurate and not available for all routes, it represents the best NYCT data for average passenger miles.

Comparison cannot be made directly with NTD Section 15 data, because NYCT's NTD sample picks one specific route only a few times per year (700 trips sampled from universe of 244 routes), and thus cannot determine average trip lengths by route.

Figure 6(b) shows AFC trip lengths for one single day (vertical axis) versus SR survey values (horizontal axis) for all NYCT local routes. R-squared of 0.75 and nearly 1:1 slope indicates good correlation between results obtained from different computation methods and data sources. Even scattering on both sides suggests neither AFC nor SR contain assumptions that systematically skew passenger miles estimations for specific route types.

Consistency of Estimated Trip Lengths

Average trip lengths by route for different dates are compared against each other. As a route characteristic, it should be fairly stable, even if passenger counts vary. Weekday AFC trip lengths in January 2010 for all routes in Manhattan (Figure 6(c)) and the Bronx (Figure 6(d)) appear fairly stable compared to the monthly averages, with coefficient of variation (standard deviation divided by the mean) generally remaining <10%. Routes having lower average miles per passenger tend to have lower day-to-day variability. Routes with lower ridership have more variability, due to greater influence of one non-routine trip pattern on a given day's results.

Some exceptions are noted. Limited/local routes are close mutual substitutes. When separated, symmetric daily activity pattern assumption may be partially violated. Passengers traveling northbound on M5 Limited may later return south on M5 Local. This traffic imbalance, coupled with lower service spans on some limited routes (therefore lower ridership and further exacerbates imbalance), creates larger variances seen sets of substitutable routes like M2/M2LTD, M4/M4LTD, M5/M5LTD, BX1/BX1LTD/BX2/BX2LTD, BX25/BX26.

M60, a Manhattan local bus serving LaGuardia Airport in Queens, also has high variance. This route loops around all airport terminals before returning to Manhattan, resulting in mismatching eastbound and westbound paths with different lengths. Algorithms for handling this situation were proposed elsewhere (*40*), but were not used in here due to its complexity. Airport passengers often make multi-day trips and therefore producing significant imbalances in ridership, especially on Mondays and Fridays. Both phenomena partially violate the symmetric daily activity pattern assumption, resulting in higher variance.

Direct Comparison with Annual Section 15 Sample

Traditional Section 15 data estimates annual passenger miles through a 700-trip annual sample, producing ±10% error at 95% confidence. When one year's AFC data has been processed and summarized, it can be compared to results from manual sampling procedures. The difference should be no greater than 10%. AFC passenger miles remain within 10% of traditional survey estimates (Figure 6(e)).

OBTAINING APPROVAL

To obtain necessary FTA approvals, NYCT summarized statistics from parallel testing together with benefits of adopting 100% AFC data for passenger mile derivation:

1. Uses 100% MetroCard data, removing sampling needs.
2. Eliminates paper data collection, manual data entry, and checking; provides higher accuracy and consistency by eliminating "human" elements.
3. Eliminates a discrepancy consistently observed between revenue AFC ridership and surveyor counts.
4. Eliminates attendance issues that require survey rescheduling.
5. Data is accessible electronically, making it easier to retrieve for audits.
6. Since sampling is not used, monthly fluctuations from sampling errors will no longer occur.

FTA approved NYCT's conversion to AFC data for bus passenger-miles reporting on December 9, 2009. At FTA request, NYCT submitted appropriate additional backup documentation:

- Overall presentation of AFC conversion project
- Data flow diagram
- Selected manual survey raw data for 2007 (Figure 6(a))
- Monthly AFC results and selected daily data for 2007
- Descriptions of AFC adjustment factors and passenger-mile algorithm

CONCLUSIONS

As previous literature indicates, developing automated analyses of AFC data streams can be a minefield full of pitfalls resulting from defective data, incorrect assumptions, poor design, and implementation issues. By applying lessons learned in NYCT's subway AFC passenger miles reporting system development (*5*), this project avoided some common processing-time and regulatory pitfalls.

Although required engineering assumptions in this fault tolerant algorithm may seem egregious at the outset, NYCT staff diligently tested and proved each hypothesis by conducting elaborate surveys, analyzing and re-utilizing readily available service planning route profile data, and running extensive parallel tests. At each stage in development, computationally efficient methods were utilized. Processes requiring extensive coding effort, processor time, hard-to-obtain data, or manual intervention were rejected. Fault tolerant alternatives and assumptions providing higher degrees of automation without introducing unacceptable estimation errors were sought. Maximum use were made of available data sources to automatically and periodically update lookup tables and correction factors. This effort paid off as regulatory approvals were vastly streamlined compared with subway passenger-miles, because appropriate test results and documentation were readily available.

As more properties embrace OD farebox data collection or AVL/APC systems, analytical algorithm needs for non-geographic data will decline. Nonetheless, even best geocoded smartcard or AVL/APC data streams can have data quality issues and continues to require interpretative assumptions, failure detection algorithms, and data correction factors for human elements like fare evasion or blocked sensors. We hope methods and ideas demonstrated in this algorithm, although specifically developed for NYCT, can be helpful and applicable as other data analysts delve into newer data streams.

ACKNOWLEDGEMENTS

Authors gratefully acknowledge support and assistance of the following during algorithm development: Gary Delorme, Federal Transit Administration; Sergio Maia, National Transit Database; Jane Bailey, Savantage. Authors thank NYCT colleagues Svetlana Rudenko, Michael Kelly, Santosh Kumar, Minh Tran, Louis Balfan, Tewfik Berri, Steve Aievoli, Ted Wang, Anthony Cramer, H. Robert Menhard (SDR), Miguel Garcia, Daniel Rodriguez (AFC), Qifeng Zeng, Stella Levin, Jennifer Cohen, William Amarosa, and Lawrence R. Hirsch (OMB). Authors also thank the reviewers for their helpful comments. Responsibility for errors or omissions remains with the authors. Opinions expressed are the authors' and do not necessarily reflect official policies of Metropolitan Transportation Authority or New York City Transit.

REFERENCES

(1) Federal Transit Administration. *2007 Annual NTD Reporting Manual* and *Circular 2710.4A: Sampling Techniques for Obtaining Fixed Route Bus (MB) Operating Data Required Under the Section 15 Reporting System*. Accessed via the National Transit Database website http://www.ntdprogram.gov/ on May 2, 2008.

(2) Zhao, J., 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.

(3) Rahbee, Adam B. Farecard Passenger Flow Model at Chicago Transit Authority, Illinois. In *Transportation Research Record 2072*, TRB, National Research Council, Washington, D.C., 2008.

(4) Barry, James J., R. Newhouser, A. Rahbee, and S. Sayeda. Origin and Destination Estimation in New York City Using Automated Fare System Data. In *Transportation Research Record: Journal of the Transportation Research Board*, Transportation Research Board of the National Academies, 2002.

(5) Reddy, Alla, A. Lu, S. Kumar, V. Bashmakov, and S. Rudenko. Application of Entry-Only Automated Fare Collection (AFC) System Data to Infer Ridership, Rider Destinations, Unlinked Trips, and Passenger Miles. TRB Paper # 09-0809. In *Transportation Research Record 2110*, Transportation Research Board of the National Academies, 2009.

(6) 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.

(7) 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.

(8) Zureiqat, Hazem, N.H.M. Wilson, and J. Attanucci. Fare Policy Analysis for Public Transport: A Discrete-Continuous Modeling Approach Using Panel Data, TRB Paper #09-1591. In *Proceedings of the 88th Annual Meeting of the Transportation Research Board*. CD-ROM. Transportation Research Board of the National Academies, 2009.

(9) Frumin, Michael. *Automatic Data for Applied Railway Management: Passenger Demand, Service Quality Measurement, and Tactical Planning on the London Overground Network*. Thesis, Massachusetts Institute of Technology, Cambridge, Mass., 2010.

(10) Gordillo, Fabio. *The Value of Automated Fare Collection Data for Transit Planning: An Example of Rail Transit OD Matrix Estimation*. Technology and Policy Program Thesis, Massachusetts Institute of Technology, Cambridge, Mass., 2006. Retrieved from *http://dspace.mit.edu/handle/1721.1/38570* on June 27, 2010.

(11) 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.

(12) 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, Pages 35-48, January, 1998.

(14) Liu, Jianfeng, J.H. Li, F. Chen, Y.Q. Zhou. Review on Station-to-Station OD Matrix Estimation Model and Algorithm for Urban Rail Transit. Presented at *Second International Conference on Computer Modeling and Simulation*, Vol. 3, pp.149-153, Sanya, China, January, 2010.

(15) Pelleter, Marie-Pierre, M. Treeplanner, and C. Moroncy. *Smart Card Data in Public Transit Planning: A Review*. Report CIRRELT-2009-46. Retrieved from *https://www.cirrelt.ca/DocumentsTravail/CIRRELT-2009-46.pdf* on June 27, 2010.

(16) 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.

(17) 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.

(18) Mulqueeny, James Jr., S.J. LaBelle, R.T. Patronsky, and J. Simonetti. What to Do With Your New Electronic Farebox Data. Presented at *73rd Annual Meeting of the Transportation Research Board*. Transportation Research Board of the National Academies, 1994.

(19) Furth, Peter G. Integration of Fareboxes with Other Electronic Devices on Transit Vehicles. In *Transportation Research Record 1557*, p. 21-17, Transportation Research Board of the National Academies, 1996.

(20) Cui, Alex. *Bus Passenger Origin-Destination Matrix Estimation Using Automated Data Collection Systems*. Thesis, Massachusetts Institute of Technology, Cambridge, Mass., 2006.

(21) Taiwan Smart Card Corporation. *What is Taiwan Tong?* (什麼是「台灣通」？) Retrieved from http://www.twnsc.com.tw/node/5 on July 2, 2010.

(22) Gilligan, James M. New Jersey Transit BRT Initiatives: Go Bus28 and Reuse of a Right-of-Way in Union County. Presented at *APTA Multimodal Operations Planning Conference*, New York City, N.Y., July 26-28, 2010.

(23) Furth, Peter G. Innovative Sampling Plans for Estimating Transit Passenger Kilometers. In *Transportation Research Record 1618*, TRB, National Research Council, Washington, D.C., 1998, pp. 87–95.

(24) Navick, David S. and P.G. Furth. Estimating Passenger Miles, Origin-Destination Patterns, and Loads with Location-Stamped Farebox Data. In *Transportation Research Record 1799*, Paper No. 02-2466, TRB, National Research Council, Washington, D.C., 2002, pp. 107–113.

(25) Fare Demonstration Project. In *Headlights*, Magazine of Electric Railroaders' Association, Inc., New York, N.Y., August, 1964.

(26) Illinois Central Railroad. *Illinois Central's Gamble at Chicago: Private Breakthrough for a Public Cause*. Chicago, Ill., circa 1968.

(27) Buneman, Kevin. Automated and Passenger-Based Transit Performance Measures. In *Transportation Research Record 992*, pp. 23-28, Transportation Research Board of the National Academies, 1984.

(28) Miller, Luther S. AFC: A Fare Deal for All – Mass Transit Automatic Fare Collection Systems. In *Railway Age*, Issue 5, Volume 195, May, 1994.

(29) Vigrass, J. William. *The Lindenwold (New Jersey to Philadelphia) Hi-Speed Line: The First Twenty Years of the Port Authority Transit Corporation (PATCO)*. West Jersey Chapter, National Railway Historical Society, Cherry Hill, N.J., 1990.

(30) Ford, Roger. Technology Update: Ticket Issuing and Revenue Control. In *Modern Railways*, Volume 41, Pages 256-257, May, 1984.

(31) Young, David. The Business of Fare Collection. In *Mass Transit Magazine*, September, 1977.

(32) Urban Mass Transportation Administration. Sampling Procedures for Obtaining Fixed Route Bus Operating Data Required Under the Section 15 Reporting System, UMTA Circular 2710.1, Washington, D.C., February 22, 1978.

(33) Donohue, Pete. You Won't Find a Free Ride Here... MTA Inspectors Keep Bronx's BX12 Fare-Beaters in Check. In *New York Daily News*, June 17, 2010. Retrieved from *http://www.nydailynews.com/ny_local/bronx/2010/06/17/2010-06-17_dont_do_the_crime__or_you_may_pay_farebeat_fine_on_bx12.html* on July 10, 2010.

(34) MTA New York City Transit. 2010 NYC Transit Service Reductions. New York, N.Y., January 27, 2010. Retrieved from *http://mta.info/mta/news/books/pdf/100125_1031_service2010-nyct.pdf* on November 9, 2010.

(35) Reddy, Alla, J. Kuhls, and A. Lu. Measuring and Controlling Subway Fare Evasion: Improving Safety and Security at New York City Transit Authority. In Press. TRB Paper #11-2016, Submitted for Publication at the *Transportation Research Board 90th Annual Meeting*, Washington, D.C., January 23-27, 2011.

(36) New York State Legislature. New York State Penal Law, §120.05 Assault in the Second Degree, Subdivision 11. Retrieved from *http://public.leginfo.state.ny.us/LAWSSEAF.cgi?QUERYTYPE=LAWS+&QUERYDATA=$$PEN120.05$$@TXPEN0120.05* on July 10, 2010.

(37) Donohue, Pete. Brooklyn Man Indicted for Pummeling Bus Driver, Could Get Up to Seven Years. In *New York Daily News*, July 3, 2010. Retrieved from *http://www.nydailynews.com/news/ny_crime/2010/07/03/2010-07-03_goon_slapped_with_rap_in_busdriver_attack.html* on July 10, 2010.

(38) McFadden, Robert D. Police Say Killer Paid No Fare and Attacked After Being Denied a Transfer. In *New York Times*, New York Section, Page A27, December 2, 2008.

(39) Lee, Jason. "Passes and Transfers, Please": Uncovering Proof-of-Payment Patterns to Reduce Fare Evasion. In Press. Submitted for Publication at the *Transportation Research Board 90th Annual Meeting*, Washington, D.C., January 23-27, 2011.

(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 1927*, pp. 207-216, Transportation Research Board of the National Academies, 2005.

LIST OF FIGURES

FIGURE 1 Transaction file layout and theoretical analysis of algorithm assumptions: (a) NYCT's MetroCard AFC farebox EU65 transaction file rounds the transaction time to the nearest 6 minutes; (b) Errors in transaction time rounding cancels out in aggregate when the difference between embarkation and disembarkation time (i.e. onboard travel time) is computed; (c) Potential passenger-mile estimation impacts of the constant average speed assumption on a typical local route; (d) Potential underestimation on an inbound suburb-to-city express commuter bus.

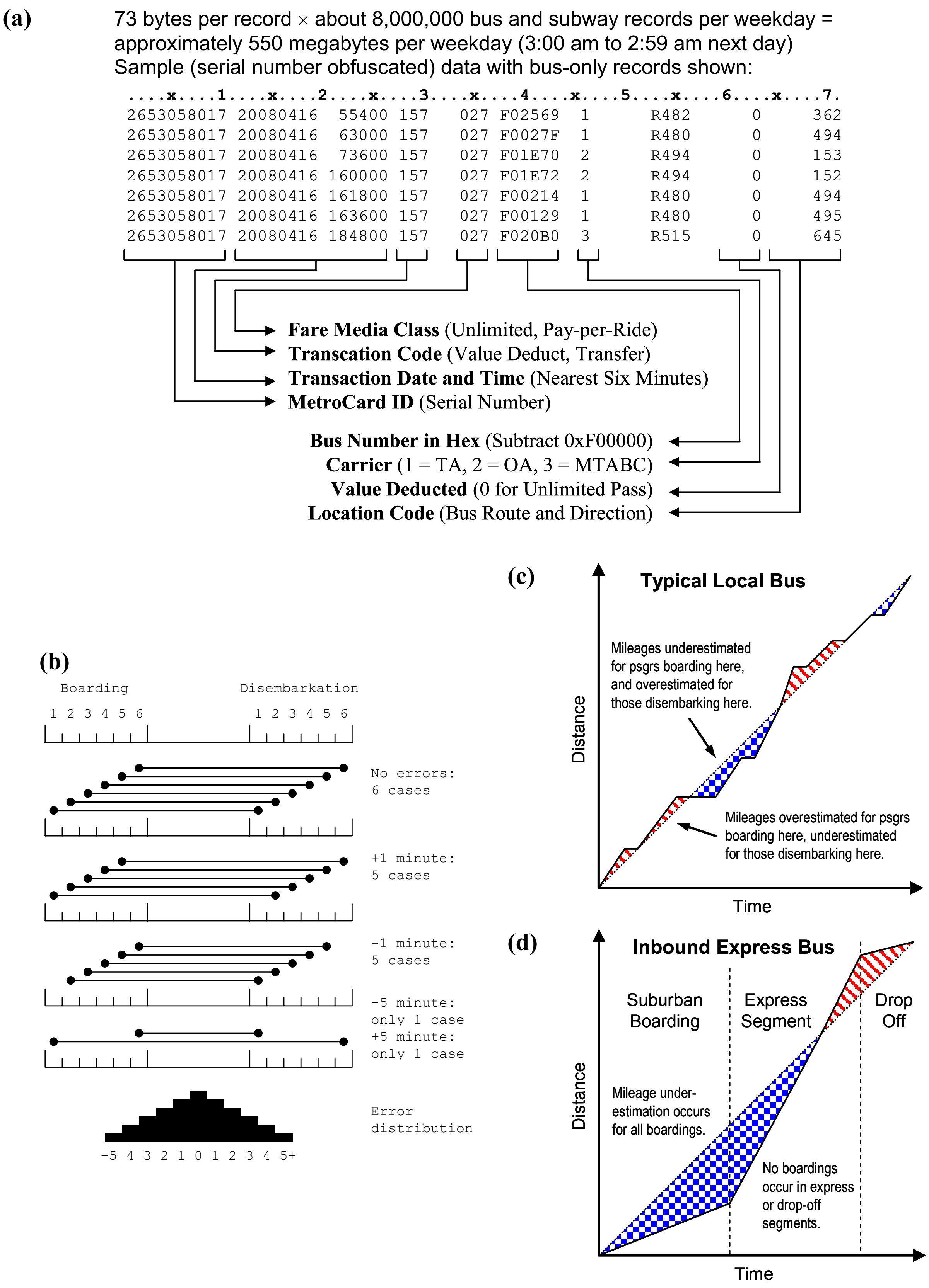
FIGURE 2 Data from NYCT's Surface Ridecheck program is used to validate the symmetric daily activity pattern assumption: (a) BX55 activity data shows good correlation between Ons and reverse direction Offs; (b) Symmetry assumption holds on the BX55 cumulative activity curves; (c) Dogbone shaped S60 is not at all symmetrical, resulting in low R-squared; (d) Baseball-cap shaped B51 has a loop in Manhattan and is limited-stop in one direction only, also violating symmetry assumptions; (e) Other routes show high daily cumulative On/Off correlation R-squared.

FIGURE 3 Graphical illustration of NYCT's AFC bus passenger mileage algorithm derivation.

FIGURE 4 AFC passenger mileage program flowchart and time-distance lookup file.

FIGURE 5 Derivation of correction factors used to adjust data for non-AFC fares: (a) Non-farebox passenger survey; (b) POP leakage survey; (c) All correction factors.

FIGURE 6 (a) Backup submittal to the FTA showing an example of manual survey data for Q58 on June 1, 2007; (b) Results from independent approaches to testing the AFC mileage algorithm: AFC estimated mileage versus Surface Ridecheck estimated mileage correlation test; (c) Internal consistency test (Manhattan); (d) Internal consistency test (Bronx); (e) Comparison with traditional Section 15 NTD sample (2007-2009).

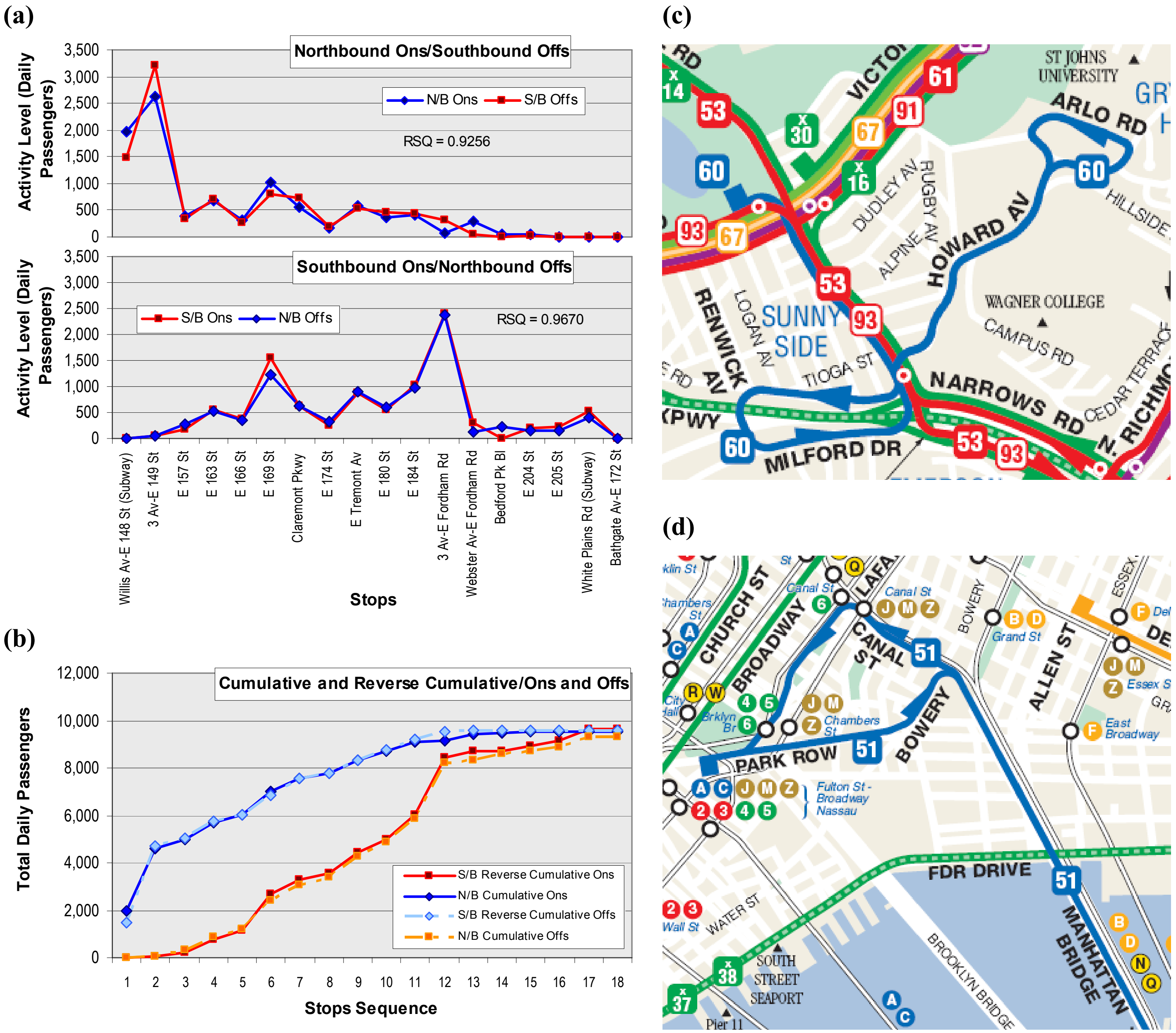


(a)

73 bytes per record × about 8,000,000 bus and subway records per weekday = approximately 550 megabytes per weekday (3:00 am to 2:59 am next day)
Sample (serial number obfuscated) data with bus-only records shown:

```
....x....1....x....2....x....3....x....4....x....5....x....6....x....7.
2653058017 20080416  55400 157   027 F02569  1      R482      0     362
2653058017 20080416  63000 157   027 F0027F  1      R480      0     494
2653058017 20080416  73600 157   027 F01E70  2      R494      0     153
2653058017 20080416 160000 157   027 F01E72  2      R494      0     152
2653058017 20080416 161800 157   027 F00214  1      R480      0     494
2653058017 20080416 163600 157   027 F00129  1      R480      0     495
2653058017 20080416 184800 157   027 F020B0  3      R515      0     645
```

FIGURE 1

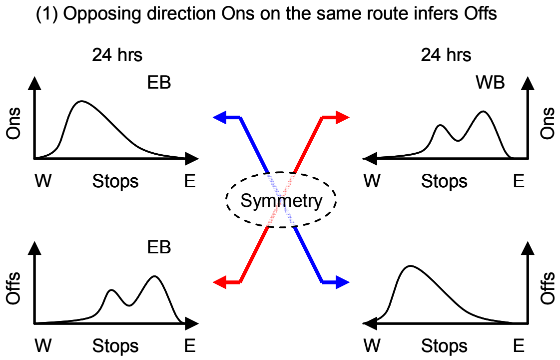


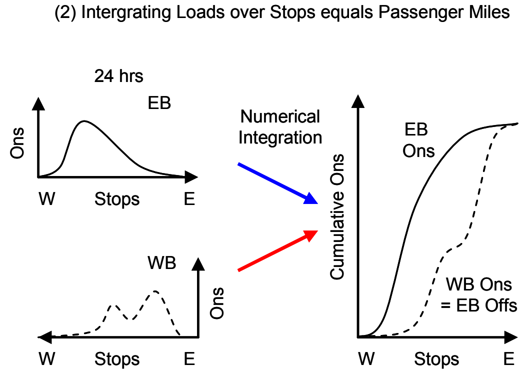
(e)

Route	Type	Boarding Direction	R-squared	Boarding Direction	R-squared	Route	Type	Boarding Direction	R-squared	Boarding Direction	R-squared
BX12	Crosstown-feeder hybrid	EB	0.992	WB	0.996	Q46	Suburban trunk	EB	0.972	WB	0.999
BX29	Double-ended feeder	EB	0.983	WB	0.981	Q74	Campus feeder/circulator	NB	0.904	SB	0.962
BX55	Limited stop trunk	NB	0.997	SB	0.999	Q79	Suburban crosstown	NB	0.982	SB	0.981
B31	Low density feeder	NB	0.992	SB	0.994	Q83	Suburban feeder	EB	0.995	WB	0.996
B42	Subway extension	NB	0.978	SB	0.985	S4090	Suburban trunk	EB	0.994	WB	0.997
B51	Bridge route	NB	0.833	SB	0.907	S52	Suburban feeder	NB	0.973	SB	0.986
M15	High volume trunk route	NB	0.987	SB	0.993	S60	Neighbourhood circulator	NB	0.698	SB	0.833
M18	Branch shuttle	NB	0.924	SB	0.978	S79	Suburban trunk	NB	0.999	SB	0.986

FIGURE 2

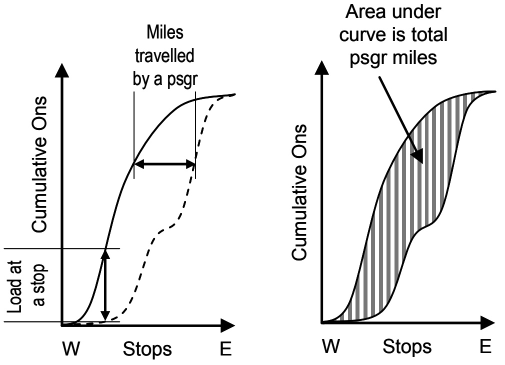
(a) The daily summary AFC observable Ons-by-Stop curves (in both directions) are reflected to produce the non-observable Offs-by-Stop curves, based on the symmetry assumption (right).



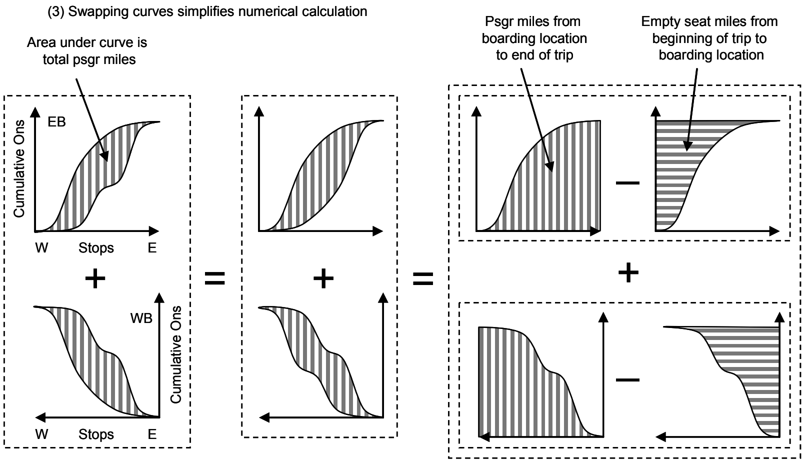


(b) The Ons-by-Stop and Offs-by-Stop curves are then integrated (i.e. cumulative values are computed) together to produce the cumulative On-Off (Load) by-Stop curve (left).

(c) The cumulative values clearly show graphically the Loads at a given stop, and the Mileage incurred by a single passenger along the route (right). The mileage incurred by all passengers is thus the sum of the area under the cumulative curve (far right).

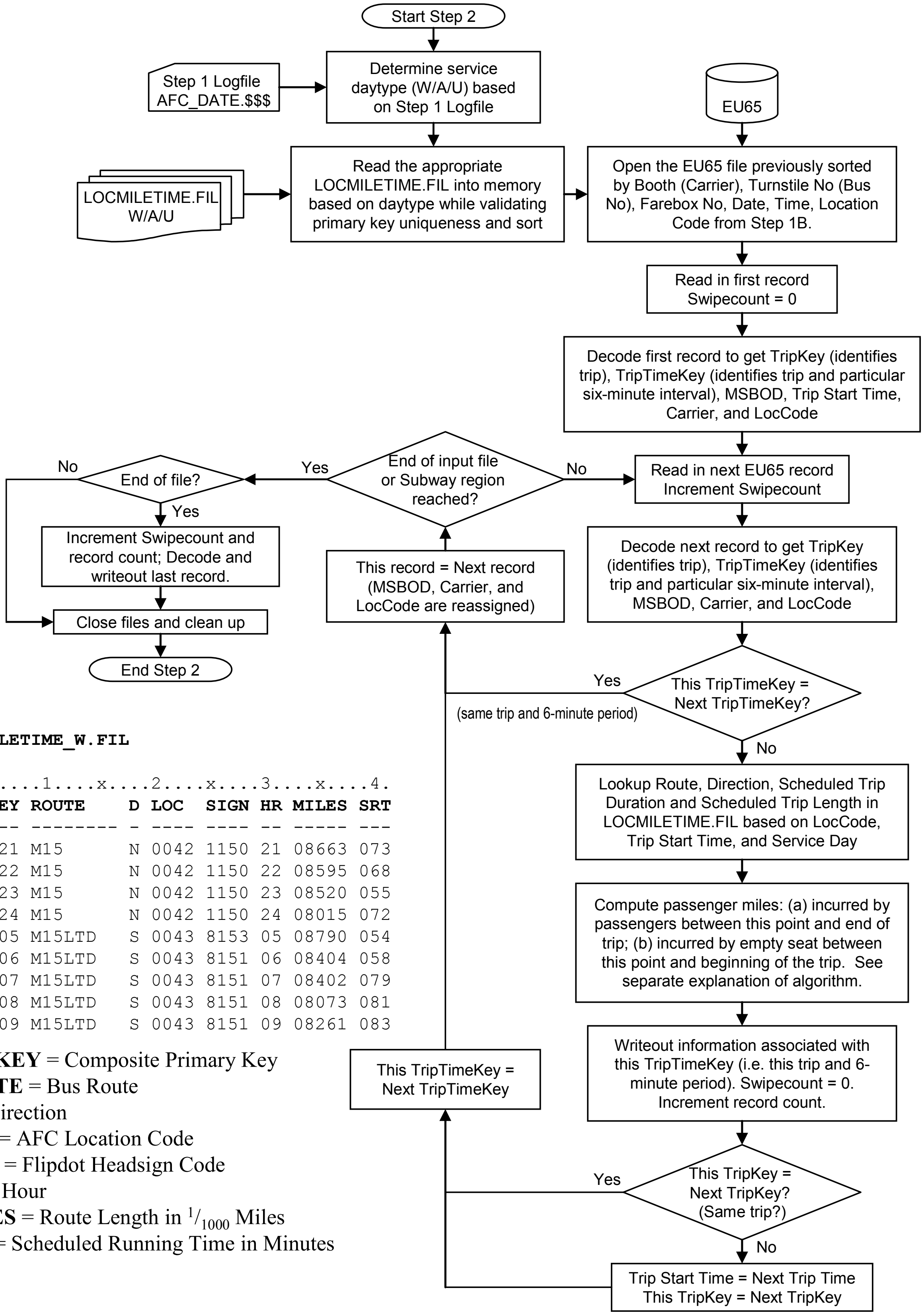


(d) Finally, to simplify computation of area for the irregular shape between the cumulative Ons and Offs curves, the curves are swapped in such a way that preserves the total area but increases the symmetry (below left). The final integral to be evaluated turns out to be the sum of all passenger miles from boarding location to the end of the trip, subtract all empty seat miles from the beginning of the trip to boarding location. These geometric transformations greatly simplify computation as the program would need to consider only two variables: the length of a trip on the route, and the location of each boarding relative to the route's beginning and end.



The four-step integral shown above is the calculation computed in the Step 2 algorithm.

FIGURE 3



LOCMILETIME_W.FIL

```
....x....1....x....2....x....3....x....4.
PRI_KEY ROUTE    D LOC  SIGN HR MILES SRT
------- -------- - ---- ---- -- ----- ---
0042W21 M15      N 0042 1150 21 08663 073
0042W22 M15      N 0042 1150 22 08595 068
0042W23 M15      N 0042 1150 23 08520 055
0042W24 M15      N 0042 1150 24 08015 072
0043W05 M15LTD   S 0043 8153 05 08790 054
0043W06 M15LTD   S 0043 8151 06 08404 058
0043W07 M15LTD   S 0043 8151 07 08402 079
0043W08 M15LTD   S 0043 8151 08 08073 081
0043W09 M15LTD   S 0043 8151 09 08261 083
```

PRI_KEY = Composite Primary Key
ROUTE = Bus Route
D = Direction
LOC = AFC Location Code
SIGN = Flipdot Headsign Code
HR = Hour
MILES = Route Length in $^{1}/_{1000}$ Miles
SRT = Scheduled Running Time in Minutes

FIGURE 4

(a) Non-Farebox Passenger Survey Results

Category	Passengers Observed	Percentage	Notes
Valid MetroCard Fare	19,630	85.4%	1. Transfers between Ⓛ Train and **B42** Bus to Canarsie Beach at Canarsie Subway station occurs within fare control. Transferring passengers are not required to swipe upon boarding, thus no AFC record is generated.
Valid Split Fare	294	1.3%	
Invalid MetroCard Fare	449	2.0%	
Invalid Split Fare	130	0.6%	
Subtotal AFC Counted Passengers	**20,503**	**89.2%**	
Front Door Non-Paying Passenger	644	2.8%	
Rear Door Non-Paying Passenger	110	0.5%	
Child Over 44" Travelling Without Fare	413	1.8%	
Paper Ticket	47	0.2%	2. BX12 Select Bus utilizes a proof-of-payment (POP) fare collection system. POP receipts are not valid on board BX12 local buses, but are occasionally accepted by drivers.
Child Under 44" Travelling Without Fare	557	2.4%	
Flash Pass, Uniform, or Official Badge	268	1.2%	
Wheelchair Travelling Without Fare	44	0.2%	
Seamless Transfers[1]	31	0.1%	
Select Bus Receipt on BX12 Local[2]	27	0.1%	
% of Trips with Farebox Malfunction	1.40%		3. Estimated passenger boardings during farebox malfunction based on 1.4% evenly distributed.
Broken Farebox Psgr Boardings[3]	336	1.5%	
Subtotal Unaccountable Passengers	**2,477**	**10.8%**	
Total Passengers	22,980	100.0%	

(b) BX12 Select Bus Service Non-Receipt Boarding Survey Results

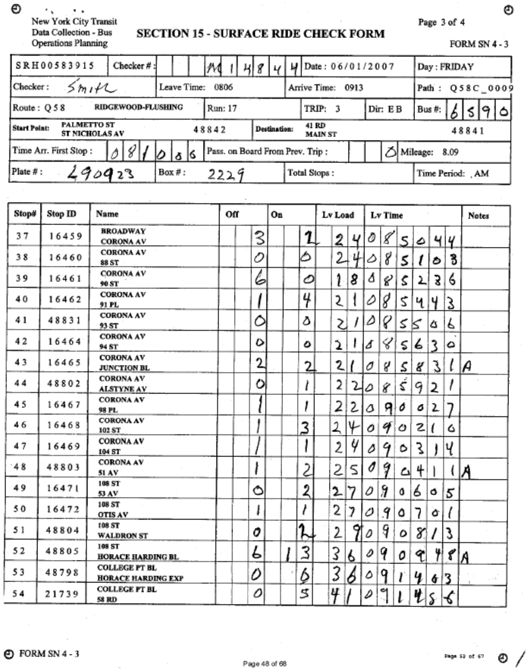
Category	Total Count	Rate	Notes
Fares Paid – MetroCard Validators	1,847		4. POP receipts are occasionally redeemed on BX12 local buses.
Fares Paid – Coin Fare Collectors	41		
Paid Passengers Boarding Local Service (Leakage)[4]	-7		
Passenger Registrations Observed	**1,881**	—	5. Includes Children under 44", and passengers with UniTickets (commuter railroad universal fare media, accepted only on feeder buses).
Front Door Entries	1,383		
Rear Door Entries	895		
Passenger Boardings Observed	**2,278**	—	
Boardings minus Registrations	397	17.4% ±2%	
Exempt (non-Receipt) Adjustment[5]	—	4.4%	
Rate of Unaccountable Boardings	—	**13.0% ±2%**	

(c) All Required Correction Factors to AFC Data

Service	Bus Passengers	Description	Factor Req'd?	Factor Value
Standard Bus	In EU65 MetroCard Transaction File	Base "raw" passenger boarding data from the MetroCard AFC system.	No	—
	Cash Passengers	Passengers not using electronic fare media.	Yes	15.4%
	Non-Farebox Passengers	Passengers not interacting with farebox due to broken farebox, fare evasion, paper tickets, flash passes, etc.	Yes	12.1%
	Farebox Data Transmission Errors	Passengers paying fare normally but data not in EU65 file due to farebox data transmission malfunction.	Yes	5.4%
	Total Adjustment Factor (Standard Bus)			**32.9%**
Select Bus Service	Revenue Passenger Data from Select Bus Service Fare Validation Machines	Number of receipts issued by the Proof-of-Payment wayside fare collection machines (Cash and MetroCard) on the BX12 Select Bus Service.	No	—
	Non-Receipt Passengers	Passengers without POP receipts due to fare evasion, paper tickets, and flash passes, etc.	Yes	14.9%
	Fare Validation Machine Data Transmission Errors	Passengers paying fare normally, but not recorded because of fare payment machine malfunction.	Yes	2.0%
	Total Adjustment Factor (Select Bus)			**16.9%**

FIGURE 5

(a)



New York City Transit
Data Collection - Bus
Operations Planning

SECTION 15 - SURFACE RIDE CHECK FORM

Page 3 of 4

FORM SN 4 - 3

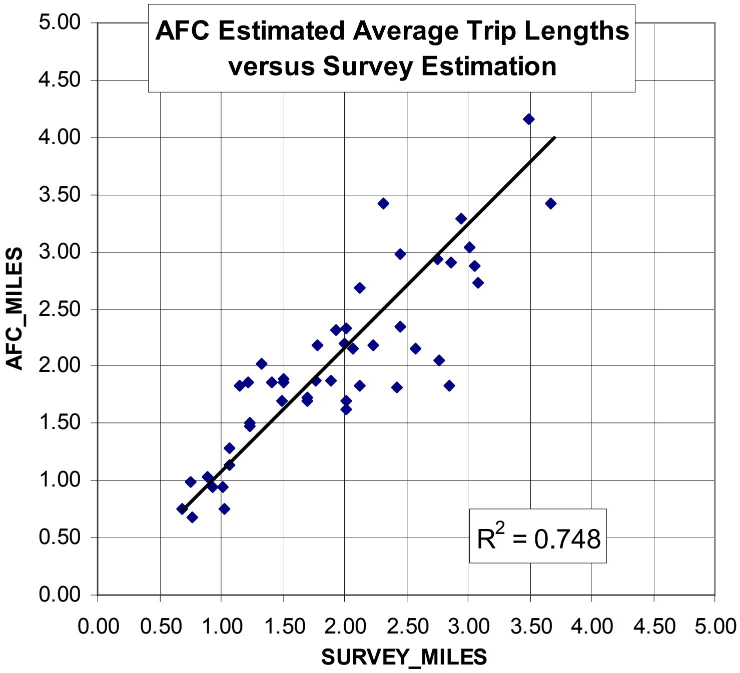
SRH00583915	Checker # : M14844	Date : 06/01/2007	Day : FRIDAY
Checker : Smith	Leave Time: 0806	Arrive Time: 0913	Path : Q58C_0009
Route : Q58 RIDGEWOOD-FLUSHING	Run: 17	TRIP: 3 Dir: EB	Bus #: 6590
Start Point: PALMETTO ST ST NICHOLAS AV 48842		Destination: 41 RD MAIN ST 48841	
Time Arr. First Stop : 081086	Pass. on Board From Prev. Trip : 0		Mileage: 8.09
Plate # : 490923	Box # : 2229	Total Stops :	Time Period: AM

Stop#	Stop ID	Name	Off	On	Lv Load	Lv Time	Notes
37	16459	BROADWAY CORONA AV	3	2	24	085044	
38	16460	CORONA AV 88 ST	0	0	24	085103	
39	16461	CORONA AV 90 ST	6	0	18	085236	
40	16462	CORONA AV 91 PL	1	4	21	085443	
41	48831	CORONA AV 93 ST	0	0	21	085506	
42	16464	CORONA AV 94 ST	0	0	21	085630	
43	16465	CORONA AV JUNCTION BL	2	2	21	085831	A
44	48802	CORONA AV ALSTYNE AV	0	1	22	085921	
45	16467	CORONA AV 98 PL	1	1	22	090027	
46	16468	CORONA AV 102 ST	1	3	24	090210	
47	16469	CORONA AV 104 ST	1	1	24	090314	
48	48803	CORONA AV 51 AV	1	2	25	090411	A
49	16471	108 ST 53 AV	0	2	27	090605	
50	16472	108 ST OTIS AV	1	1	27	090701	
51	48804	108 ST WALDRON ST	0	2	29	090813	
52	48805	108 ST HORACE HARDING BL	6	13	36	090948	A
53	48798	COLLEGE PT BL HORACE HARDING EXP	0	6	36	091463	
54	21739	COLLEGE PT BL 58 RD	0	5	41	091456	

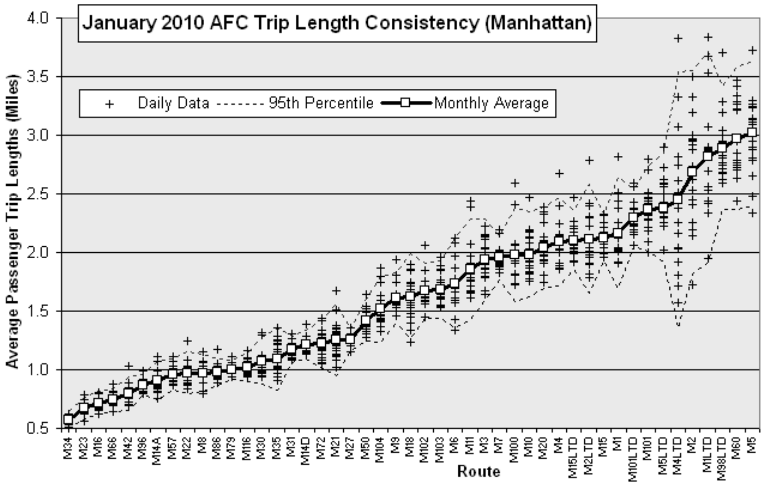
FORM SN 4 - 3

Page 48 of 68

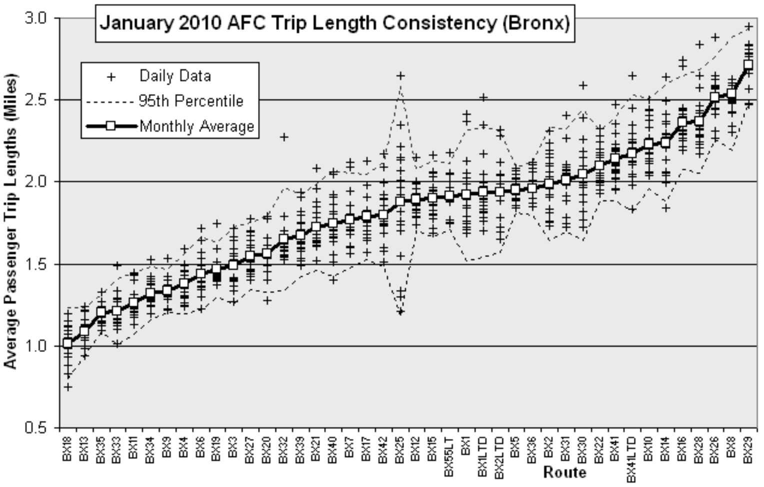
(b)



(c)



(d)



(e)

PARALLEL TEST

Surveyor's (Sample) versus MetroCard AFC Data (100%)

Passenger Miles & Unlinked Trips -- NTD Bus (MB): 2007 - 2009

	2007		**2008**		**2009**	
	Surveyor	**AFC Data**	**Surveyor**	**AFC Data**	**Surveyor**	**AFC Data**
REVENUE RIDERSHIP	738,039,531		746,977,406		726,433,247	
% Change (from previous year)			1.21%		-2.75%	
UNLINKED TRIPS	862,630,526	863,838,154	902,640,956	868,638,444	906,529,603	842,865,961
% Change - MetroCard AFC to Surveyor Data		0.14%		-3.77%		-7.02%
% Change (from previous year)			4.64%	0.56%	0.43%	-2.97%
PASSENGER MILES	1,812,108,125	1,887,689,579	1,861,302,947	1,892,676,473	1,865,303,339	1,838,901,551
% Change - MetroCard AFC to Surveyor Data		4.17%		1.69%		-1.42%
% Change (from previous year)			2.71%	0.27%	0.21%	-2.84%
TRIP LENGTH (miles)	2.10	2.19	2.06	2.18	2.06	2.18
% Change - MetroCard AFC to Surveyor Data		4.03%		5.67%		6.03%
% Change (from previous year)			-1.84%	-0.29%	-0.21%	0.13%

FIGURE 6 (Parts (c) and (d) Not Publication Resolution)

ENDS