Proceedings of the 2021 Joint Rail Conference
JRC2021
April 20-21, 2021, Virtual, Online

JRC2021-59109

ESTIMATION OF PRE-COVID19 DAILY RIDERSHIP PATTERNS FROM PAPER AND ELECTRONIC TICKET SALES DATA WITH ORIGIN-DESTINATION, TIME-OF-DAY, AND TRAIN-START DETAIL ON A COMMUTER RAILROAD: QUICK-RESPONSE BIG DATA ANALYTICS IN A WORLD STEEPED WITH TRADITION

Alex Lu
Metro-North Commuter Railroad
New York, N.Y.
(corresponding author)

Thomas Marchwinski
Metro-North Commuter Railroad
New York, N.Y.

Robert Culhane
Metro-North Commuter Railroad
New York, N.Y.

Xiaojing Wei
MTA Construction & Development
New York, N.Y.

ABSTRACT

Our niche method independently estimates hourly commuter rail station-to-station origin-destination (OD) matrix data each day from ticket sales and activation data from four sales channels (paper/mobile tickets, mail order, and onboard sales) by extending well-established transportation modelling methodologies. This algorithm's features include: (1) handles multi-pack pay-per-ride fare instruments not requiring electronic validation, like ten-trip paper tickets "punched" onboard by railroad conductors; (2) correctly infers direction-ality for direction-agnostic ticket-types; (3) estimates unlimited ride ticket utilization patterns sufficiently precisely to inform vehicle assignment/scheduling; (4) provides integer outputs without allowing rounding to affect control totals nor introduce artifacts; (5) deals gracefully with cliff-edge changes in demand, like the COVID19 related lockdown; and (6) allocates hourly traffic to each train-start based on passenger choice. Our core idea is that the time of ticket usage is ultimately a function of the time of sale and ticket type, and mutual transformation is made via probability density functions ("patterns") given sufficient distribution data. We generated pre-COVID daily OD matrices and will eventually extend this work to post-COVID inputs. Results were provided to operations planners using visual and tabular interfaces. These matrices represent data never previously available by any method; prior OD surveys required 100,000 respondents, and even then could neither provide daily nor hourly levels of detail, and could not monitor special event ridership nor specific seasonal travel such as summer Friday afternoons.

Keywords: commuter rail, origin-destination matrix, ticket sales, ridership estimation algorithm, travel pattern

1. INTRODUCTION

This paper describes a novel method to estimate commuter rail station-to-station origin-destination (OD) matrix at an hourly level of granularity (assignable to specific trains), separately and independently for each day. This algorithm combines and extends well-established transportation modelling methodologies and applies it to a niche problem with important practical implications that has nonetheless seen limited attention from the expert community. Special features of this algorithm includes:

- Handles multi-pack pay-per-ride fare instruments that do not require electronic validation at the time of use, such as a ten-trip commuter rail paper ticket that is "punched" by a railroad conductor at the time of ride, with no electronic record being made of ticket usage.
- Infers directionality for direction-agnostic ticket-types, such as monthly tickets sold between pairs of stations.
- Sensitive to day-to-day changes in travel conditions, such as weather, special events, and network disruption.
- Ability to deal with a cliff-edge sudden change in demand or ridership patterns, such as one faced by commuter rail operators following the COVID19 related lockdown.
- Estimate utilization patterns of unlimited-ride tickets in a sufficiently sophisticated fashion to inform operations planning decisions (e.g., vehicle assignment, scheduling, stopping patterns, connections) in a useful way.
- Provides output in terms of whole numbers of passengers for each OD pair during each hour, without allowing rounding to affect overall control totals (e.g. daily total ridership) whilst keeping the probability of each origin, destination, and hour combination proportional to their

fractional ridership estimate over the long term, such that rounding artifacts do not appear in long-run average data.
- Allocates hourly traffic to each train-start (i.e. schedule number) based on passenger choice.

Data processing capabilities to analyze ticket-sales data this way had existed for approximately fifteen years (e.g. [1-5]), however, conditions that generated the required input datasets for a commuter railway (specifically relating to the electronic ticket activation data) was unavailable until recently. eTix data on the target system is equivalent to a 40% sample of the overall ridership population; this large and dispersed sample allows us to accurately estimate behaviours of the remaining 60%.

2. MODELLING METHODOLOGY

2.1 Input Data Description

This algorithm takes as input a series of different data sources. Not typical for metro systems, but common on mainline railways, is the existence of multiple sales channels which issues different fare media and for which sales transactions are recorded on different systems (e.g. Figure 1). The target system has four main sales channels and those are reflected in the input data streams:

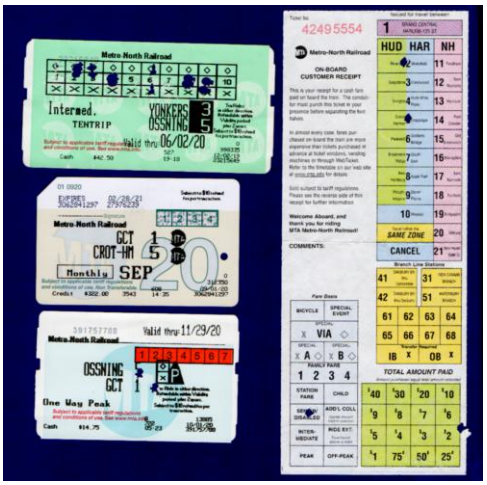

FIGURE 1: DIVERSITY OF PAPER FARE MEDIA

Self-service ticket vending machines at stations and ticket printers operated by station agents generate sales transaction records in a centralized database which records the two points between which the tickets are valid, type of ticket (one-way, multi-trip, periodic unlimited ride, special discount), time of sale, and a unique ticket identifier. Fare media, once printed, are handled manually (visual inspection, ticket punch cancellation) and no electronic records are generated at the point-of-use. This is very different from a subway environment where turnstiles provide electronic records at the point-of-entry.

Mobile tickets (eTix), which is a proprietary application installed on the passenger's smart phone, transmit over the public cellular communications network sales transaction records providing much the same information as the previous source, but also, crucially, adds activation information detailing when each ticket was used. Pay-per-ride (PPR) tickets must be "activated" prior to use and generate a use record. Unlimited ride tickets generate an activation record when the app is opened for visual inspection, and do not generate additional records even if the app is repeatedly opened and closed within a time window approximating the maximum length of a trip.

Mail order tickets have a customer database that stores standing orders from monthly ticket purchasers, and a refund database that records returns of unused and unexpired tickets.

Onboard ticket sales generate a database detailing one-way tickets and upgrades sold (e.g. off-peak to peak fare, or extensions of ride).

2.2 Description of Fare Types

The target system has more than sixty ticket types described in a detailed tariff document, however, for the purposes of modelling travel behaviour, in most cases it is only necessary to differentiate between four basic categories: monthly commutation, weekly commutation, ten-trip, and à-la-carte single/return tickets. PPR tickets are further subdivided into peak, off-peak, intermediate, child, senior, family fare, and special discount schemes. This distinction is only important in certain cases, discussed later.

2.3 Key Assumptions

Prior to the introduction of eTix, it was impossible to know, once a ticket is sold, when and how the customer actually used it. At infrequent intervals, surveys are carried out to determine how much monthly ticket holders utilize their tickets, but there is no time-of-day nor day-of-week detail. The eTix activation records provide a means to understand this at an extensive level of detail—as well as when and how multi-ride tickets are used in the wild. The eTix ridership comprises approximately 40% of total ridership, although it varies by ticket type, with more monthly tickets remaining on paper and more PPR tickets sold as eTix.

It is necessary to assume that eTix passengers utilize their tickets in approximately the same way as paper ticket holders. We know from surveys that the demographic of paper ticket holders trend older, but we have little reason to believe their temporal travel patterns or frequency of use are on aggregate significantly different from eTix users.

We also make the assumption that paper single tickets are purchased immediately prior to use. It is difficult to verify this assumption in any meaningful way because eTix passengers do not have any point-of-sale issues (e.g. queuing, agent involvement) that may require them to purchase tickets in advance, and there are no discounts for advance-purchase tickets. Based on anecdotal experience, we believe very few single ticket users purchase their tickets in advance. (Not true for unlimited ride tickets, and we account for that separately.)

2.4 General Concept of "Lookback Windows"

It is useful to think of PPR tickets as inventories of trip-coupons that are held by customers, which are cancelled upon fulfillment of transportation. Indeed, historical accounting practices reflect this tradition: tickets sold were carried as liability on the company's books ("transportation owed"), until tickets were lifted and cancelled by traincrews, and physically returned to the accounting department for booking as revenue earned. Therefore, to understand the probability that a given ticket would be cancelled today, it is necessary to model the customer's outstanding unused inventory of tickets. Because of business process changes, revenue accrual today no longer depend on the physical return of collected tickets, thus no reasonable way existed to know when a paper ticket was used.

Figure 2 shows the cumulative fraction of ten-trip eTix rides that were cancelled within *N* weeks of purchase. Although technically these tickets are valid for six months, the data shows 90% of all rides were consumed within ten weeks. For processing time reasons, we limit the probability modelling to 92-days of historical ten-trip sales data.

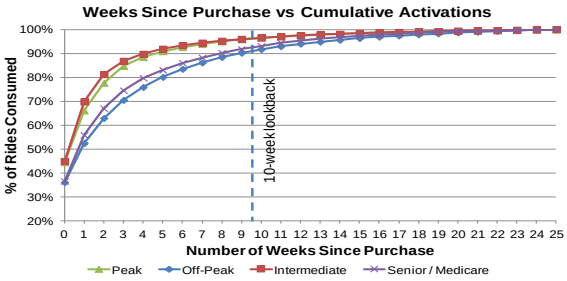

FIGURE 2: PERCENTAGE OF RIDES CONSUMED AS A FUNCTION OF WEEKS SINCE PURCHASE OF TEN RIDE ELECTRONIC TICKETS

A similar analysis examined round-trip eTix, where the data showed 97.5% of return portions were collected within seven days of purchase. We chose to limit the modelling to a lookback window of 31-days.

Having observed the distribution of "days each ticket is held by the customer" from the eTix data, the number of rides taken today is actually just the sum product of {probability of usage after *N* days}, and {number of tickets sold exactly *N* days ago} during the lookback period. Figure 3 explains this calculation graphically.

Conceptually this is a very simple calculation, but numerous practical challenges exist: (1) it is necessary to choose the categories within which we group the distributions together, i.e. define appropriate strata within the 40% uncontrolled sample; (2) it may be necessary to modify this basic model to fit the usage and behaviour patterns for each ticket type; (3) since results are probabilities and not observed ticket usage, they can be very small decimals in a given market during a given hour; (4) due to the sheer volume of data required, computational processing time must be considered when specifying the model. In the following sections, how this basic concept is applied to each ticket type will be described.

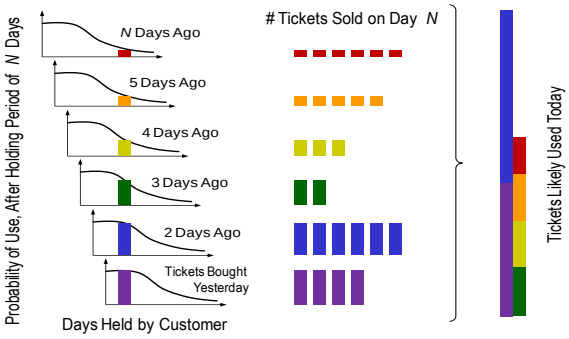

FIGURE 3: BASIC CONCEPTUAL MODEL RELATING TICKETS EXPECTED TO BE CANCELLED TODAY AND CUSTOMER INVENTORY OF TICKETS

2.5 Ten-Trip Ticket "Weeks Since Purchase" Model

We know from past experience that ten-trip tickets are generally utilized by customers who have occasional needs to travel between fixed points (e.g. between their home station and a place where they occasionally do business). For this reason, we believe their date and time of travel are driven by when they need to do business, rather than strictly based on how long they have been in possession of fare media. However, we also know that customers try to use their tickets sooner rather than later, due to their inventory carrying costs, so the days-held concept will still apply.

To provide a compromise model, we chose to first compute the number of paper tickets expected to be utilized during the current week, based on a distribution of weeks since initial purchase. For ten-trip tickets, this distribution varies subtly by ticket subtype (peak, off-peak, intermediate, and senior), as Figure 2 shows, because off-peak and senior tickets (valid off-peak only) are less likely to be used by occasional commuters (e.g. attorneys with occasional Manhattan court appearances).

Having computed number of tickets used during this week, the model then sprinkles the ticket usages over the course of the week based on a combined distribution of day-of-week and time-of-day, by ticket subtype. Figure 4 is a representative illustration how the usage of tickets differs by time-of-day, day-of-week, and subtype. For instance, fewer peak tickets were used on Fridays compared to Wednesdays.

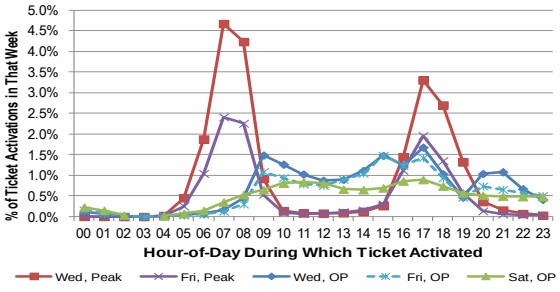

FIGURE 4: HOUR-OF-DAY DURING WHICH TEN TRIP TICKETS WERE ACTIVATED FOR PEAK AND OFF-PEAK TICKETS ON WEDNESDAYS, FRIDAYS, AND SATURDAYS

Note that this model does not depend on origin and destination station information provided in the ticket data because we found that time of travel is not significantly affected by geography. This is not true in intercity travel where a journey time comprises a significant portion of the passenger's day (therefore giving rise to "sleeper" and "mid-afternoon express" type services), but in a commuter environment our data shows that this effect is minimal.

2.6 Return (Round-Trip) Ticket Model

It is a commonly-held belief within the commuter rail industry that round-trip passengers in essence fall within two distinct markets: a day-return market that tends to leave early and come back late, and a period-return market where two trips are practically independent in terms of departure time, where the return ticket is merely a device to obtain discounted fares or avoid the inconvenience of another transaction. The latter market tends to be more elastic as it is predominantly driven by leisure travel. It is not entirely obvious where one market ends and the other begins, particularly when the outbound trip falls on a Friday or Saturday. We thus sought to answer this question with data.

Figure 5 shows a typical pattern. Those who travel outbound in the AM peak tend to return in the PM peak, although some return during the AM peak the following day. Those who start their trip midday usually return either during the PM peak or late evening hours. Those who travel outbound in the PM peak typically return in the late evening, the following AM peak, or the following PM peak. Generally speaking, those patterns held steady for Mondays through Thursdays, but a distinct pattern is seen each for Friday, Saturday, and Sunday. Finding a distinct pattern for those three days is not uncommon [6]. In each case, these patterns peter out after about 48 hours.

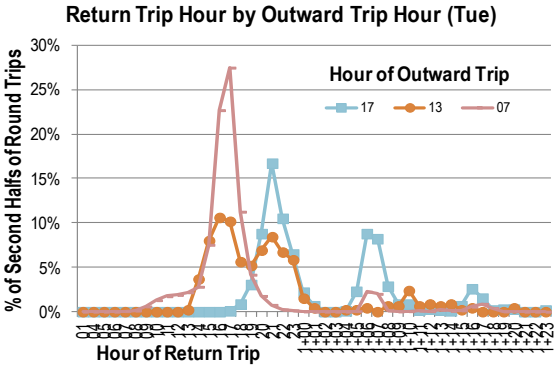

FIGURE 5: HOUR OF RETURN TRIP AS A FUNCTION OF HOUR OF OUTWARD TRIP (AM PEAK, MIDDAY, AND PM PEAK) IN PAY-PER-RIDE RETURN TICKETS

For this model, we therefore chose a segmented approach. For paper tickets purchased today and yesterday, we distribute the time of return trip based on day-of-week and time-of-day when the ticket was sold (i.e. a proxy for when outbound trip was taken). We then apply a filter to discard expected return travel that doesn't occur today (such that tickets bought yesterday and used for return travel yesterday doesn't interfere with today's results). We call this portion the "48-hour model".

Although we do have peak and off-peak information in the sales records, we decided not to use that information for two reasons: (1) because we distribute expected ticket usage hour-by-hour, that distribution should capture the peak/off-peak distinction anyway, because user behaviour is reflected in the distribution; (2) peak tickets are valid during off-peak periods, and off-peak tickets are often used during peak periods upon payment of appropriate step-up charges. Although we could in theory process all of this data (correlating paper ticket sales with onboard sales), it makes the model unnecessarily complex.

For tickets purchased earlier than yesterday, we apply the basic "days away" logic by day-of-week (because more passengers stay two or more nights when the outbound travel occurs on a Friday) to determine the fraction of tickets expected to be used today. Once the expected daily total ticket usage is found, we then sprinkle that based on the distribution of return trip hours from all return portions of round-trip tickets where passengers stayed at their destination for at least two days. The use of this distribution, rather than a general distribution of hour-of-day in which tickets are used, is significant, because we found that the use of return portions after at least one night's stay is likely to be biased towards earlier in the day (see Figure 6). We call this the "period-return model".

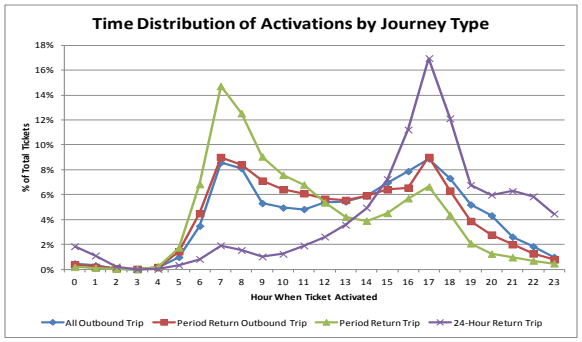

FIGURE 6: HOUR OF RETURN TICKET USAGE AS A FUNCTION OF OUTBOUND/RETURN PORTIONS AND DAY RETURN VERSUS PERIOD RETURN

In the off-peak, two passengers travelling together often purchase a return ticket and use both portions for the same journey (permissible). This shows in the data as activation of both portions within minutes of one another, and requires the origin and destination records to be swapped for the return portion.

The 48-hour model covers 87% of the market overall for us, but on Fridays its coverage drops to 78%. Both models are important in providing a complete view of travel behaviour.

2.7 Monthly Ticket "Day-of-Week Zone Hour" Model

Initially we thought that geography might affect unlimited-ride ticket utilization, as passengers with longer commutes (some in excess of two hours each way) might be expected to utilize their monthly ticket less frequently or intensively than

someone whose commute is only a half-hour long. Figure 7 shows this is not really the case.

We knew from prior work [7] that the number of weekdays and holidays in each month has a significant and measurable impact on monthly ticket utilization, we therefore model monthly ticket utilization as a two-stage problem: (1) given today is, e.g. a Friday in January, what is the fraction of monthly ticket issued that I would expect to "see" on the system; (2) given that a specific ticket is seen today, how many trips do we expect that ticket to redeem? This is in contrast to a more traditional approach that might measure average trips per month. Multiplying the number of tickets issued by both of these factors provides total rides expected to be consumed today by holders of these tickets. These factors are computed separately for each month, and for each day of the week. By doing this, the impact of holidays (e.g. Martin Luther King day, July 4, Thanksgiving week, Columbus Day, etc., all of which have distinct levels of commuter ticket usage) are automatically and specifically accounted for. In determining these factors it was necessary to enumerate count of each daytype within each month, requiring some complex calendar maths.

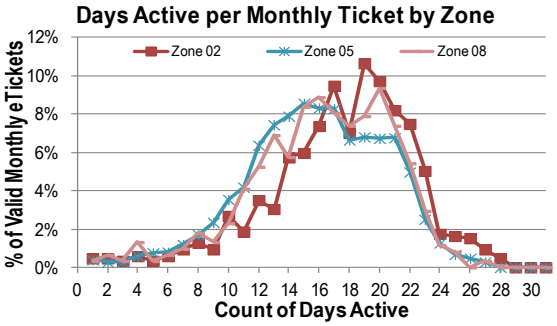

FIGURE 7: DAYS MONTHLY TICKETS ARE SEEN TO BE ACTIVE BY FARE ZONE (ZONE 2 = CITY, ZONE 5 = PRIME SUBURBAN TERRITORY, ZONE 8 = EXURBAN AREAS)

Monthly tickets are sold beginning on the 25th of the prior month, and continue to be sold up until around the 15th. The lookback window for monthly tickets is therefore 38 days, and a filter is applied to find only those tickets sold relevant to the current month. In addition to ticket vending machine data, sales data from mail order tickets are also merged in. eTix monthlies are treated separately because we can directly observe activations of those "tickets".

Figure 8 shows two interesting phenomena: (1) morning commutes begin earlier for those living further away from downtown (i.e. Zones 09/10); (2) on the system's extremities where service frequency is sparse, afternoon ticket usage is tied to specific popular train departures (Zone 10). Although we do not need to model ticket usage frequency as a function of geography, we need it for accurate estimation of hour-of-day.

Having determined how many rides would be redeemed today (preserving origin-destination information in sales data), we sprinkle these rides amongst the 24-hour day based on relevant hourly distributions for that specific day-of-week, and origin and destination fare zones. The day-of-week is included because temporal travel patterns are different for Friday, Saturday, and Sunday, even amongst season ticket holders.

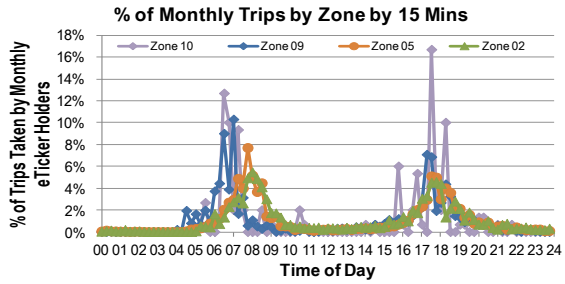

FIGURE 8: PERCENTAGE OF TRIPS TAKEN BY MONTHLY ETIX HOLDERS AS A FUNCTION OF TIME-OF-DAY, BY ZONE

Despite the aggregation of geography from station-to-station to the larger farezone-level (one fare zone contains multiple stations), there were nonetheless some OD zones on specific days of the week for which no eTix activations were observed at all. In those cases, the estimated paper ticket usage risks getting lost unless there is a "residual" process to recover them. In cases where no time-of-day pattern could be found (~0.2%), we divide them according to a generic zone-independent hour-by-hour activation distribution of all monthly tickets.

2.8 Weekly Ticket Model

The weekly ticket model is basically identical to the monthly model, except that the "lookback period" is 12 days, and utilization factors are separately computed for each of the 52 weeks of the year.

2.9 Origin-Destination Model Structure

At this point, it is useful to review the estimation model's overall structure. In summary, eTix sales and activation data are used to generate various distributions, separately for each ticket type, representing how the tickets that were sold, would end up being used. These patterns determine the number of rides redeemed, and the date and time of such rides relative to when the tickets were sold or the ticket's validity period. These patterns do not determine the ticket's geography, which is recorded at the time of sale and typically not altered except through on-board extension-of-ride supplemental fares.

Separately, sales records from various non-mobile sales channels (from which ticket usage cannot be observed) were combined, summarized, and multiplied by these distributions based on variables relevant to each ticket type. This produces a daily hour-by-hour, station-by-station partial OD matrix that represents the probability that tickets with that OD would be used during that hour on that day. This is combined with sales data for single ride tickets and outbound portions of return tickets, which represent actual ODs observed. Figure 9 shows a high-level block diagram.

Probabilities are usually decimals, and fractional passengers are not necessarily useful in transport planning. In commuter rail, small ODs at unsociable hours have very sparse

demand, seeing perhaps one or two trips a month, translating into 0.03 trips per hour per day. If a one-way ticket sale was seen during that hour for that OD, then we know for sure that someone did travel that day, but a method is still needed to definitively assign that one probabilistically estimated marginal passenger to one specific hour, to represent the possibility that someone perhaps did travel that day, but this must be done in such a way as not to affect either long-term averages for that specific hour in that OD, nor to affect daily "roll-up" control totals. This is the "Fractional Passenger Dithering Process".

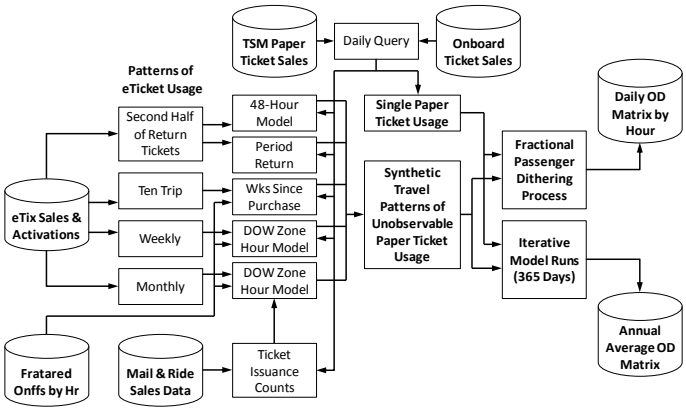

FIGURE 9: PHASE I BLOCK DIAGRAM SHOWING HOW PARTS OF THE TICKET RIDERSHIP ESTIMATION MODEL WORK TOGETHER WITH DIFFERENT DATA SOURCES

Figure 10 represents the solution space for the system's OD matrix during the seventeenth hour for a normal weekday. Grey pixels represent OD pairs that contain nonzero values representing less than one passenger. Black pixels are more than one passenger, and white pixels are either invalid ODs (meaning tickets are not sold for that market) or an OD having exactly zero passengers assigned.

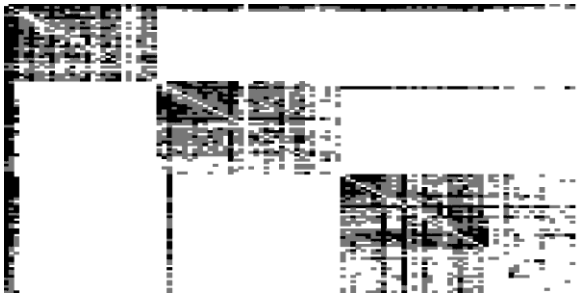

FIGURE 10: GRAPHICAL REPRESENTATION OF TYPICAL PROBABILISTIC DAILY-BY-HOUR COMMUTER RAIL OD MATRIX—GREY CELLS REPRESENTS BETWEEN ZERO AND ONE PASSENGERS DURING ANY ONE GIVEN HOUR

Commuter rail demand is so focused on travel in and out of important urban centres, including between central business district and suburbs, that even during height of the rush hour many markets show less than one daily passenger per hour. (Although, it should be noted that with 4,646 valid OD pairs, and about 250,000 trips per weekday, the average hourly ridership per OD market assuming completely uniform distribution would only be 2.24 passengers; these issues are inherent in modelling low density traffic.)

2.10 Error Diffusion for Passenger Counts

The algorithm for finding individual integer solutions for each cell that fit within the overall ridership picture borrows from an established methodology in digital audio and image processing called "dithering" [8]. Dithering is the intentional addition of noise (random or systematic) to prevent large scale quantization errors when converting from a continuously variable source to one with discretely defined levels. Both truncating (using only the integer part of output) and 4/5 rounding leads to this predictable and repetitive form of error, and won't preserve overall control totals either within margins of the matrix, or across multiple days or hours of OD data.

Our error diffusion algorithm treats the OD matrix sequentially, sorting fractional outputs by origin station, destination, and hour-of-day. Consequently, accumulated errors are in effect first moved to adjacent hours, and then to adjacent destinations if necessary. The algorithm keeps track of errors from truncating each non-integer value, and augments the output by one passenger in that OD-hour combination when a specific criterion is met. To keep control totals constant, it is important to augment exactly one market by one passenger for each 1.000000 worth of accumulated truncation errors. This approach is conventionally termed "bucket rounding". However, we must not simply augment markets within which the 4/5 rounding condition is reached, because doing so could lead to systematic and repetitive errors, which would be visible in the data and could artificially inflate ridership significantly in those markets over the long term, when, for instance, a whole month's worth of OD matrix data is summed.

One approach to solving this issue is to essentially randomize the point at which accumulated errors are added to the market, such that overall probability of "adding one" is proportional to the contribution of truncation errors from each market. Although this would lead to proper outputs on average, it would not guarantee that daily control totals would be maintained, and that the results would not be deterministic so model runs would not be reproducible. This approach is unacceptable for transport planning models and is also known to produce undesirable artifacts in image and audio processing, causing random noise (sometimes swamping actual data) especially in slowly varying, low-amplitude regions like intermediate ridership on the railroad's branches.

We therefore developed an algorithm for deciding when to "add one" that varies on a repeating, but very long cycle that also contains shorter cycles, but the two cycles are set up to drift in and out of phase. It is the mathematical equivalent of having a cam that rotates over a number of followers, but the frame itself is also rotating much more slowly, such that the entire assembly does not end up in the same position until the lowest common multiple of the two cycles are reached. The shorter cycles guarantee that data artifacts would be evenly distributed between destinations within each origin market,

whereas the longer cycles ensure that artifacts are evenly distributed between days and months across each OD market. To accomplish this, we chose prime numbers to drive divisors in establishing when "add one" occurs within each 1.000000 of accumulated errors, and the probability of adding one remains proportional to each market's contribution to cumulative error.

```
CulErr = 0; LastCulErr = 0;
Param2 = 0; Param2Cycle = 67;            //Prime no
Param4 = 0; Param4Cycle = 1009;          //Prime no
Param3Date = ((RunDate Mod 7) / 7 +
   (RunDate Mod 31) / 31 + (RunDate Mod 365) / 365);
Param3 = Param3Date - Int(Param3Date);
While OD_hr_Markets_Remaining do {
   IntTrips = Int(EstTrips);
   ResidualTrips = EstTrips - IntTrips;
   CulErr = CulErr + ResidualTrips;

   Param2 = (Int(CulErr) Mod Param2Cycle) / Param2Cycle;
   Param4 = (Int(CulErr) Mod Param4Cycle) / Param4Cycle;
   SumParams = Param2 + Param3 + Param4;
   Threshold = SumParams - Int(SumParams);
   If ((LastCulErr < Threshold + Int(LastCulErr)) and
       (CulErr >= Threshold + Int(LastCulErr))) or
      ((LastCulErr < Threshold + Int(CulErr)) and
       (CulErr >= Threshold + Int(CulErr))) Then
      NewTrips = IntTrips + 1
   Else
      NewTrips = IntTrips;
   If NewTrips <> 0 Then Writeout(NewTrips, database);

   LastCulErr = CulErr;   RetrieveNextRecord;
}
```

FIGURE 11: PSEUDOCODE FOR QUANTIZING PROBABILISTIC PASSENGER COUNT DATA

As this is the most complex aspect of this algorithm, the pseudocode necessary to implement the dithering algorithm is provided in Figure 11. This algorithm could be applied to any set of sequential transport planning data that needs to be quantized while distributing truncation errors to nearby buckets in a predictable and proportionate-probability fashion.

2.11 Directionality Issue

After full implementation, we discovered a directionality issue. Multi-ride tickets and passes on commuter rail are sold such that they are valid for travel in either direction between two defined points. Passengers therefore tend to flip origins and destinations whimsically. On a daily level, the issue washes out on average because commuter traffic tends to be directionally balanced. On an hourly level, however, e.g. during the morning peak, a suburban station such as Irvington may originate 95% of the total suburb-CBD traffic, and terminate 5% (due to the proximity of a small business park). But about 60% those travellers held nominally inbound tickets, whilst the other 40% held outbound tickets. Fractions of inbound and outbound passengers by hour must thus be determined independently of origin and destination stations shown on the ticket.

Luckily, due to a ridership census (see Section 3.6 below), a dataset existed for the target system that provided boarding, disembarkation, and leave load counts for each scheduled train. This data was not easy to collect, as it involved stationing one surveyor per coach to check ride the entire trip gathering data for all 800 daily train starts, separately for weekday, Saturday, and Sunday. The entire data collection effort took place over three calendar years.

This data is fed into a standard "frataring" (iterative proportional fitting, IPF, see e.g. [9]) algorithm to synthesize a directionally-correct train-level OD matrix (albeit "smeared" over the three-year data collection period, during which many schedule adjustments took place), summarized by origin-destination and hour, then expressed as an inbound/outbound fraction. This fraction (where it exists) is then used allocate the observed passengers in the main OD model output, in each origin-destination market for each hour. This process preserves daily passenger-count and ticket-type information by hour from the ticket data, but reallocates the fraction of inbound/outbound passengers by hour such that directionality is more accurate, applying only to directionally-ambiguous ticket types. A summary is provided in Figure 12, represented by the "Fratared Onffs by Hr" canister in Figure 9.

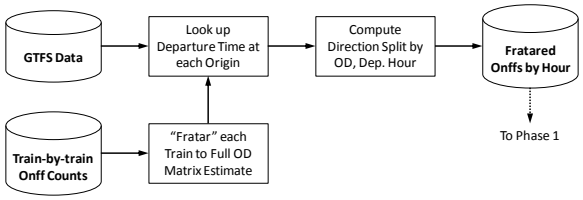

FIGURE 12: PHASE IA BLOCK DIAGRAM SHOWING HOW TRAIN-BY-TRAIN ONFF COUNTS ARE USED TO INFORM DIRECTIONALITY OF TICKET-SALES INFERRED TRAFFIC (GTFS = GENERAL TRANSIT FEED SPECIFICATION)

3. RESULTS AND DISCUSSION

3.1 Practical Implementation

We implemented this model in Microsoft Access 2000 using a combination of the Jet database engine and the built-in Microsoft Visual Basic for Applications (VBA) compiler. There were obviously many tools for this task that may have been better (e.g. Oracle database, Python script, "R" analytics software, SAS, or cloud-based tools), however, newer tools either took too long to set up (because it would involve the I.T. department) or would have a learning curve (due to existing skillsets of current personnel), that we decided it was faster to work with older tools that we know we could work with reliably and accurately. When developing this type of algorithms, it is more important not having to doubt and double-check that the tool is performing your intended commands, than to have a theoretically elegant solution that may require experimenting with the tool as opposed to focusing on the algorithm.

One of the practical limits with this toolset is that largest table that could be handled is 2 GB in size. This means eTix sales, activation, and ticket vending machine data must be exported from the enterprise database in two-month chunks. This was not a terrible handicap, as we simply read in relevant datasets from multiple files during the first stage of processing. This has the advantage of being able to discard in advance (with program code) data records not required during the current stage of processing, speeding up query execution in the more complex parts of the model. For instance, the ten-trip

model pre-processor must read in 92 days' worth of contiguous ticket sales data from numerous source files, but discard all sales transaction that doesn't concern ten-trip tickets, dramatically reducing the data from 4.5 million records (per quarter) to ~85,000 records. This is a common technique in big data processing [10] from an era when modern computer tools were not widespread.

3.2 Execution Time Performance

This model runs in three stages: (1) compute directionality factors, i.e. "frataring"; (2) summarize required information from eTix sales and activation data into distributions used to drive the model, called the "calibration stage"; (3) apply distributions to synthesize an OD matrix for a given day, termed the "execution stage". Table 1 below shows the time required to run all of 2019's data.

Phase	Model Step	Run Time	Runs Req'd
Frataring	Directionality	1 Min 30 Secs	1
Calibration	Round-Trip	3 Minutes	16
	Ten Trip	15 Seconds	12
	Monthly	3 Mins 30 Secs	18
	Weekly	5 Seconds	18
Execution	Pre-Processing	2 Minutes	365
	Single/Return	4 Minutes	365
	Ten Trip	1 Min 15 Sec	365
	Monthly	5 Seconds	365
	Weekly	3 Seconds	365
	Mobile Ticket	45 Seconds	365
	Dithering	20 Seconds	365

TABLE 1: MODEL RUN-TIME PERFORMANCE STATISTICS

Execution time was timed on a Core 2 Q8300 at 2.5 GHz and 8 GB of RAM (a twelve-year-old computer), connecting to local databases residing on an external USB hard drive. Execution time is actually a critically important part of any transportation model's performance, particularly in applications involving big data. If model calculations are too slow, planners will not be able to do scenario analysis and will simply not use it to inform decision making, which is not useful.

Total time requirements of 10 minutes per day's worth of data after calibration implies the entire year could be run in approximately 60 hours. Whilst slower than ideal, it is well within the range that model outputs can be considered useful. Some models take many hours to run a single day's worth of data, which would be far less useful.

3.3 Sample Results

OD matrices can be difficult to present in tabular form. Classic output shows origins on one axis and destinations on the other, looking somewhat like mileage tables appended to old-fashioned highway atlases, or fare tables in commuter rail timetables. Representing an hourly OD matrix is nearly impossible, as it would require three basic dimensions (O, D, and hour) to be displayed simultaneously. This could be done as computer animations cycling data through each hour, but that might not be analytically fruitful. Summary views would have to be developed from this database based on service planning questions being asked.

Figure 14 shows the whole-day OD matrix for the Hudson, Harlem, and New Haven Divisions (including Branches) in a colour-coded way, with station codes along both axes running from south to north then onto branches, left-to-right and top-to-bottom. The sheer domination of Grand Central Terminal is readily apparent; however, the Hudson Line graphic also shows two stations of secondary importance (Yonkers and Marble Hill) that connects strongly with virtually every other station on the line; tellingly, those were stations that received off-peak diesel express service during 1994~2012. This is the sort of insights that inform service planning when drilled down into the hourly level whilst contemplating which trains could make additional stops to provide more journey opportunities where demand exists in temporal and geographic space, even if it doesn't necessarily tell us the causal directionality.

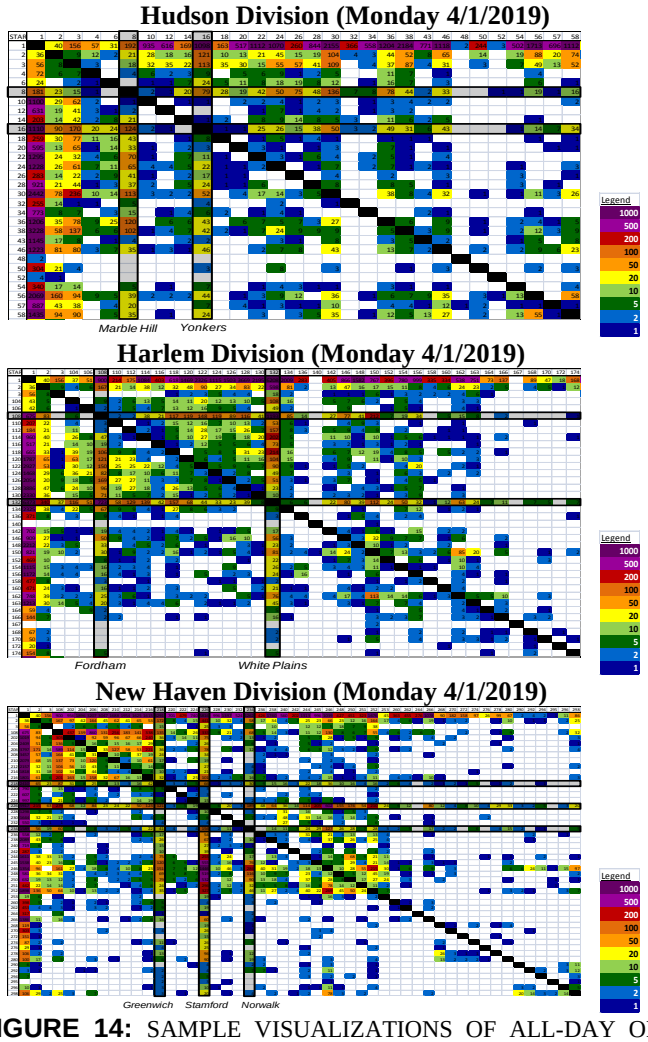

FIGURE 14: SAMPLE VISUALIZATIONS OF ALL-DAY OD MATRICES FROM TICKET RIDERSHIP ESTIMATION MODEL

The Harlem Line graphic reveals suburban hubs at Fordham and White Plains. Melrose appears strongly connected with stations south of White Plains but not well

connected with others. This could be a function of the existing service plan or an indicator of underlying travel demand.

On the New Haven Line, Stamford, Greenwich, Norwalk, and Fordham show strong connections with all other stations, affirming our understanding of current travel patterns.

3.4 Marginal Summaries

One way of visualizing and utilizing OD data is to flow it over a network model to determine link-loads and use those "load profiles" to make service decisions. At this time, the target system does not have a network model, but it is nonetheless possible to understand passenger loads along a line (although not a specific service route) by sorting results by station location, then integrating boardings and disembarkations at each station to determine onboard loads. This analysis is sometimes termed "marginal" analysis because it requires summing data both down and across the OD matrix, with the results often shown in margins of the table.

Figure 15 shows load profiles of the three lines attributable to those holding single and return tickets. The peak load points in those cases were not actually near the central business district, because non-CBD ridership tends to heavily favour pay-per-ride tickets. When these results are combined with data from the unlimited ride ticket model, the peak load point shifts substantially towards the CBD.

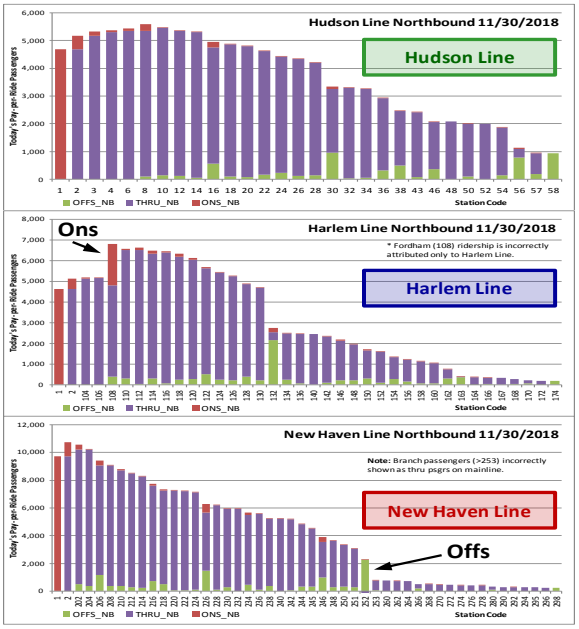

FIGURE 15: NORTHBOUND PAY-PER-RIDE PAPER TICKET LOAD PROFILES FOR ALL THREE LINES

3.5 Graphic Representation

One original motivation of developing this model is to provide a graphical representation of OD travel demand on the target system. Figure 16 shows a map-based representation of the OD matrix. Because the travel pattern is so dominated by travel to CBD, this visualization turned out not to be analytically informative unless filters were implemented. We are currently working with a vendor to develop a web application that would allow users to select how they want to visualize this data, which would fully unlock the analytical value of these ridership estimates.

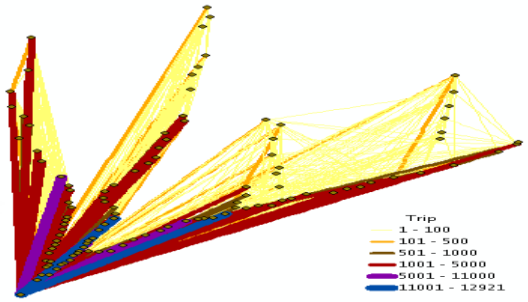

FIGURE 16: MAP OF THE SERVICE TERRITORY SHOWING RELATIVE DEMAND DENSITIES OF ALL OD PAIRS

3.6 Model Verification and Validation

Validating this model, like other big data analytical results, is practically impossible. In theory, results derived from big data sources should be 100% accurate provided the analytical algorithm is implemented correctly. However, many issues can arise in practice, including corrupted and missing data (often due to field communication issues with data gathering devices), data misinterpretation, etc. Data gathered from transport systems typically must travel over code lines (or ethernet, or wireless networks) subject to all sorts of weather-related disruptions; systems often have different designs for dealing with code line down conditions and can error-correct or re-transmit to different extents. Ticket sales data are amongst the most reliable field-collected data because they involve financial transactions where errors are not tolerated by either the company or customers.

The type of algorithm described here is particularly problematic because it relies on estimation to deal with information that the equipment does not collect. Short of manually collecting a very large sample (which is impractical, and in any case would be subject to data collection errors), there is basically no way to verify the outputs of this model.

We came up with two different approaches to validate the model to some extent. The system had performed an origin-destination survey (perhaps better termed a "census") several years ago with more than 100,000 respondents and a 40% total population response rate [11]. Comparing model results with the survey should provide some level of assurance. Although this sample size sounds very impressive, the reality in commuter rail is that CBD-based travel so dominates the traffic pattern that 5% of total ridership accounts for the bottom 72% of OD markets (in this case, each market accounts for fewer than 50 daily passengers). Attempting to ascertain accurate ridership counts in these markets using a sample survey methodology is practically impossible.

Figure 17 shows the comparison between survey data and four days' worth of model output. The model is fairly internally consistent, but shows some level of deviation from survey data. On average, the survey reports ridership that is 5%

higher, but there are some notable exceptions. Nonetheless, the curve fit shows very good correlation with R^2 values in excess of 0.93 in all cases. We believe some visible deviations are due to errors in the survey, rather than issues with the model. However, based on available data, there can be no definitive conclusion one way or another.

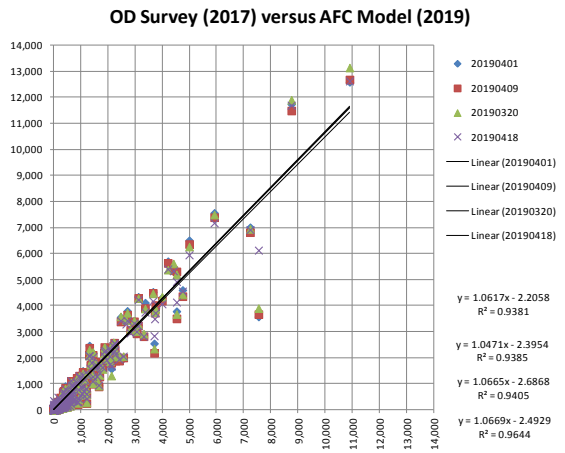

FIGURE 17: CORRELATIONAL ANALYSIS OF OD SURVEY DATA WITH TICKET RIDERSHIP-GENERATED OD MATRICES

Another method of validating the model is to compare it with "official" ticket count data. This approach will not detect ticket data integrity issues, but could flag logical errors in the model algorithm if they exist.

Ticket Type	AFC Model	"Official"	Difference
Single/Return	2,331,589	2,457,162	5.1%
Ten Trip	558,579	569,460	1.9%
Monthly	3,266,408	4,020,667	18.8%
Weekly	168,012	213,100	21.2%
Total	**6,324,588**	**7,260,389**	**12.9%**

TABLE 2: COMPARISON OF TICKET RIDERSHIP ESTIMATE MODEL VERSUS "OFFICIAL" RIDERSHIP COUNT

Table 2 shows April 2019 total passenger counts from our model compared to an "official" count. The official counts explicitly assume each monthly ticket is used for exactly 40 trips, but our model suggests that number is closer to 33 trips. We had recently conducted a passenger survey that suggested monthly riders telecommuted on average twice a month (in the pre-COVID condition), implying monthly ticket utilizations closer to 36 trips. We also had other information suggesting that somewhere around 2~5% of ticket activations may never reach the server, due to intermittent communication issues from user mobile devices, particularly affecting users of unlimited ride tickets. The official ridership is inferred from ticket sales alone, which may have its own sources of errors. For purposes of ridership estimation, we consider these differences to be wholly acceptable. If we are eventually able to determine the sources of errors and measure their impacts, we can apply correction factors to the results, perhaps to both the model's output and official ridership statistics.

The discrepancy between the model and "official" count on pay-per-ride tickets is more problematic. In theory, each ticket counts for one transaction/activation, and totals from both sources should tie out. We are continuing to investigate this; we are currently unable to rule out either duplicates in official counts, or missing data from our download. Importantly, these issues relate to practical implementation rather than defects in the theoretical concept or algorithm.

3.7 Applying Model to COVID Ridership Estimation

The design of this model utilizes the pattern of eTix user behaviour to estimate paper ticket usage. Therefore, when a cliff-edge condition arises in travel demand, such as that occurred on March 15, 2020 when we entered a "New York State on Pause" COVID19 induced lockdown, this assumption no longer held true. As nonessential employees began a prolonged period of working-from-home, rider behaviour prior to that date held little relevance to ticket usage after. Thus, the model calibration must be "flushed" and the model re-calibrated for post-COVID travel conditions.

Due to the high level of granularity required of distributions used to estimate paper ticket passenger behaviour, and sharply reduced travel due to the lockdown, we thought conservatively that a six-month post-COVID sample period for observations (i.e. one pre-COVID month's equivalent volume) of eTix activation data would be required before we could generate meaningful distributions for application to paper ticket sales data. Whilst travel behaviour is continuing to evolve as parts of the economy is reopened, other indicators (including "official" ridership statistics, and turnstile utilization on the subway system) has indicated that ridership has stabilized at about 20% of pre-COVID levels. The hour-by-hour ridership estimates are particularly critical for COVID response efforts because it gives us an early indication where potential exists for onboard social distancing to approach guideline capacity limits.

We have retrieved all post-lockdown source data from the date range April 1 through September 30, 2020 and are currently working on re-calibrating the model using the post-COVID ridership patterns.

3.8 Connecting to Network Model for Arc-Loads

Our next step in the development of this model is to connect the OD matrix, which is necessarily a point-to-point representation of travel demand, with electronic train schedule data now available, to essentially flow the traffic over the network using a methodology similar to [12,12]. Doing this accurately is particularly important in a commuter rail setting because of the preponderance of skip-stop and zone-express services. Figure 18 shows the proposed process, where it is necessary to calibrate a "train choice" model based on journey time, headway between trains, and transfers (see, e.g. [14]). We are currently using a working draft model that distributes the hourly traffic based on elapsed minutes between successive departures that provide service to the specific destination, then

applying a positive or negative bias based on journey time relative to the daily average within that OD market.

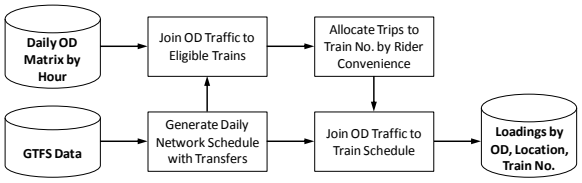

FIGURE 18: PHASE II BLOCK DIAGRAM SHOWING THE PROCESS TO ALLOCATE HOURLY TRAFFIC TO TRAIN STARTS AND THUS DETERMINE LINK-LOADS ON EACH TRIP

The OD matrix provides a view of emerging markets that could be better served with skip-stop services, perhaps by adding a stop to existing non-stop zonal expresses, or by removing train-stops that do not serve any markets effectively. Matching OD traffic to service arcs would allow us to understand if the current service patterns are effective in serving or creating that travel demand. It would also allow us to use data to inform day-to-day scheduling decisions such as number of railcars required on each zone-express train.

3.9 Implication of Loadweigh and Camera Count Data Availability

The target system recently announced [15] that approximately one-third of the electric railcar fleet have recently been modified to report continuously in real-time, via a wireless modem, the pressure required to inflate air suspension to a set level, which is an approximate measure of laden weight utilized by onboard systems to compute brake force required to decelerate the train at a specified rate. This data has been utilized to infer passenger occupancy on a coach-level in real-time, but the error margins are significant and may require frequent re-calibrations.

Due to the inherent limitations of this approach, including the inapplicability to non-EPB (electro-pneumatically braked) rolling stock, work is currently ongoing (by others) to use computer vision algorithms to process image data gathered from the ten onboard security video cameras in each carriage to literally and automatically count passengers in real-time. This data can be algorithmically combined with the loadweigh data to produce the most accurate loading estimates.

When complete, these direct observations will be the best data on coach occupancies, and when combined with consist information, excellent daily train-by-train loading data. However, they provide no market intelligence in terms of customers' ODs, transfers, ticket types, nights' stay, repeat system usage, trip purpose, or passengers travelling together. Ticket data continue to be an important source of market information, although their role in inferring train loadings will necessarily become more limited. We envision the current algorithm will be helpful to those railroads having advanced ticketing systems, but chose not to install onboard cameras with 100% coverage for other reasons.

4. CONCLUSION

We described a novel method to estimate commuter rail station-to-station OD matrix at an hourly level of granularity, separately and independently for each day, using traditional ticket sales data, and usage data from electronic tickets. We allocated the traffic to each train-start using a train-choice model and determined the correct direction for multi-ride tickets utilizing historical ridecheck data. The basic idea is fairly straightforward: distributions of observable patterns are used to model unobservable ones. Practical interpretations in choosing variables, mathematically describing likely customer behaviour for each ticket type, converting probabilities into whole riders, and relating ridership patterns to subtle but ever-present minor schedule adjustments, are somewhat more complicated. We hope that the thought processes outlined here contributes to the transport modelling community in demonstrating a fairly complex case of analysis applied to a niche market. Business practices specific to commuter rail are generally not well understood outside of specialized practitioners, and we hope to shed some light for those who are not specialists in this classic transport mode. This model can estimate post-COVID ridership once sufficient sample of travel habits are collected; necessary model re-calibration work is currently in progress.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the contributions of the following members of the Metro-North family to making this research possible: C. Planck, C. Roberts, L. Rodriguez, S. Hawkins, M. Collins, Z. Zhong, F. Lennon, J. Fiegerman, V. Jones, K. Smith, D. Fogel, J. McGovern, T. Pawlowski, S. Marra, S. Astacio, S. Chidambaranathan, R. Marino, M. Goulard, P. Donnelly, J.E. Kennard, M.J. Shiffer, and J.E. Kesich. This work was internally funded by Metro-North Commuter Railroad. Responsibility for errors or omissions remains with the authors. The opinions expressed or implied are the authors' and do not necessarily reflect official policy or positions of the New York State Metropolitan Transportation Authority.

REFERENCES

[1] Barry, J.J., Newhouser, R., et al., 2002, "Origin and Destination Estimation in New York City with Automated Fare System Data," *J Transp. Res. Board*, Issue 1817, pp. 183–187.

[2] Lai, Y.-C. (R.), Huang, C.-W., and Hsu, Y.-T., 2018, "Estimation of Rail Passenger Flow and System Utilization with Ticket Transaction and Gate Data," *Trans. Plan. & Tech.*, **41** (7), pp. 752-778. doi:10.1080/03081060.2018.1504184

[3] Wang, W., 2010, "Bus Passenger Origin-Destination Estimation and Travel Behavior Using Automated Data Collection Systems in London, U.K.," M.S.T. Thesis, Massachusetts Institute of Technology.

[4] Ro, W.Y. (羅惟元), 2008, "Using Taipei EasyCard Transaction Data to Explore the O-D Table of Bus Passengers," Masters Thesis, Graduate Institute of Transportation Management, Tamkang University, Damtsui, Taiwan.

[5] Zhao, J., Rahbee, A., and Wilson, N.H.M., 2007, "Estimating a Rail Passenger Trip Origin-Destination Matrix Using Automatic Data Collection Systems," *Comp.-Aided Civ. and Infra. Eng.*, No. 22, pp. 376–387.

[6] Lu, A. and Reddy, A.V., 2012, "Strategic Look at Friday Exceptions in Weekday Schedules for Urban Transit," *J Transp. Res. Board*, Issue 2274, pp 30–51.

[7] Hickey, R., 2005, "Impact of Transit Fare Increase on Ridership and Revenue: MTA, New York City," *J Transp. Res. Board*, Issue 1927, 2005, pp 239–248.

[8] Roberts, L.G., 1961, "Picture Coding Using Pseudo-Random Noise," S.M. Thesis, Masschusetts Inst. of Tech.

[9] Deming, W. E., Stephan, F. F., 1940. "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known," *Annals of Mathematical Statistics*, **11** (4), pp 427–444. doi:10.1214/aoms/1177731829.

[10] Lu, A. and Reddy, A.V., 2011, "Algorithm to Measure Daily Bus Passenger Miles Using Electronic Farebox Data," *J Transp. Res. Board*, Issue 2216, pp 19–32.

[11] "2017 Metro-North Railroad Customer Origin-Destination Survey," accessed October 9, 2020, http://web.mta.info/mta/planning/data.html .

[12] Zeng, Q.F., Reddy, A.V., Lu, A., and Levine, B., 2015, "Development of Application for Estimating Daily Boarding and Alighting Counts on NYC Buses," *J Transp. Res. Board*, Issue 2535, pp 1–14.

[12] Stasko, T., Levine, B., Reddy, A.V., 2016, "Time-Expanded Network Model of Train-Level Subway Ridership Flows Using Actual Train Movement Data," *J Transp. Res. Board*, Issue 2540, pp 92–101.

[14] Slagmolen, M. H., 1980. "*Train Choice: Measurement of the Time-Table Quality of Rail Services Based on an Analysis of Train Choice Behaviour*," Ph.D. Dissertation, University of Rotterdam.

[15] Metropolitan Transportation Authority, 2020. "*MTA Unveils New Capacity Tracking and Real-Time Location Features in Metro-North TrainTime App*," Press Release, retrieved from http://www.mta.info/press-release/metro-north/mta-unveils-new-capacity-tracking-and-real-time-location-features-metro on November 24, 2020.