ASME® 2021 Joint Rail Conference

RTD and JRC: 100 Years of Tracking and Promoting Railroading Progress

VIRTUAL CONFERENCE
APRIL 20–21, 2021

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, Thomas Marchwinski, Robert Culhane, and Xiaojing Wei

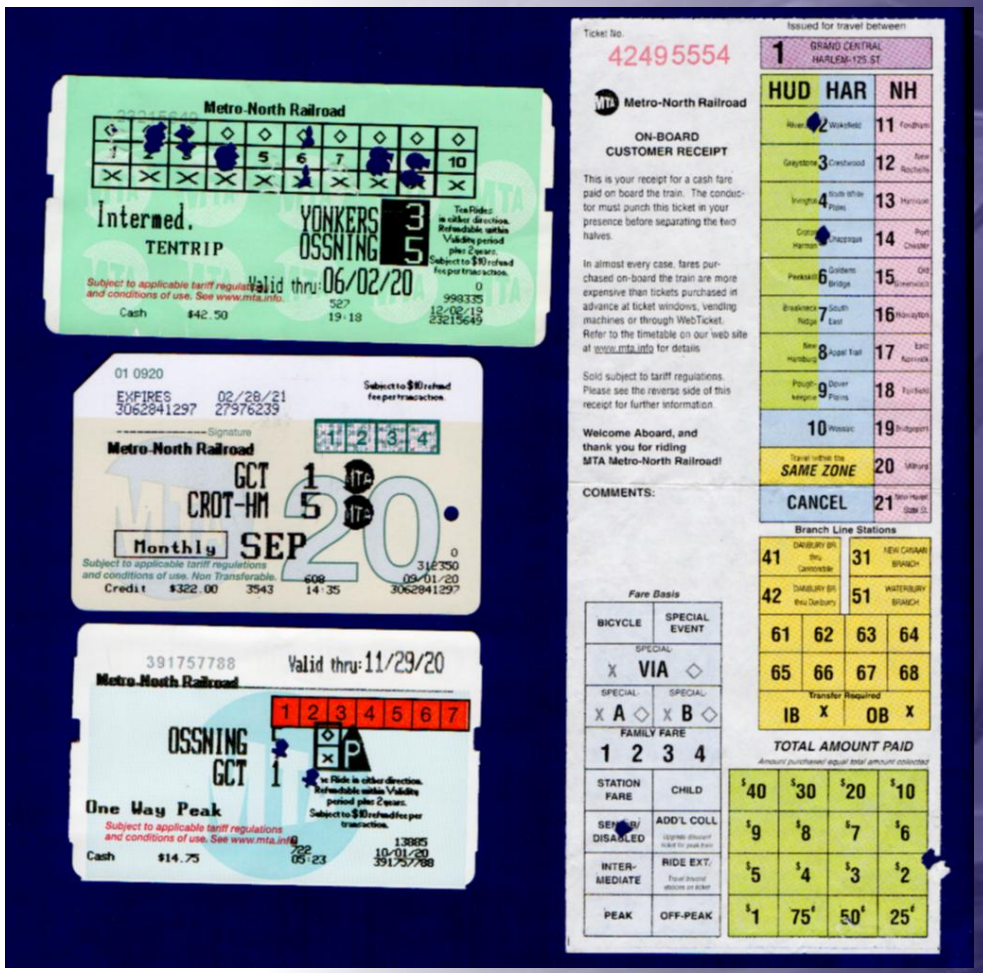
Metro-North Commuter Railroad

Introduction

- Method to estimate commuter rail station-to-station origin-destination (OD) matrix at hourly level, separately and independently for each day. Features:
  1. Handles multi-pack pay-per-ride fare instruments not requiring electronic validation
  2. Infers directionality for direction-agnostic ticket-types
  3. Sensitive to day-to-day changes in travel conditions (e.g. weather, special events, etc.)
  4. Deals with a cliff-edge sudden change in demand (e.g. COVID-19 lockdown)
  5. Estimate utilization patterns of unlimited-ride tickets
  6. Provides output in terms of whole numbers of passengers
  7. Allocates hourly traffic to each train-start (i.e. schedule number)

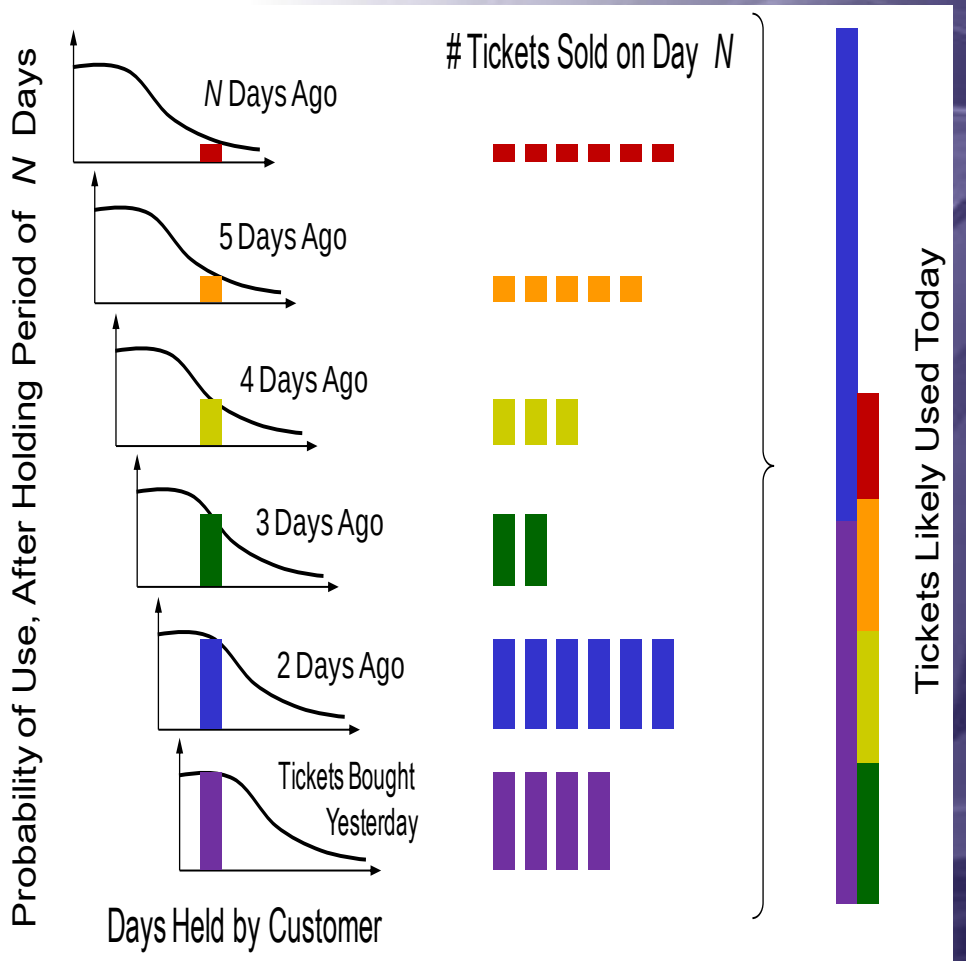
Data Description

- Multiple sales channels w/ different fare media and systems:
  1. Self-service ticket vending machines
  2. Mobile tickets (eTix)
  3. Mail order tickets
  4. Onboard ticket sales
- More than sixty fare types:
  1. Only four basic categories necessary:
     a) monthly commutation,
     b) weekly commutation,
     c) ten-trip, and
     d) à-la-carte single/return tickets.
  2. Pay-per-Ride (PPR) tickets are further subdivided into peak, off-peak, intermediate, child, senior, family fare, and special discount schemes



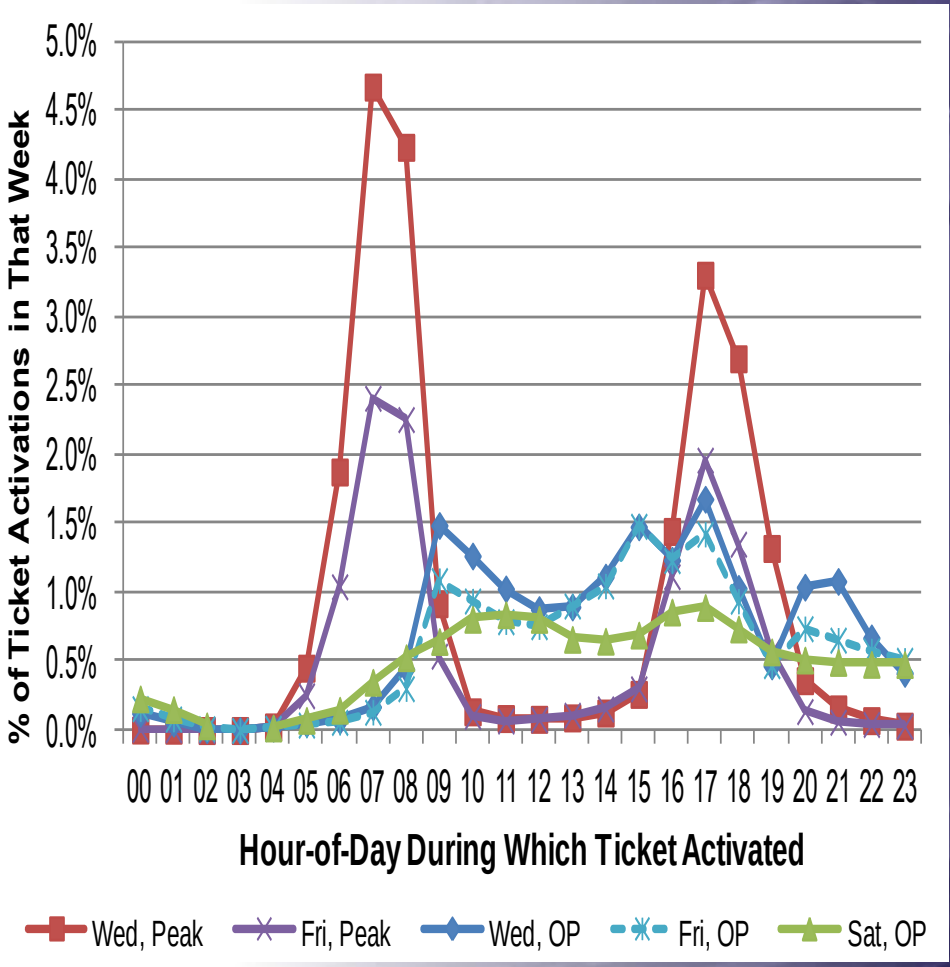
"Lookback Windows"

- Key Assumptions:
  1. eTix passengers use their tickets approximately the same way as paper ticket holders
  2. Paper singles purchased immediately prior to use
- PPR tickets are inventories of trip-coupons held by customers, cancelled upon fulfillment of transportation:
  1. 90% of Ten Rides consumed within ten weeks
  2. 97.5% of Round-Trip return portions used within seven days
- eTix data provides distribution of "days each ticket is held by the customer"
- Number of rides taken today = sum product of {probability of usage after N days}, and {number of tickets sold exactly N days ago}



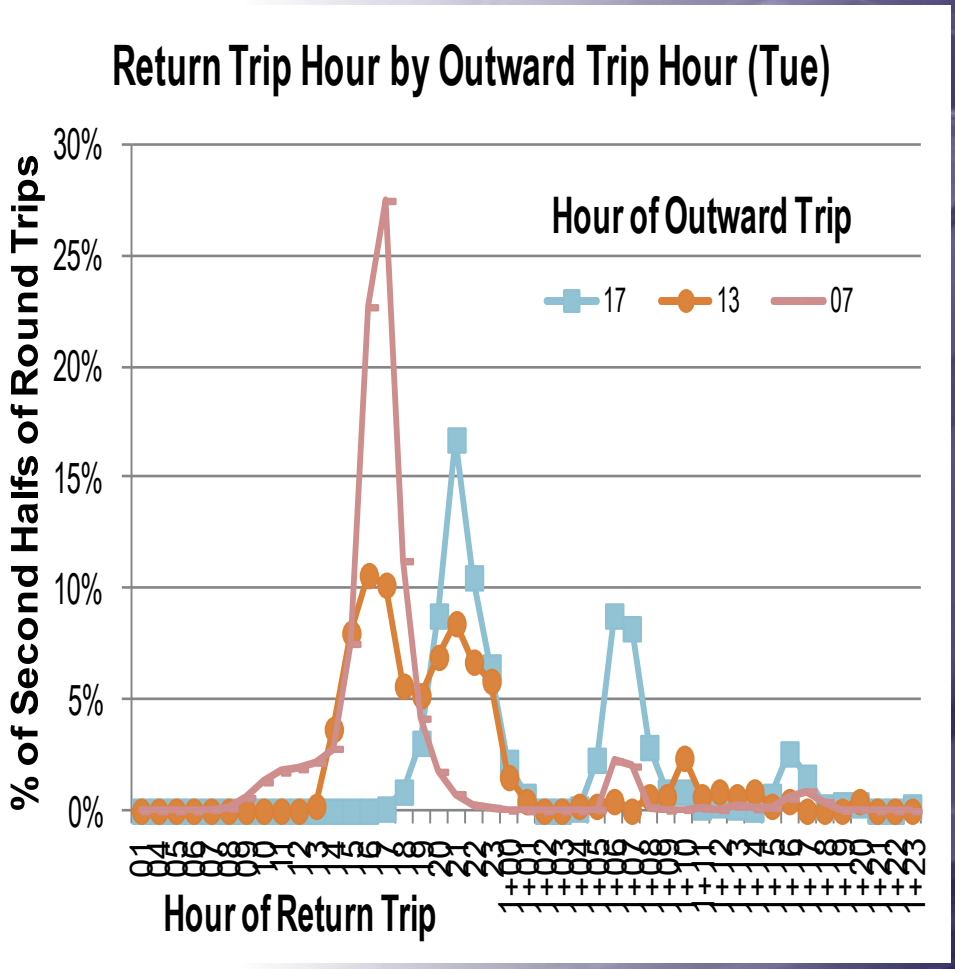
"Weeks Since Purchase" Module

- Ten-trips are utilized by customers having occasional needs to travel:
  1. Date and time of travel driven by customer business
  2. Not how long they have possessed fare media
  3. But customers try to use tickets sooner rather than later
- Compromise model:
  1. Compute paper tickets expected to be used during current week, using distribution of weeks since purchase
  2. Sprinkles rides this week using combined distribution of day-of-week and time-of-day, by ticket subtype
- Model does not use origin and destination stations:
  1. Time of travel is not significantly affected by geography



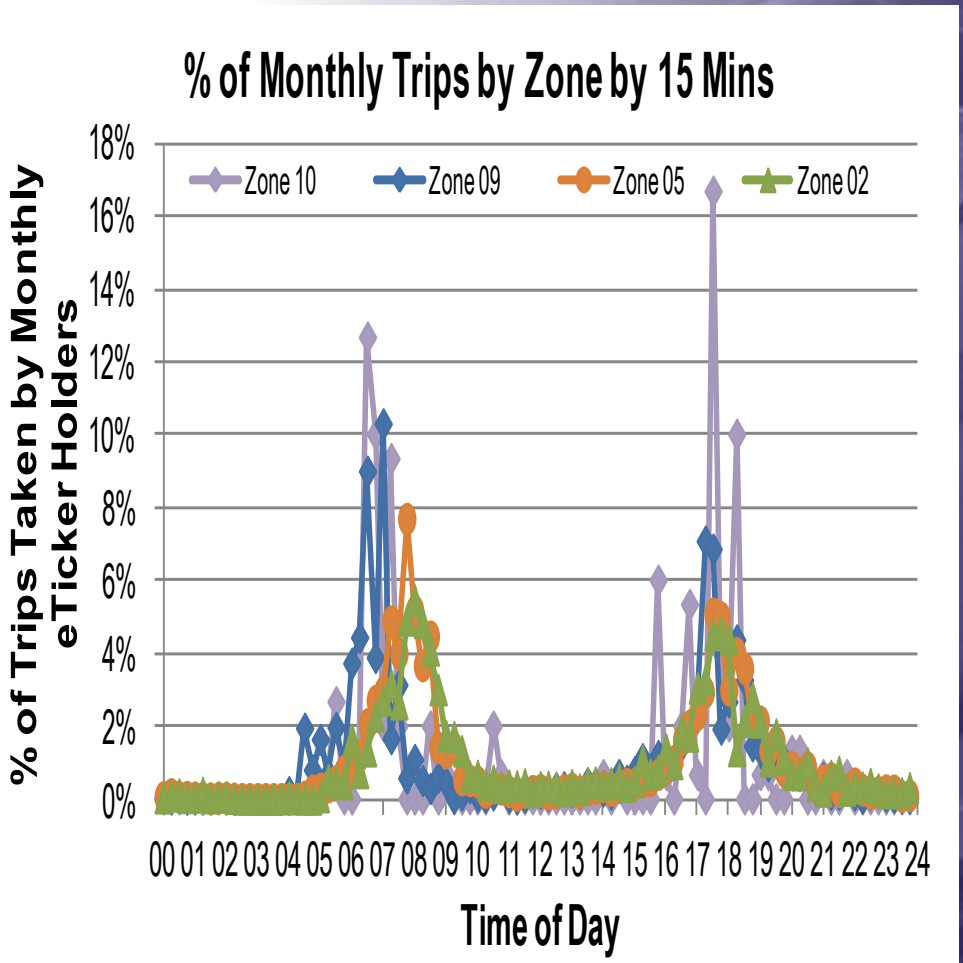
"48-Hour/Period Return" Module

- Round-trip passengers fall in two distinct markets:
  1. day-return market: leave early and come back late
  2. period-return market: two trips with independent departure times
- Travel patterns by outward trip hour:
  1. Held steady for Mondays through Thursdays
  2. Distinct pattern is seen each for Friday, Saturday, and Sunday.
  3. Peter out after about 48 hours.
- Segmented approach:
  1. Tickets bought today and yesterday (48-hour model):
     a) distribute return-time based on day-of-week and time-of-day when the ticket was sold
  2. Tickets purchased earlier than yesterday (period return model):
     a) apply basic "days away" logic by day-of-week to determine fraction of tickets used today
     b) sprinkle daily used return trip hours distribution of all return portions where passengers stayed for at least two days



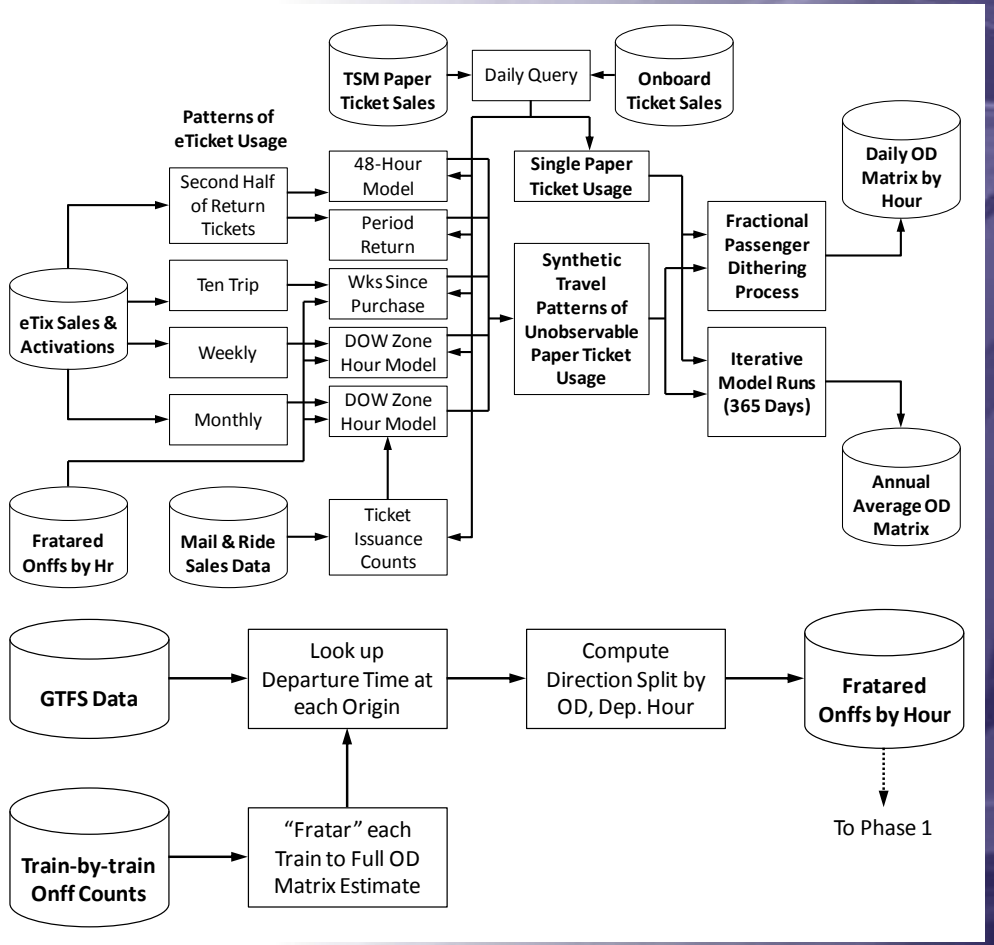
"Day-of-Week Zone Hour" Model

- Trends observed in the data:
  1. Geography (travel distance) does not affect unlimited-ride ticket utilization
  2. Number of weekdays and holidays in each month a significant driver of monthly ticket utilization
  3. Morning commutes begin earlier for those living further away from downtown
  4. On the system's extremities, afternoon trips are tied to specific train departures
- Two-stage problem:
  1. Given day-of-week and month (e.g. Friday in January), compute fraction of monthly tickets expected to be "seen"
  2. Given ticket is seen today, how many trips do we expect that ticket to redeem? Sprinkle rides amongst the 24-hour day based on relevant hourly distributions for that {day-of-week, and origin and destination fare zones}



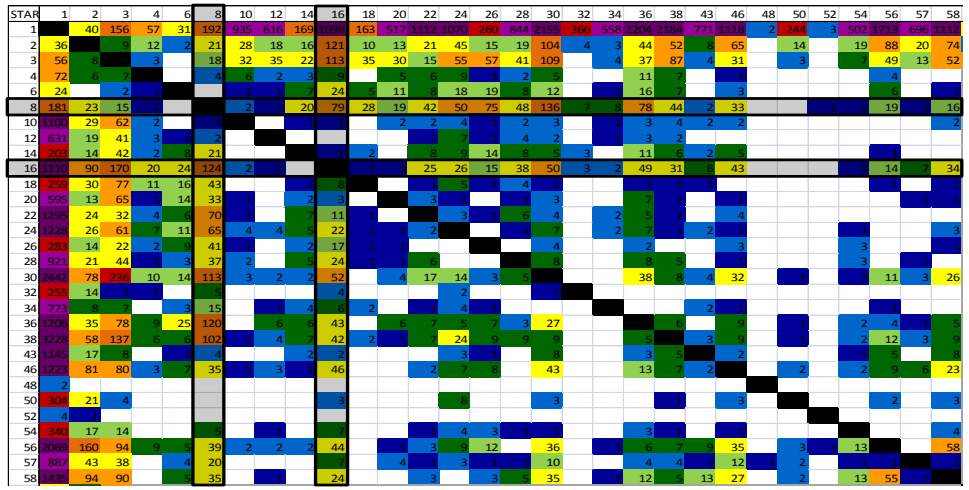
Other Required Processing

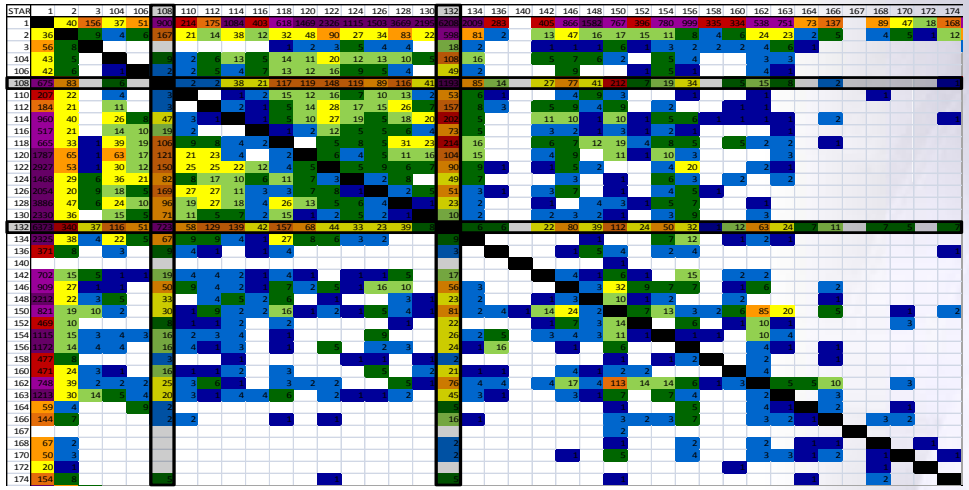
- Model Structure:
  1. eTix sales and activation data are used to generate distributions
  2. Separately, sales from various non-mobile channels were combined, summarized, and multiplied by these distributions
- "Fractional Passenger Dithering Process"
  1. Small ODs at unsociable hours have sparse demand; assigns probabilistically estimated marginal passenger to specific hour
- Directionality Issue:
  1. Multi-ride tickets and passes are valid for travel in either direction
  2. Rdership census data "frataring" (iterative proportional fitting) algorithm synthesizes directionally-correct train-level OD matrix
  3. Fraction used allocate observed passengers, in each origin-destination market for each hour:
     a) Preserves daily passenger-count and ticket-type information by hour
     b) Reallocates fraction of inbound/outbound passengers by hour

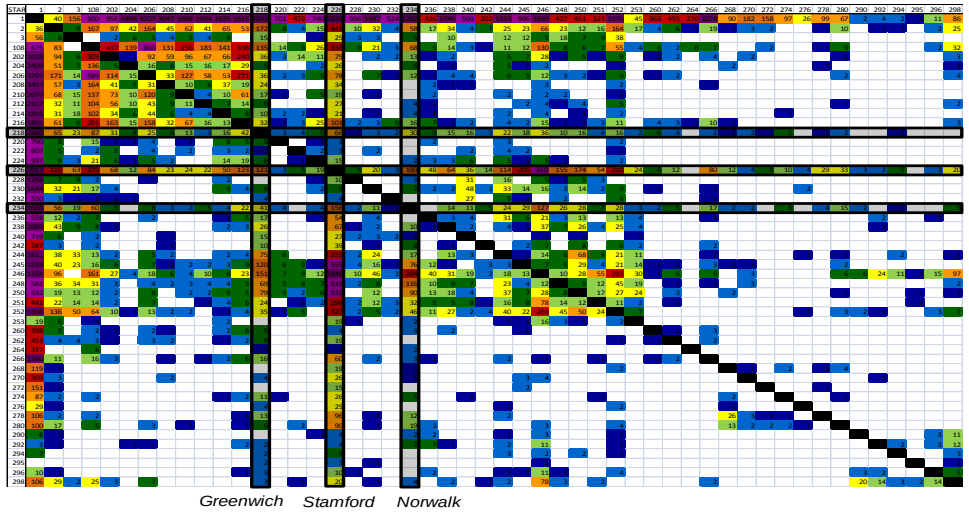


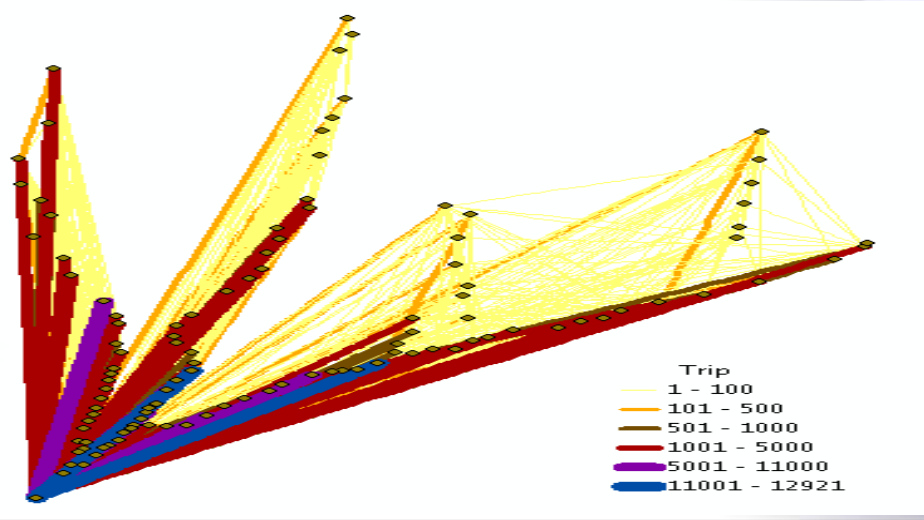
Origin-Destination Matrix Results

ASME® 2021 Joint Rail Conference
Virtual | April 20–21, 2021









Legend

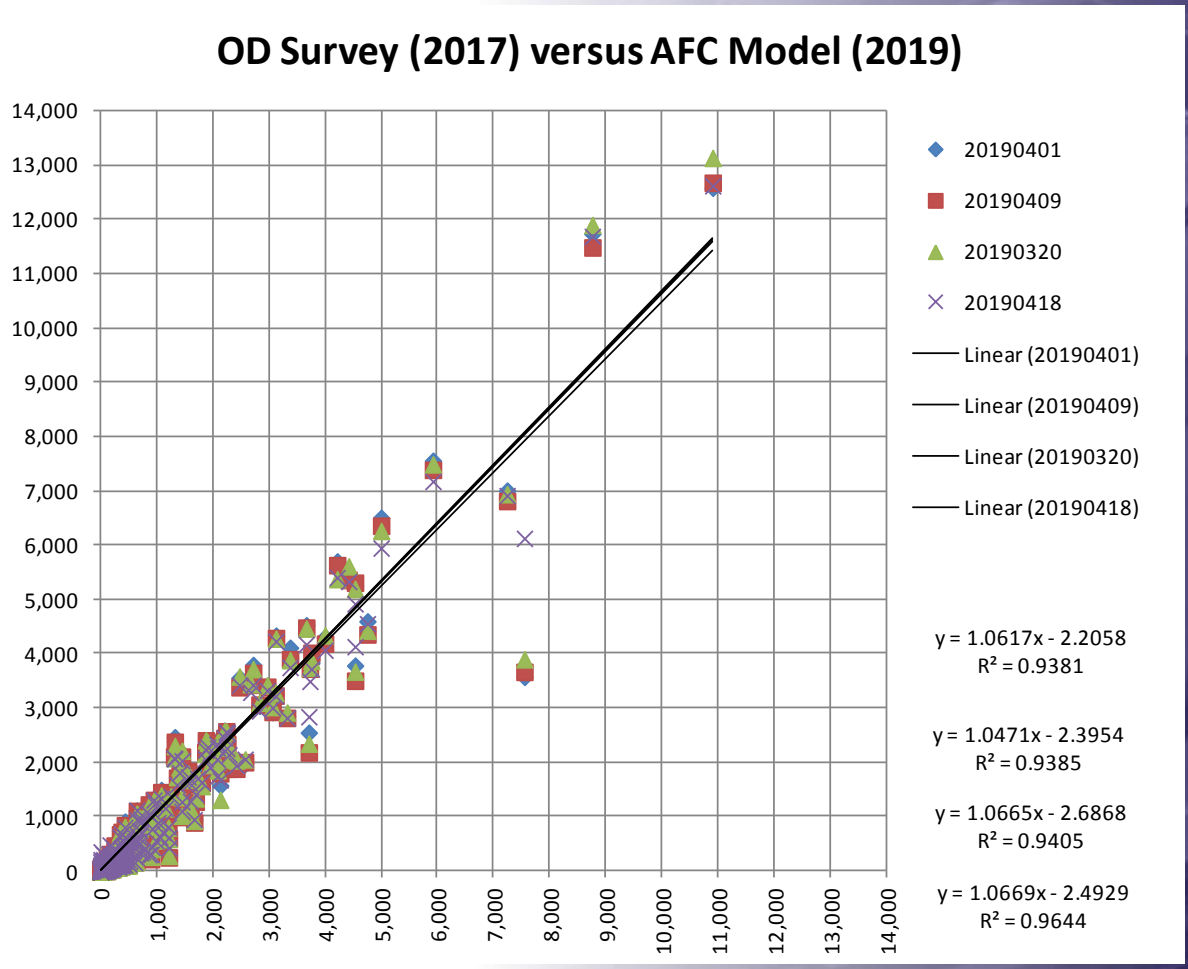
1000
500
200
100
50
20
10
5
2
1

Key

Hudson	Harlem
N.H.	System

Verification and Validation

- Validating virtually impossible (normal for "big data").
- This algorithm type particularly problematic:
  1. Estimated data (not collected by equipment)
  2. Large manual sample impractical
- Two approaches:
- Compare with ridership "Census":
  1. Model internally consistent
  2. Some deviation from survey data, but R^2>0.93
  3. Visible deviations could be survey error!
- Compare with "official" counts:
  1. Official counts higher, but assumptions differ



Conclusion

- Model design uses eTix behaviour pattern to estimate paper ticket usage
  - Calibration must be “flushed” for post-COVID travel conditions
- Next steps:
  1. Connect OD matrix to electronic train schedule data and flow traffic over network
  2. Approximately one-third of electric railcar fleet now fitted with airbag loadweigh sensors
  3. Ongoing work (by others) to use computer vision to count passengers in real time
  4. When complete, these direct observations will be the best data on coach occupancies
- Onboard “counts” provide no market intelligence, e.g.
  - Customers’ ODs, transfers, ticket types, nights’ stay, repeat system usage, trip purpose, or passengers travelling together
- Ticket data’s role in inferring train loadings will necessarily become more limited
- This algorithm useful to railroads having advanced ticketing systems, but chose not to install onboard cameras with 100% coverage

Thank you!

Alex Lu
lexcie@gmail.com

Thomas Marchwinski
thomas.marchwinski@outlook.com

Robert Culhane
rculhane@mnr.org

Xiaojing Wei
xiaojingwei@yahoo.com

Note: 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.