TRB Paper Manuscript #09-0809

**Application of Entry-Only Automated Fare Collection (AFC) System Data to Infer Ridership, Rider Destinations, Unlinked Trips, and Passenger Miles**

*Alla Reddy*, Alex Lu, Santosh Kumar, Victor Bashmakov, and Svetlana Rudenko*
* Corresponding author

Alla Reddy
Senior Director, System Data & Research (SDR)
Operations Planning
New York City Transit
2 Broadway, A.17.92
New York, NY 10004
Tel: (646) 252-5662
Email: *Alla.Reddy@nyct.com*

**Word Count:** 249 (Abstract) + 6,609 (Text) + 2 * 250 (Figures) = 7,358 Words

## TABLE OF CONTENTS

**ABSTRACT**

All U.S. transit agencies receiving Federal Transit Administration (FTA) Urbanized Area Formula Program (UAFP) funding under §5307 (Section 15) report service consumption statistics (revenue passenger miles and unlinked trips) to the National Transit Database (NTD). Passenger miles is an incentive-based funding element that generates millions of dollars annually for New York City Transit (NYCT).

Originally, Section 15 random sample data was collected by surveyors gathering passenger destination information, followed by manual distance calculation based on judgment of likely travel paths. This method was costly, inefficient, inconsistent, and not always reproducible despite rigorous auditing and certification.

NYCT modernized this process by directly retrieving passenger origination information from the Automated Fare Collection (AFC) system, inferring destinations with a second swipe, and automating passenger-mile calculation using schedule-driven shortest-path algorithms. While using state-of-art data collection and computation methods, NYCT retained FTA-approved sampling methodology to maintain comparability of data going forward.

Success of automated data reporting is maximized by developing algorithms first using small datasets, followed by clearly documented parallel testing with full involvement of data consumers, including relevant regulatory authorities. Software development is iterative, and computation time should be monitored to ensure scalability.

Building on this work, NYCT is developing AFC-based methodologies to infer bus passenger origins and destinations, train loads, and adapting signal system data for routine operating performance monitoring. Reporting automation to a point where live data can be used for scheduling and service planning without modelling or special analyses will enable service to be monitored much more frequently and extensively.

## INTRODUCTION

MTA New York City Transit (NYCT) operated 5,253 subway cars and 3,866 buses in maximum service in 2006. The subway system ran 350 million car-miles, carried 1.9 billion unlinked trips (and 8.3 billion passenger-miles), required coordinated efforts of 32,000 employees, and generated $1.9 billion in farebox revenues – but cost $2.7 billion to operate. The bus network ran 120 million bus-miles, carried 930 million trips (1.9 billion passenger-miles), required the diligent work of 17,000 employees, generated $0.8 billion in farebox revenues, and cost $1.9 billion in operating expenses.

How does NYCT keep track of usage statistics on such a mammoth system? Moreover, how does NYCT continually generate funds to cover operating deficits, replace capital assets as they become life-expired, and perform maintenance to keep the system in good repair? As the statistics clearly demonstrate, farebox revenues do not nearly cover operating expenses, let alone generate needed capital funds to continually reinvest in a great system that serves 8 million people 24-hours, seven days a week. The widely-known physical decay of the 1970s shows the severe consequences of deferred maintenance for customers (*1*,*2*).

Fortunately for NYCT, funds are available from local, state, and federal governments to address these needs. This paper focuses on Federal Transit Administration's (FTA) Urbanized Area Formula Program (UAFP), also known as §5307 (colloquially, Section 15) funds. All U.S. transit agencies receiving funding under Section 15 (over 660 properties) report service statistics to U.S. Department of Transportation's (USDOT) National Transit Database (NTD) pursuant to §5335. NTD data is used to apportion over $4 billion in FTA funds annually (*3*). In 2007, NYCT received about $530 million under this program, approximately 13% of which was performance based. Two variables, unlinked trips and revenue passenger miles, represent a major annual data collection effort to satisfy NTD and FTA data needs – and a critical determinant in the incentive portion of Section 15 formula appropriation.

Section 15 funding is based on system characteristics and usage. The fixed infrastructure portion is based on cost parameters such as fixed guideway mileage and vehicle revenue miles. The incentive portion is based on usage indicators like passenger miles and unlinked trips. Collection and calculations of this data is very carefully monitored. Historically, these important indicators were obtained by dedicated teams of surveyors through manual observation and interactive passenger survey. The process is audited by external auditors, scrutinized by NTD staff, and certified as accurate by each transit agency's Chief Operating Officer. Additionally, data is reported monthly through NTD Internet Reporting for statistical purposes.

Since the early 1980s, Urban Mass Transportation Administration (UMTA) – predecessor to the FTA – has required a minimum confidence of 95%, and precision level of ± 10% (*3*). If not using an FTA-approved sampling technique, a qualified statistician must determine that sample meets FTA's prescribed standards.

**Using Electronic Data for Section 15 Reporting**

Passenger miles is total distance traveled by all subway passengers. Originally, surveyor data was analyzed by manual calculation of average distance travelled. Annual passenger miles is extrapolated from sample total by multiplying with annual total subway riders.

For NYCT's highly complex network of subway services featuring local and express trains on most mainlines, up to 8,200 daily train-starts, and around 80 transfer points, sampling or inferring unlinked trips is difficult. Calculation of passenger-miles is even more challenging, as multiple viable paths likely exist between two given stations, and path choice will depend on customer preference and service patterns at time of travel. This paper chronicles the story of how NYCT modernized this effort from a dated system of interviews, manual calculations, and analyst judgment of preferred travel paths – to an automated, efficient, accurate, and accountable data collection system utilizing a modern Automated Fare Collection (AFC) system, state-of-art shortest path algorithms, and automated destination determination.

Passenger origination information is directly retrieved from the AFC system. Destinations are inferred from a second subway swipe, and passenger-mile calculation is automated using schedule-driven algorithms. While using modern data collection and computation methods, NYCT retained FTA-approved sampling methodology to maintain and ensure comparability of data going forward. Interwoven in the modernization process is an appreciation of how disparate data systems originally designed for different purposes can be made to work harmoniously together to maximize the value of information collected.

While destination inference algorithms are well documented for multimodal origin-destination modelling (*8,9*), NYCT uses a modified algorithm designed to unambiguously capture entry and exit points for only the subway portion of trips.

## SECTION 15 DATA REQUIREMENTS AND SAMPLING METHODOLOGY

Section 15 Annual Report in 2007 consisted of six modules of 17 forms and three declarations that provide a certified summary of transit characteristics (*3*). Although data requirements have been revised essentially every year, two performance indicators remained remarkably consistent since the early 1980s: unlinked trips and revenue passenger miles. These indicators are currently reported on Form S-10, Lines 18 and 20. Passenger miles is also reported on Form FFA-10, Line 09. Together with other indicators such as route miles, revenue vehicle miles, operating expenses, and farebox revenues, these statistics enable social scientists to perform comparative studies of the nation's state and effectiveness of transit, and measure social benefits of Federal transit programs. While direct comparisons between metropolitan areas are practically impossible and can be somewhat misleading, these indicators are nevertheless important basic data that most governments and transportation carriers compile on an ongoing basis. Similar statistics and sampling methodology is used by scheduled passenger airlines, intercity bus operators (including New York's Chinatown buses), and even freight railroads (loads and revenue ton-miles) to determine the effectiveness of their service plans.

Much research was conducted in the early 1980s on how such extensive system usage data could be collected economically on a high frequency, low-cost system – where payment of a fare could

be simple as dropping coins into a farebox with no 'ticket' issued, and passengers do not declare their destinations to the carrier due to lack of 'zone fares'. For rapid transit systems, where free transfers are typically accomplished by walking from one platform to another without further payment of fares, estimation of unlinked trips and passenger miles is particularly problematic. The research culminated in the dissemination of a standard sampling methodology (*4*) that continues to be used today by subway systems without zone fares and not equipped with entry-exit AFC systems.

**The 1984 Sampling Methodology and NYCT Practice**

The 1984 sampling methodology calls for 1,200 sample responses in 30 clusters – requiring a minimum response rate of 40 passengers per hour. To prevent biases due to correlations between trip length and time-of-day, the methodology requires stratification of the 30 sample clusters by time periods, proportional to ridership within each time period.

NYCT obtained specific approval in 1989 from UMTA to use sampling techniques for passenger mile data, based on this methodology (*5*). The subway random sample approved by UMTA is a stratified sample consisting of seven strata with at least 50 control areas (station entrances) per strata. Thus, UMTA actually required 350 sample clusters on an annual basis, theoretically allowing 95% ± 10% error margins to be achieved on a time-period basis.

Subsequently, NYCT voluntarily increased sample sizes due to higher between-station variances. Strata known to have large variations are sampled more heavily. Clusters of random passengers are surveyed in 700 hour-long sessions each year (*4*). NYCT uses a spreadsheet random function to select control areas, hours, and dates to survey – subject to time-of-day stratification constraints.

For each response, the surveyor asked each passenger a single question: their last station on this trip. When applied consistently, this methodology yields average trip lengths and transfer rates, which is applied to systemwide total turnstile boardings to estimate total unlinked trips and passenger miles. The surveyor field procedure was: (*5*)

1. Record all turnstile entry readings at start time
2. For one hour, remain in paid area, interview and record final destinations for as many incoming passengers as possible
3. Count all non-fare paying passengers entering over and under turnstiles, and through agent operated gates
4. Record all turnstile entry readings at end time

When all forms were collected and returned, they were individually reviewed for common errors. The analyst certified proper execution of each survey by signature. To determine the mileage traveled by each responding passenger, analysts used subway maps, schedules, and distance table to manually determine where transfers took place, and summed mileages for all route segments traveled. Once all mileages and transfers were computed for all respondents in one assignment, the following formulae were used to estimate total passenger miles and unlinked trips:

$$\text{Entrants} = \sum \left( \text{Turnstile Readings at End} - \text{Turnstile Readings at Start} \right) + \text{Non-Turnstile Passengers}$$

$$\text{Average Miles for Respondents} = \frac{\Sigma \text{ (Miles Traveled) by Respondents}}{\text{(Count of Responses)}}$$

$$\text{Average Transfers for Respondents} = \frac{\Sigma \text{ (Transfers) by Respondents}}{\text{(Count of Responses)}}$$

$$\text{Total Passenger Miles for hour-long survey} = \left( \text{Average Miles for Respondents} \right) \times \left( \text{Entrants} \right)$$

$$\text{Total Unlinked Trips for hour-long survey} = \left( 1 + \left( \text{Average Transfers for Respondents} \right) \right) \times \left( \text{Entrants} \right)$$

When all assignments in the required sample were available, data was extrapolated to estimate total unlinked trips and passenger miles in the reporting period, using these formulae:

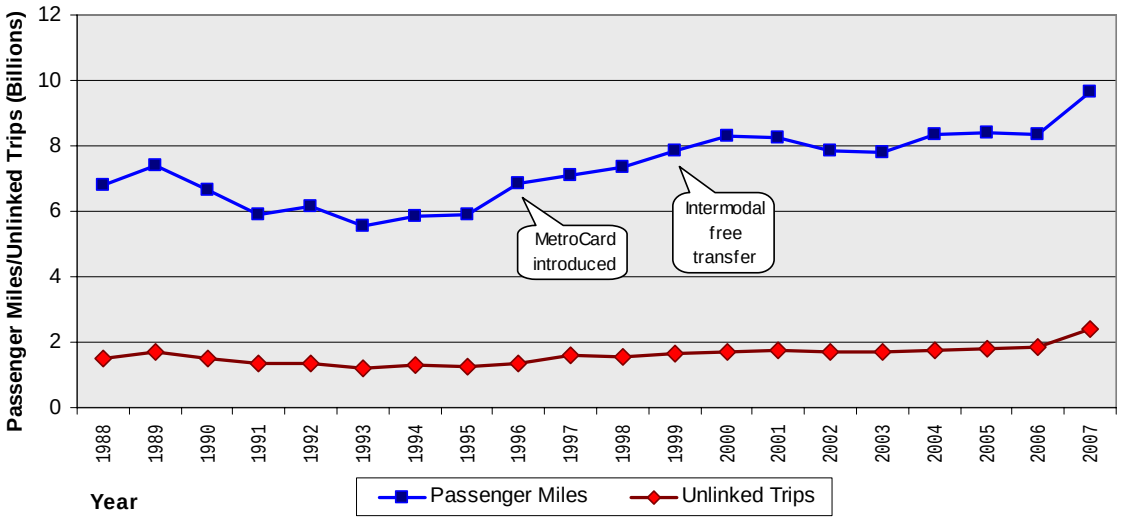
$$\text{Passenger Miles During Reporting Period} = \left( \text{Total Registrations During Reporting Period} \right) \times \frac{\Sigma \text{ (Passenger Miles) by hour-long surveys}}{\Sigma \text{ (Registrations) by hour-long surveys}}$$

$$\text{Unlinked Trips During Reporting Period} = \left( \text{Total Registrations During Reporting Period} \right) \times \frac{\Sigma \text{ (Unlinked Trips) by hour-long surveys}}{\Sigma \text{ (Registrations) by hour-long surveys}}$$

NYCT has used UMTA-approved sample with the 1984 methodology to calculate unlinked trips and passenger miles since 1988. During most of the 1990s, both passenger miles and unlinked trips grew steadily; however, average trip length has remained fairly constant, between 4 and 5 miles per trip. Between 2007 and the lowest point in 1993, passenger miles grew by 73% while unlinked trips grew by 103% (Figure 1). In 2007, annual subway (linked) ridership of 1.56 billion was the highest since 1951. It is remarkable that all ridership growth was accommodated on substantially the same subway infrastructure as the early 1990s, through increased train frequencies and increased utilization of underutilized parts of the system.

Fare control modernization and the introduction of MetroCard enabled tariff innovations based on new technology, including: volume discounts, free transfers between subway and buses, and periodic passes with arbitrary expiration dates. Substantial capital investment and a drop in crime citywide also contributed to the ridership rebound (*1*).

**Figure 1**
**NYCT Section 15 Reported Passenger Miles and Unlinked Trips, 1988-2007**




## POTENTIAL DATA COLLECTION ISSUES

During the twenty years of manual data collection, NYCT identified many practical problems. These issues ultimately led to development of support systems required to substitute AFC data for surveys. However, prior to using AFC data, NYCT developed a series of methods to combat these issues – some more successful than others.

### High Data Collection Costs

UMTA required a very rigorous and statistically random "probability sample" (*5*). Surveys are assigned to random times and locations, requiring 24-hour surveyor coverage, geographic range of 250 square miles, and travel times of up to 90 minutes from the central office location. 700 samples consumed six part-time surveyors to provide coverage for day-off relief, absenteeism, late starts and early departures. This represented an annual cost of about $150,000 in data collection alone. Additional costs were incurred in supervision and field spot-checking efforts.

In 1995 and 2000, internal audits revealed inconsistencies in data collection practices, and NYCT committed to FTA to improve supervision. In 2000, NYCT combined Section 15 data collection with another analytical unit, enabling higher surveyor utilization, more intensive supervision, reduced absenteeism, and lower costs. Nonetheless, data collection efforts continued to require many surveyors.

### Difficulty of Processing Large Passenger Volumes

NYCT's top 200 control areas process more than 1,000 passenger entries per hour during busiest times of the day. The busiest control area is R138 at 34 Street-Pennsylvania Station 1 2 3 (IRT Broadway-7 Avenue Line), processing on average 5,000 AFC swipes per hour during the morning peak. This poses a significant challenge for surveyors attempting to conduct a destination survey by gathering verbal responses. Assuming a continuous stream of passengers, each conversation may take 20 seconds to complete and record, resulting in maximum survey

throughput of 180 responses per hour. Because of practical limitations, many responses are missed due to sheer passenger volume. This factor alone resulted in only 28% of all sampled passengers being actually interviewed by surveyors in 2007.

### Difficulty of Gathering Responses

Passenger surveys in New York are particularly difficult for three major reasons: (a) Passengers do not take the time to respond to surveys during rush hours; (b) Passengers do not talk to strangers during off-hours due to concerns about personal safety; (c) Some subway riders have limited English proficiency and simply do not understand the surveyor. These three factors combine and result in a survey yield (responses ÷ passengers surveyed) of 77%.

The net impact of surveying difficulties result in a response rate (responses ÷ total entries) of 21% in 2007. Thus, substantial response bias exists in this dataset.

### Missed Assignments

Field observations are inherently difficult and attendance problems occur in survey workforces. Aside from unscheduled absences (48% of dropped assignments in 2007), other conditions can result in missed samples:

1. Scheduling difficulties – no surveyors available at time needed (17%);
2. Surveyors delayed en-route – service disruption (10%);
3. Surveyors completing form incorrectly, or observe incorrect location (3%).

To combat these issues, 'reserve' samples were added such that equivalent random sample data could be substituted for assignments missed. Typically 15%~25% of assignments were replaced with reserves.

### Data Interpretation Issues

Apart from the common difficulty of interpreting surveyors' handwriting, an issue specific to NYCT is the presence of very similar or identical subway station names. A passenger's response of '59 Street' may refer to the 4 Avenue Subway N R station in Brooklyn, the 8 Avenue IND A C B D station at Columbus Circle in Manhattan, the 4 5 6 station on the Lexington Avenue (East Side) IRT, or 59 Street-5 Avenue N R W station under Central Park in Midtown. Two '36 Street' stations exist on the R, one in Brooklyn and one in Queens. '111 Street' may refer to the 7 station in the Corona neighborhood of Northern Queens, the J Z station in Richmond Hill (Eastern Queens), or the A station in Ozone Park (Southern Queens). These can have significant impacts on passenger miles: travel from Myrtle Avenue J M Z to 111 Street in Richmond Hill is a one-seat ride; traveling to 111 Street in Corona involves two transfers and many reasonable choices of paths. Where final destinations cannot be ascertained, analysts assumed no interline transfers, resulting in a systematic bias lowering unlinked trips and passenger miles.

Even though surveyors asked riders about their 'final destination', many riders responded with their first transfer point. Inbound L riders on the Canarsie elevated overwhelmingly reported on their destination as 'Broadway Junction', even though field observations showed that most riders

disembarking there transferred to Manhattan bound A C J Z trains. This phenomenon also contributed to under-reporting.

**Inconsistencies in Data Collection**

In 1995, Operations Planning found inconsistencies in passenger survey data. All data collected by those employees was discarded. Although additional data was available from other surveyors (due to three-hour minimum duration for survey assignments), the sample was declared invalid because consecutive hours at the same location was considered a 'non-probability sample'. This problem resulted in lost funding assessed against NYCT based on lowest passenger miles of the prior three years.

Following that incident, NYCT implemented new procedures for supervising surveyors. Additional 'spot checks' by undercover analysts were implemented, in addition to routine field supervisor visits. Even more 'reserve' samples were added to provide replacement 'probability samples' should inconsistencies arise. 700 reserve assignments were prepared in 2007; 430 were performed, of which 115 was used. While these safeguards improved reliability to an acceptable level, substantial additional staffing resources over and above the budget were required. Data collection under these constraints became quite inefficient.

## THE METROCARD AUTOMATED FARE COLLECTION (AFC) SYSTEM

NYCT introduced an AFC system, known as *MetroCard*, beginning in 1994. Initially piloted at Whitehall and Wall Street stations in Lower Manhattan, the MetroCard project introduced a paradigm shift to revenue processing and accounting, all the way from station booths to the Money Room.

Replacing mechanical turnstiles that accepted transit tokens ('good for one fare') vended from booths, a new generation of "smart" turnstiles designed to use electronic fare media and discourage fare evasion were installed. The turnstiles and supporting computer systems, designed by Cubic Transportation Systems, featured many innovations. Credit-card sized magnetic farecards (MetroCard) are sold at self-service MetroCard Vending Machines (MVMs). Passengers are required to *swipe* their MetroCard to enter to the subway. During a swipe, the MetroCard is read, re-written to, then check-read to verify correct encoding (*6*). The read/write innovation allows stored-value fares to be processed and eligibility for transfers automatically determined, completely replacing tokens and paper transfers.

The AFC system consists of approximately 3,300 turnstiles, 4,500 bus fareboxes, 785 MetroCard readers (MRMs), 750 token booth terminals, 1,645 MVMs, 607 MetroCard Express Machines (MEMs), 476 High Entrance-Exit Turnstiles (HEETs), and Autonomous Farecard Access System (AFAS) gates at 49 stations. 460 station controllers and 30 depot computers facilitate communication amongst AFC devices. MetroCard is the form of payment for New York City Subway, New York City Bus (including routes operated by Atlantic Express under contract), MTA Bus Company, Long Island Bus, Port Authority Trans-Hudson (PATH), the Roosevelt Island Tram, AirTrain JFK, and Westchester County's Bee-Line.

### How the MetroCard Works

Each MetroCard is assigned a unique, permanent ten-digit serial number when manufactured. Value is stored magnetically on the card, while transaction history is centrally held in the AFC database. When purchased, the machine encodes card value and updates the database, identifying cards by serial numbers. The AFC database is necessary for maintaining transaction records, to track cards.

MetroCard contains two electronic records describing the last two fare transactions. Transaction data contains control area (station booth) of entry, transaction time, fare paid, and value remaining. All turnstiles in a control area are controlled by a Station Controller computer. As swipes occur, turnstiles transmit transaction information to the station computer. It in turn transmits the information to the Area Controller, an IBM mainframe where the AFC database is held. Transaction data, including fare purchases and expenditures, is stored on the mainframe for a defined period.

NYCT's Office of Management and Budget (OMB) uses transaction data to generate revenue, ridership, and MetroCard usage statistics (*7*). Revenue analysts and internal auditors use the data to detect suspicious patterns of financial and operating activity, assisting with investigations where required.

All token-based transactions were phased out by 2003, allowing NYCT to capture MetroCard data for approximately 95%~97% of all subway fare transactions. The remaining 3%~5%, for which data is not captured, include Single Ride Tickets, other paper fare media, and fare evaders.

### Capturing non-MetroCard Ridership

Tickets and Bus Transfers are made from card stock (rather than polyester polymer used to fabricate MetroCards) and are cheaper to produce. The magnetic stripe is overwritten after use. Tickets have no serial numbers and are not treated as MetroCard transactions. For each swipe, a turnstile (or farebox) 'audit register' is incremented. Total registration count at each turnstile is recorded by the mainframe at predetermined times of day. To capture ticket ridership, registration counts when the period begun is subtracted from registration counts at the end.

Fare evaders and other non-turnstile passengers (including paper transfers, flash passes, children under 44", and large school groups entering using the service entrance) are not captured automatically by AFC registration counters. Registrations are factored and adjusted based on monthly station agent reports of non-turnstile entries.

### Quality Assurance Processes

MetroCard transaction data is subject to much scrutiny, primarily because it is used to process credit card sales, report operating revenues, keep track of values remaining on stored-value cards, and reconcile sale revenues with rides (revenue accrual). The system is designed with the following multiple redundancies to assure correct functionality and high quality data:

1. **High Reliability, High Availability Mainframe Infrastructure:** AFC uses industry-standard financial processing computer equipment to achieve high availability and

reliability, ensure data integrity, and minimize passenger disruptions from AFC malfunction.

2. **Off-line Data Storage:** Should a communications failure occur between station computers and the mainframe, station computers operate in an autonomous, stand-alone mode. Data is stored until communication is restored and the mainframe confirms that data is securely stored. (*6*) During network outages resulting from events of September 11, 2001, data remained on station computers for a few weeks until emergency dial-up links were installed to allow data upload and processing.
3. **Communication Validation Features:** Data is validated during each transmission. Data corruption occurs due to such environmental events as corrupted farecard magnetic stripe (e.g. unreadable serial number) and turnstile malfunction. Station computers perform integrity checks as data is transmitted, and invalid data is automatically refused. Data failing transmission remains on the turnstile until fault is cleared by AFC Maintenance. Later, the mainframe performs checks as station computers upload data.
4. **Database Validation Features:** Transaction database design has dedicated fields for data verification and validation, preventing common database synchronization errors such as lost or duplicate transactions.
5. **Daily Processing Data Validation:** AFC data on the mainframe is processed nightly and stored on the AFC FrontOffice system. During this process, rule-based and table-based validation is performed. Fraud detection algorithms are then executed. Exceptions found are immediately flagged and addressed.
6. **Random Audits:** MTA Audit Services routinely conduct audits of AFC operations. AFC was designed with internal auditing features facilitating consistency checks and silent field observation. Discrepancies in AFC usage and/or suspicious transactions are reported and corrected.
7. **OMB Data Validation:** Additionally, OMB performs range checks and statistical outlier searches prior to using AFC data to report ridership and registration counts.

## DAILY AFC DATA PROCESSING FOR SECTION 15 SAMPLE

Subway AFC requires one swipe per trip. Repeated swipes are not necessary for free within-system transfers. Research shows unlinked trips are systematically undercounted by AFC systems (*10*). To accurately infer unlinked trip and passenger mile aggregate statistics, each passenger's origin, destination, transfer count, and miles travelled must be calculated. The estimation process has six steps, detailed below.

### Download AFC Data

The AFC mainframe is a live system processing millions of transactions everyday. Data must be downloaded to a stand alone system for analysis, to minimize disruption to live data. An export process generates a flat file each night. About 550 megabytes of swipe data is generated daily, and downloaded using a mainframe terminal emulator. PCs download about 200 megabytes of data per hour. For historical data, the mainframe re-generates export files from archives. When large volumes of data is needed, special workstations with direct mainframe connections can download about 5 gigabytes of data per hour.

### Split and Merge Program

Export file contains transaction records from midnight to 23:59:59. However, a MetroCard "Revenue Day" runs from 03:00:00 to 02:59:59 the following day. The export file is split then merged back together in the required sequence using a C++ program, SplitMetroCard. Fast local storage device (e.g. external hard drive) is used, to minimize network traffic and improve throughput. It takes approximately 20 minutes to process one day.

### Determine Turnstile Registrations

Originally, surveyors recorded turnstile register readings ('registrations') for every turnstile. To eliminate this error-prone manual observation and keying process, registrations data is extracted directly from AFC:

- MetroCard transaction counts are available at six-minute intervals by turnstile.
- Single Ride Ticket entries by turnstile are available at four hour intervals. Entry counts are used to produce an adjustment factor for the selected time period.
- Registrations during the sample hour, based on exact start and end times, is retrieved by query. Difference in registrations is passenger entries during the sample hour.

### Determine Passengers' Final Destinations

Riders' destination is inferred from swipes using an industry standard algorithm. Origin is where passenger swiped into fare control. Destination is where the second-swipe into subway occurred. In New York, this algorithm is more than 90% accurate when compared to travel diary surveys (*8*). The extraction process:

- Retrieves unique MetroCard serial numbers for every swipe during sample hour at sample booth.
- Searches AFC data for the next subway swipe of that card, within the Revenue Day. If no next swipe is found, the algorithm searches for the first swipe of that day. (The first swipe corresponds to home station and thus last swipe destination.)
- Retrieves the booth where the swipe occurred, and assigns that station as rider destination. If all other swipes in the 24-hour period occurred onboard buses, rider's subway destination cannot be ascertained, and swipe is treated as 'no response'.
- Where no matching serial number is found, swipe is also 'no response'.
- In practice, non-responses accounts for about 20% of trips (about 12% of swipes), compared to manual surveys where the rate was 80%.

This program produces a list of swipes and destinations, an analogous dataset to that produced by surveyors' Section 15 interviews. Usable swipes are those with determinate destinations. Total passengers is derived by combining turnstile registrations, booth clerk count of non-turnstile passengers, and special counts of trial 'PayPass' passengers.

PayPass is a contactless credit-card payment system being tested at all Lexington Avenue Subway stations and five stations in the Bronx, Brooklyn, and Queens. The keytag-based Pay-per-Ride system is designed to ease congestion at turnstiles and MVMs by reducing time spent paying fare.

### Calculate Trip Distance and Transfers

Prior to computerization, trip distances were calculated in Excel by analysts, who determined path taken by survey passengers based on judgment and experience. Manual mileage calculation was very tedious and prone to errors and typos. Originally designed to automate survey calculations, an Oracle program developed by System Data & Research computes distance travelled for each response. Analysts entered turnstile readings at start and end times, and passenger responses recorded by surveyors. Oracle validated data to reduce data entry errors. Today, AFC dataset is imported directly into Oracle. The program automatically:

- Calculates all possible paths from origin to destination for each trip, using subway schedules in effect at time of survey.
- Looks up mileages for all paths using a distance look-up table.
- Shortest distance path with least transfers is selected as preferred path.
- Where shortest path has more transfers than a longer path, path with lowest transfers per mile is picked.
- Where transfer count is equal, preference is given to transfers requiring shortest walking distance. Cross-platform transfers are chosen over long underground passageways.

This program exactly replicates the old paper process. Through increased algorithmic automation, human judgment elements are removed. Computerized analysis is consistent and reproducible. Passenger miles and unlinked trips are minimized in the overall results, because shortest distance path with the least transfers is always picked.

### Compute Average Miles and Unlinked Trips

The procedural language program "Section 15 Rapid" accessing Oracle calculates passenger miles and unlinked trips for each sample assignment using the same 1984 UMTA-approved formulas, based on individually calculated miles and transfers for each swipe. While data collection and analysis was improved to take advantage of new technologies, sample methodology and factoring algorithms remain faithful to approved original specifications.

At present, this process cannot be used for 100% data. The AFC system generates in excess of 1.5 billion transactions annually. This algorithm was designed to enumerate a small number of responses accurately and precisely, resulting in heavy per-response computational resource requirements. To discretely enumerate transfer counts and miles travelled for all swipes is presently impractical.

## GAINING FTA APPROVAL FOR AFC DATA COLLECTION

Although at the heart of this process is an industry standard algorithm, implementation using AFC data and adaptation to prescribed sampling methodology entailed substantial development work. The major obstacles that were overcome during development are discussed in this section.

**Shortest Path Optimization – Algorithm Development**

New York City's subway system can be counterintuitive to outsiders. *The Map* provides a visual display of the route network, but actual network changes quite substantially depending on time of day. Complication is introduced because not all transfers are equal. Free transfers initially designed during system construction are generally convenient without excessive walking distances or difficult elevation changes. Transfers retrofitted later, particularly new transfers opened after the 1940 unification of the City's three formerly separate rapid transit systems, can be more awkward. Multiple elevation changes, long walks in underground passageways or along active platforms are not unusual. While passageways provide paid-area connections for the convenience of Pay-per-Ride customers, regular commuters prefer easier and quicker transfers.

The shortest-path program was designed, implemented, and parallel-tested from 2004 to 2007 against original mileage calculation spreadsheet. First program versions did not account for transfer distance when calculating paths. Counterintuitive results arose in certain odd cases, e.g. transfers between A and 1 at 42 Street via long passageway between Port Authority Bus Terminal (8 Avenue) and Times Square (7 Avenue), instead of a more direct transfer at Columbus Circle (59 Street-8 Avenue), because 42 Street Transfer minimized on-board passenger miles. NYCT's complex and interconnected subway system meant the algorithm required substantial refinement prior to acceptance.

**Approvals Process**

Section 15 data is a Federal funding element. Data-gathering methods are highly prescriptive. Changing sampling methodology can have funding implications not only for one agency but for all others relying on apportioned funds. Any change to Section 15 data collection, even as simple as replacing manual surveys with electronic data, is subject to high levels of scrutiny from FTA.

Using AFC data to determine sample passenger destination does *not* change the sample methodology. AFC-generated data is incorporated in lieu of human interaction (subway rider interviews) while still maintaining FTA-approved sample. Initially, May and October 2005 data was downloaded from the mainframe as proof-of-concept. Results were calculated and compared with passenger miles and unlinked trips calculated using surveyor data. Results of the two 'pilot' months were presented to FTA for discussion in early 2007.

FTA was interested in NYCT's efforts to upgrade from error-prone manual surveys to automated AFC data downloads. However, FTA was concerned about data quality and impacts on result continuity and comparability from year to year. As a trial, FTA requested full parallel testing of this new methodology. The remaining ten months of 2005 were chosen as a test period for baseline comparison with surveyor passenger interviews.

**Parallel Testing**

In May 2007, annual and monthly 2005 results and comparison was sent to FTA for review (Figure 2). Together with statistical results, NYCT prepared a list of benefits from adopting AFC-based data collection:

1. Provides higher accuracy and consistency by eliminating the "human" element of data collection.
2. Swipe data is synchronized with sources of annual registrations used in extrapolating sample data.
3. All control areas are measurable including areas converted to HEET and Paypass (smartcard) turnstiles. This is especially important because full time and part-time control areas have been and will continue to be reduced.
4. Eliminates paper data collection and surveyor interpretation of final destination. AFC data also captures non-English speakers who would not normally respond to surveyors.
5. Data is electronically accessible, making it easier to retrieve for audits.
6. Reduces four months lag for unlinked passenger trips calculations for monthly NTD Internet reporting to FTA.

**Figure 2**

**SECTION 15 SUBWAY DATA: JANUARY - DECEMBER 2005**

**PASSENGER MILES & UNLINKED TRIPS**

**Surveyor's vs MetroCard O/D Data**

| | | TOTAL REGISTRATION | RESPONSE | PASSENGER MILES | UNLINKED TRIPS | Ratio of Unlinked Trips to Total Registration |
|---|---|---|---|---|---|---|
| ❶ | Suerveyor's data | 1,468,463,274 | 59,442 | 8,402,147,333 | 1,804,034,331 | 1.23 |
| ❷ | MetroCard O/D Data | 1,468,463,274 | 195,283 | 8,984,910,917 | 2,172,124,952 | 1.48 |

| COMPARE | PASSENGER MILES | UNLINKED TRIPS |
|---|---|---|
| ❶ | 8,402,147,333 | 1,804,034,331 |
| ❷ | 8,984,910,917 | 2,172,124,952 |
| % (+/-) | **6.94%** | **20.40%** |

**Note:**

**1) Surveyor's data is based on 2005 yearly final submission numbers.**

**2) The increase in unlinked trip is due to the more accurate destination determination based on MetroCard algorithm.**

While generally FTA required traditional 'traffic checks' with manual observations for new technology data system trials, NYCT downloaded historical AFC data for exactly the same sample and calculated same statistics. In that sense, it was a true parallel test of AFC data versus surveyors.

The results were surprising. The unlinked trips and passenger miles calculated from the same sample with AFC data was higher than statistics calculated from surveyor data. Although data collection methodology was changed, sampling and tallying methodology did not change. What might have caused the difference?

After extensive investigation conducted by NYCT responding to NTD and FTA concerns, it was determined that traditional methodology (surveyors) resulted in reporting biases that accounted for the difference. Errors in the surveyor data arose due to data collection issues (too many passengers, non-responsive passengers, passengers with no English, and incorrect destination),

and data analysis problems (data interpretation, transcription errors). Surveyor and AFC turnstile entry counts did not agree due to AFC being based on an exact one hour period, whereas surveyors report to stations early to obtain turnstile readings prior to commencing interviews, extending time period beyond one exact hour. Manual surveys were unable to measure ridership via HEETs because no turnstile readings are displayed. Together these factors explain the differences in the data found during parallel testing.

FTA was satisfied with data comparison and requested a similar calculation for 2006 AFC data. In September 2007, AFC versus surveyor comparison for 2005–06 was submitted. NYCT also provided supporting documentation for FTA review. The supporting information included a long list of data and process description:

- Flowchart and brief description comparing current process (passenger interviews) and proposed process (AFC data)
- Summary comparison
- Full Section 15 submission based on surveys (as originally submitted for years 2005 and 2006), including: annual extrapolation worksheet, summary registration information, formulas used, complete random sample, and a selection of survey forms gathered
- Full proposed submission based on AFC data, including: all sections corresponding to original submission, and printouts from Section 15 shortest path program showing destination station, miles travelled, and transfers for all responses in sample hour, and summary statistics for every sample hour

Supporting documentation comprised about 100 pages and enabled side-by-side comparisons between AFC and surveyor data.

**FTA Special Request**

On January 17, 2008, FTA approved the use of AFC data for NYCT's Section 15 sampling conditionally on the complete submission of revised NTD monthly data based on AFC sampling for 2005–07 by January 31, the normal deadline for 2007 NTD submission. The 2007 data, consisting of 700 samples and more than 200,000 AFC trips, needed to be processed in just over two weeks. This was a very difficult challenge:

- 200 gigabytes must be downloaded from mainframe on network infrastructure normally designed to download 0.5 gigabytes per day.
- SplitMetroCard program must be executed 365 times sequentially, consuming up to half-hour per run and requiring computers with ample local disk storage "scratch space".
- Shortest path program, which relies heavily on remote queries executed on NYCT's enterprise server, must be executed for 700 sample assignments. Execution time per sample ranged from a few minutes to twelve hours, depending on passenger response count and subway system complexity as viewed from sample control area. Stations with many lines and transfer possibilities have a large 'pyramid' (hierarchical listing of all trip possibilities), and require longer calculation time than stations with simple pyramids. Execution time is difficult to predict prior to pyramid creation. Experience gained during parallel testing suggests this is the 'slow step' in calculation.

NYCT's Senior Director in System Data & Research (SDR) personally committed to and led the implementation of this arduous task. The incentive given to staff was either to meet the deadline or return to manual data collection and processing. Finishing the job would normally take six months. To accelerate computation, SDR developed a manual parallel processing infrastructure:

- A master book of sample assignments (January through December), showing all 700 samples and identity of assigned PCs, was used to track computation progress. Date and time of process start, resume, and termination was recorded.
- Shortest path program was installed on 15 PCs to obtain maximum parallelism. Up to five threads were started per PC, and about 10 PCs were devoted to calculations concurrently. Each thread independently managed calculations associated with one sample hour, but multi-threading allowed different sample hours to be computed concurrently.
- The office was staffed round-the-clock and weekends, in three shifts. NYCT's enterprise server offered better performance during off-hours due to reduced activity in other departments. Calculation throughput improved markedly during overnight periods. Staff presence was necessary to ensure PCs stayed up and to re-start any processes that terminated abnormally. During this firestorm, every one of fourteen analysts and managers in SDR (including Senior Director) staffed the computer cluster for at least one shift. Successful outcome depended on a real team effort.
- Abnormal program termination was caused by: (a) anti-virus software interference; (b) security policy forced 2 a.m. logoffs; (c) Oracle server occasionally dropped connections due to restrictions in maximum number of parallel client connections; (d) maintenance related network outages.

After 14 days and nights, calculations were complete and data entered on NTD website on January 30. As a result of this unprecedented effort demonstrating that NYCT is capable of calculating passenger miles automatically on a production basis, and detailed side-by-side comparisons showing exactly the same sample and processing methodologies, FTA approved AFC-based annual (appropriation) submission for 2007 and monthly (statistical) submission for 2008. Starting 2008, NYCT submits AFC unlinked trips data monthly.

## LESSONS LEARNED

During the twenty-year span when manual data collection was the method of choice, NYCT identified many practical problems. Aside from typical problems such as high costs, low response rates, missed assignments, and inconsistencies, issues specific to NYCT such as data interpretation for identically named stations were observed.

Developing reporting systems for automatically collected data is an obvious next step for many transit agencies that have commissioned new technology data systems in the past decade. Aside from AFC data, many systems collect useful data streams such as signal system data, maintenance monitoring data, event recorder data, and even automated passenger counter and vehicle locator data. These disparate data streams could be used for internal reporting, monitoring, planning, troubleshooting, and external reporting. Using automatically collected data sidesteps problems and costs associated with traditional data collection and processing.

Using automated data collection isn't without pitfalls. The NYCT experience suggests some issues to consider while implementing processes that replace manual data collection:

1. **Use Sample Data First:** Although extensive data is collected cheaply and ubiquitously, a sample dataset should be used on a trial basis to develop reporting processes. Sample dataset is easier to work with. Faults in data quality, algorithms, and processes are easily detected when working with smaller datasets.
2. **Involve Regulatory Authorities:** NYCT obtained regulatory approval to use revised data collection process because FTA and NTD were involved from the project's conception and were kept abreast of new development and results. While data auditing is still necessary, this was made easier because FTA and NTD officials were fully on-board, understood the project's purpose and development process.
3. **Expect and Use Full Parallel Testing:** Parallel testing is a necessary part of commissioning new systems. If comparable data is not available, baseline data should be collected manually. Present automatic and manual datasets side-by-side in a scientific comparison. This process will detect anomalies within data (for both datasets), and serve to convince data consumers of its accuracy once all errors are worked out.
4. **Design for Scaling Up:** Sample data can be processed in reasonable time even by simple algorithms running on commodity PCs. Extensive data demand enormous computing resources and efficient data processing methods. When designing the process for sample data, consideration should be given to efficiency and computing requirements, to allow scaling-up in production mode. Explicit computation time, storage capacity, and network bandwidth calculations should be performed to determine if scaled-up algorithms can run in reasonable time. In this case, a simplified algorithm using an approximate look-up table (perhaps split into matrices based on originating time-of-day) could be a first step towards achieving higher throughput.
5. **Use Spiral Model and Software Prototyping:** The process evolved incrementally over years of planning and design. What began as very basic mileage and transfer calculation automation culminated in a fully automated reporting system.

**Future Work**

NYCT is currently applying these lessons learned to other data processing projects, using a rich set of automatically collected data currently available:

1. Using AFC data to generate cordon counts (passengers crossing into Manhattan's Central Business District)
2. Using AFC to determine passenger loads and peak load points for buses, supplementing surveyor 'Ridechecks'.
3. Using signal data to compute on-time performance and headway regularity, for rail lines where Automated Train Supervision is enabled.

The difference between these projects and data mining typically conducted by planning agencies is the development of automated processes that use real 'live' data to produce reports routinely – instead of using historical data to calibrate a 'model' used for long-range decision-making. Using 'live' data for operations decision-making (even though decisions are not real time) is an

exciting and pioneering development. It is hoped that systems can be developed to allow schedule and service planners to monitor service performance much more frequently than currently possible due to inherently high cost of manual data collection.

**ACKNOWLEDGEMENTS**

The authors would like to acknowledge the support and assistance of the following individuals during the development of this project and preparation of this paper: Gary Delorme, Federal Transit Administration; Sergio Maia, National Transit Database; Miguel Garcia and Daniel Rodriguez, MetroCard Office; Lawrence Hirsch and Qifeng Zeng, Office of Management and Budget; Anthony Cramer, Robert Menhard, Thomas Chennadu, and Chi Chan, Operations Planning. Thanks to John Allen, Regional Transportation Authority Chicago and Chandra Buie, New York Transit Museum for research assistance. Responsibility for errors or omissions remains with the authors. Opinions expressed in this paper are those of the authors and do not necessarily reflect official policy of Metropolitan Transportation Authority or MTA New York City Transit.

## REFERENCES

(1) K. J. Hom. Reinventing Transit: The Twenty-Year Overnight Success Story of NYC Transit. Presented at the Metropolitan Conference on Public Transportation Research, University of Illinois at Chicago, June 11. (1999).

(2) D. L. Gunn. Subway Returns to a State of Good Repair. Special Feature: Developing Metros, Railway Gazette International. (1988).

(3) 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 on May 2, 2008.

(4) G. Kocur. *Revised Procedures for Collecting Section 15 Service Consumption Data on Rail Rapid Transit Systems.* Charles River Associates, Boston, Mass. (February 24, 1984).

(5) Urban Mass Transportation Administration. *Untitled Docket I.D. #2008*, dated January 16, 1989 from Rhoda Shorter and addressed to John Kaiser, Director, MTA Grant Management, 347 Madison Avenue, New York, N.Y. (1989).

(6) Riazi, Atefeh. *MetroCard: Automating New York City's Public Transit System.* Proceedings of the First International Conference on Urban Transportation Systems, American Society of Civil Engineers, Miami, Fla. (March 21-25, 1999).

(7) New York Transit Museum. *Show Me the Money: From the Turnstile to the Bank*, Final Exhibit Text. New York, N.Y. (2008).

(8) Barry, J.J., Newhouser, R., Rahbee, A., and Sayeda, S. *Origin and Destination Estimation in New York City Using Automated Fare System Data*, Transportation Research Record (2000).

(9) Zhao, J., A. Rahbee, and N.H.M. Wilson. *Estimating a Rail Passenger Trip Origin-Destination Matrix Using Automatic Data Collection Systems.* Computer-Aided Civil and Infrastructure Engineering **22**, pp. 376-387 (2007).

(10) Utsunomiya, M., J. Attanucci, and N. Wilson. *Potential Uses of Transit Smart Card Registration and Transaction Data to Improve Transit Planning.* Transportation Research Record 1971, pp. 119-126 (2006).