Estimating Passengers’ Path Choices Using Automated Collection Data in Urban Rail Systems

Zhenliang (Mike) Ma
Baichuan Mo, Haris N. Koutsopoulos, Yiwen Zhu, Yunqing Chen
Crowding

• Issues and challenges
  – Near capacity operations
  – Passengers’ denied boarding
  – Safety

• Passengers’ perception of crowding
  – In-vehicle: 1 min standing (up to 10 min) = 2.04 min uncrowded seating*

Building Blocks for Transit Data Analytics

**Monitoring**
- Customer experience
  - Denied boarding
  - Reliability
  - System performance
  - Crowding

**Behavior**
- Route choice
  - Habitual
  - Interventions
  - Disruptions
  - Information

**Prediction & Planning**
- Real-time prediction
  - Demand
  - Individual mobility
- Strategic planning
  - TDM
  - NPM (assignment)
  - Micro-simulation
Problem Definition

- General **methodology** to estimate the route choice fractions of OD pairs and discrete choice model parameters at different time periods considering crowding impacts using AVL and AFC data in the system.
Impacts of Crowding on Passengers

- Platform crowding / denied boarding
- Route choice behavior
Passenger OD Journey Time Distribution
Challenges

- A passenger with a long journey time
  - chose a longer route, or
  - denied boarding multiple times on a shorter route

Observed
### Overview: Denied Boarding Approaches

<table>
<thead>
<tr>
<th>Approach</th>
<th>Data</th>
<th>Level</th>
<th>Applications</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Passenger-to-Itinerary-Assignment</td>
<td>AFC (tap-in &amp; out) AVL Walk distance/speed</td>
<td>Station</td>
<td>Performance measurement</td>
<td>Needs access/egress time distributions Sensitive to model parameters Unsupervised learning 5-10 mins to run</td>
</tr>
<tr>
<td>(Zhu et.al. 2018)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Structured Mixture Model</td>
<td>AFC (tap-in &amp; out) AVL</td>
<td>Station</td>
<td>Performance measurement</td>
<td>No external data needed Unsupervised learning 2-5 seconds to run</td>
</tr>
<tr>
<td>(Ma et.al. 2019)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regression</td>
<td>AFC (tap-in) AVL Denied boarding observations</td>
<td>Station</td>
<td>Performance measurement Prediction</td>
<td>Requires observations of denied boarding for calibration Supervised learning</td>
</tr>
<tr>
<td>(Miller et.al. 2018)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Network Assignment</td>
<td>OD flows Path choice fractions AVL Capacity</td>
<td>Network</td>
<td>Performance measurement Planning</td>
<td>Applied at the network level Various crowding metrics Requires capacity</td>
</tr>
<tr>
<td>(Ma et.al. 2019)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Overview: Path Choice Approaches

<table>
<thead>
<tr>
<th>Model</th>
<th>Data</th>
<th>Applications</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AFC</td>
<td>AVL</td>
<td>Monitor</td>
</tr>
<tr>
<td>Sun et al (2016)</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>J. Zhao et al (2017)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sun et.al. (2015)</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Zhang et.al. (2018)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hörcher, et al. (2017)</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>OD Journey time based approach</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Network based approach</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
</tbody>
</table>
Denied Boarding Estimation

- AFC and AVL data for a given OD pair (with no transfer and no route choice).
- The origin is the station of interest for which denied boarding metrics will be estimated.

Ma et al., 2019
Path Choice Estimation: MLE Formulation

\[
\text{maximize } \beta \quad LL(x_1, x_2, \ldots, x_N; \beta, T) = \log \left( \prod_{n \in N} P(x_n; \beta, T) \right) \\
= \sum_{n \in N} \log \left( \sum_{m \in M_n} P(x_n|m; T) \times \pi_h^m \right)
\]

where:

\[
\pi_h^m : \text{Path fraction for route } m \text{ at time period } h \quad \pi_h^m = \frac{\exp(\beta z_h^m)}{\sum_{m' = 1}^M \exp(\beta z_h^{m'})}
\]

\[
x_n : \text{Observation of journey time of a passenger} \\
m : \text{Path } m \text{ for passenger } n \\
z_h^m : \text{Attribute vector for path } m \text{ at time period } h \text{ (e.g. travel time, crowd, etc.)} \\
T : \text{Train run times (i.e. arrival and departure time from platform)} \\
\beta = (\beta_1, \beta_2, \ldots, \beta_M) : \text{Unknown parameters to be estimated}
\]
Probabilistic Process

\[ P(x_n|m; T) \]
Time-Space Diagram and Itineraries

Entry gate  Origin platform  Transfer Platform  Destination Platform  Exit gate

Tap-in  Access time  Tap-in

Tap-in  Access time  Tap-in

Tap-out  Egress time  Tap-out

Line 1
- Train 1
- Train 2
- Train 3
- Train 4

Line 2
- Train 1
- Train 2
- Train 3
- Train 4
Likelihood of an Observation Given a Path

\[ P(x_n|m; T) = P(t_{in}, t_{out}|\text{path}) = \sum_{\text{itinerary}} P(t_{in}, t_{out}|i^{th} \text{ itinerary}) \]

\[
= \sum_{\text{Itinerary}} \left\{ P(\text{access}|i^{th} \text{ itinerary}) \times \left[ \prod_j P(\text{transfer}^j|i^{th} \text{ itinerary}) \right] \times P(\text{egress}|i^{th} \text{ itinerary}) \right\}
\]

The denied boarding, access time, transfer time and egress time distributions can be pre-calculated based on AFC and AVL data or from survey data. (Zhu, 2017)
Model Characteristics

\[
\text{maximize } \quad LL(x_1, x_2, \ldots, x_N; \mathbf{\beta}, T) = \log \left( \prod_{n \in N} P(x_n; \mathbf{\beta}, T) \right)
\]

\[
= \sum_{n \in N} \log \left( \sum_{m \in M_n} P(x_n | m; T) \times \pi_n^m \right)
\]

\[
= \sum_{n \in N} \log \left( \sum_{m \in M_n} c_{m,n} \frac{\exp(\mathbf{\beta} z_h^m)}{\sum_{m' \in M_n} \exp(\mathbf{\beta} z_h^{m'})} \right)
\]

\[
= \sum_{n \in N} \log \left( \frac{1}{\sum_{m' \in M_n} \exp(\mathbf{\beta} z_h^{m'})} \sum_{m \in M_n} c_{m,n} \exp(\mathbf{\beta} z_h^m) \right)
\]
Model Characteristics

\[
\sum_{n \in N} \log \left( \frac{1}{\sum_{m' \in M_n} \exp(\beta z_{h}^{m'})} \sum_{m \in M_n} c_{m,n} \exp(\beta z_{h}^{m'}) \right)
\]

\[
= \sum_{n \in N} -\log \left( \sum_{m' \in M_n} \exp(\beta z_{h}^{m'}) \right) + \log \left( \sum_{m \in M_n} c_{m,n} \exp(\beta z_{h}^{m'}) \right)
\]

-LogSumExp Function is concave
LogSumExp Function is convex

- This problem is known as DC (Difference of Convex) Programming
- DCP can be solved globally by methods such as branch and bound, which may be slow in practice. A locally optimal (approximate) solution can be found through the many techniques of general nonlinear optimization. (Shen et. al 2016)
Implementation

• Select OD pairs
• Prepare inputs
  • Denied boarding distribution (Ma et.al, 2019)
  • Approximated walking distance (station layout)
  • Walking speed distribution (Zhu et.al, 2017)
  • Path attributes (network)
• Solve the MLE problem
Case Study Using Synthetic Data

- **Network**
  - Walk distance 30-50 m (access, egress, transfer)

- **Train operations**
  - Headway 2 min
  - Headway deviation (-10, 10) sec
  - Travel time deviation (-20, 20) sec

- **Passengers**
  - Walk speed distribution, logn(0.1, 0.4)
    - Mean 1.2 m/s, std 0.25 m/s
  - Denied boarding at station 2
    - Distribution (0.2, 0.5, 0.3)
Setting

\[ V_i = \beta_1 In_{-}Veh_{-}Time_i + \beta_2 Out_{-}Veh_{-}Time_i + \beta_3 No\_of\_Transfer_i + \beta_4 Denied\_Waiting_i \]

<table>
<thead>
<tr>
<th>Variables</th>
<th>Beta values</th>
</tr>
</thead>
<tbody>
<tr>
<td>In_{-}Veh_{-}Time</td>
<td>-0.246</td>
</tr>
<tr>
<td>Out_{-}Veh_{-}Time</td>
<td>-0.439</td>
</tr>
<tr>
<td>No_of_Transfer</td>
<td>-1.707</td>
</tr>
<tr>
<td>Denied_Waiting</td>
<td>-0.480</td>
</tr>
</tbody>
</table>

9 OD Pairs: 
\{1, 5, 4\} -> \{3, 6, 7\}

True path share for OD pairs
Results: Denied Boarding Estimation

Denied boarding times

Probability

TRUE
ESTIMATE
### Results: Discrete Choice Model Parameters

<table>
<thead>
<tr>
<th>Variables</th>
<th>Beta values (true)</th>
<th>Beta values (estimated)</th>
</tr>
</thead>
<tbody>
<tr>
<td>In_Veh_Time</td>
<td>-0.246</td>
<td>-0.201</td>
</tr>
<tr>
<td>Out_Veh_Time</td>
<td>-0.439</td>
<td>-0.411</td>
</tr>
<tr>
<td>No_of_Transfers</td>
<td>-1.707</td>
<td>-1.548</td>
</tr>
<tr>
<td>Crowd_Waiting</td>
<td>-0.48</td>
<td>-0.525</td>
</tr>
</tbody>
</table>

**Bar Chart**

- **TRUE**
- **ESTIMATION**
Sensitivity Analysis: Parameter Initialization

Estimated parameters

Difference between initialization and true parameters

- beta_tt
- beta_ovt
- beta_not
- beta_cw
Sensitivity Analysis: Denied Boarding

<table>
<thead>
<tr>
<th>Denied</th>
<th>Set_1</th>
<th>Set_2</th>
<th>Set_3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.2</td>
<td>0.2</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0.5</td>
<td>0.6</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.3</td>
<td>0.2</td>
<td>0</td>
</tr>
</tbody>
</table>

Path_1 fraction vs OD pairs
Sensitivity Analysis: Walking Speed

<table>
<thead>
<tr>
<th>OD pairs</th>
<th>Patt_1 fraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1_3</td>
<td>0.80</td>
</tr>
<tr>
<td>5_3</td>
<td>1.00</td>
</tr>
<tr>
<td>4_3</td>
<td>0.40</td>
</tr>
<tr>
<td>1_6</td>
<td>0.30</td>
</tr>
<tr>
<td>5_6</td>
<td>0.60</td>
</tr>
<tr>
<td>4_6</td>
<td>0.50</td>
</tr>
<tr>
<td>1_7</td>
<td>0.20</td>
</tr>
<tr>
<td>5_7</td>
<td>0.70</td>
</tr>
<tr>
<td>4_7</td>
<td>0.10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Set</th>
<th>Mean</th>
<th>Std</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set_1</td>
<td>1.2m/s</td>
<td>0.5m/s</td>
</tr>
<tr>
<td>Set_2</td>
<td>1.4m/s</td>
<td>1m/s</td>
</tr>
</tbody>
</table>
Summary

• Probabilistic model for path choice estimation considering denied boarding at key stations
  • Flexible to incorporate different variables and model structures
  • Robust to parameter initialization and small errors in inputs
• Neglecting denied boarding can lead to biased estimation of path fractions

• Future work
  • Real-world data testing
  • Learning behaviour under intervention (e.g. network expansion)
Zhenliang (Mike) Ma
mike.ma@monash.edu
https://www.monash.edu/engineering/mikema
Network Based Path Choice Estimation

Initial parameters and OD entry demand

NPM Engine (Assignment)

OD exit flows

No, update parameters

Difference between model and “true” OD exit flows is small enough

Path choice parameters

Simulation based optimization
Preliminary results (Bayesian method)

- Synthetic data generated using MTR network

<table>
<thead>
<tr>
<th>Variable</th>
<th>Beta (true)</th>
<th>Beta (Estimate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>In vehicle time</td>
<td>-0.0663</td>
<td>-0.0798</td>
</tr>
<tr>
<td># of transfer</td>
<td>-0.438</td>
<td>-0.293</td>
</tr>
<tr>
<td>Transfer time</td>
<td>-0.183</td>
<td>-0.210</td>
</tr>
<tr>
<td>Commonality Factor</td>
<td>-0.941</td>
<td>-0.996</td>
</tr>
<tr>
<td>Map distance</td>
<td>-0.0767</td>
<td>-0.0671</td>
</tr>
</tbody>
</table>

Discrete choice model provided by MTR using 2012 survey
Preliminary results (Bayesian method)

Best point

‘Brute-force’
Analytical Model

\[
\begin{align*}
\min_{\beta} & \quad w_1 \sum_{i_m, j_n} (q^{i_m, j_n} - \tilde{q}^{i_m, j_n})^2 + w_2 \| \beta - \tilde{\beta} \|^2 \\
\text{s.t.} & \quad q^{i_m, j_n} = \sum_{r=1}^{R(i,j)} q^{i_m, j_n}_r \quad \forall i_m, j_n \\
& \quad q^{i_m, j_n}_r = q^{i_m, j} \cdot p^{i_m, j}_r \cdot \mu^{i_m, j}_r \quad \forall i_m, j, r \\
& \quad p^{i_m, j}_r = \exp(\beta x_{r, m}) / \sum_{r'=1}^{R(i,j)} \exp(\beta x_{r', m}) \quad \forall i_m, j, r \\
& \quad \mu^{i_m, j}_r \text{ satisfies the NLC} \quad \forall i_m, j, r \\
& \quad \sum_r p^{i_m, j}_r = 1 \quad \forall i_m, j \\
& \quad 0 \leq p^{i_m, j}_r \leq 1 \quad \forall i_m, j \\
& \quad q^{i_m, j_n}_r \geq 0 \quad \forall i_m, j, r
\end{align*}
\]

\(\tilde{\beta}\) is the prior parameters for the choice model

\(\tilde{q}^{i_m, j_n}_r\) is the observed OD entry-exit flow.

Nonlinear equation constraints

Non-analytical constraints
Final Model Formulation

- Sub-problem 1(a)

\[
\begin{align*}
\min_{p_r^{im,jn}} & \quad w_1 \sum_{i_m,j_n} (q_r^{im,jn} - \bar{q}_r^{im,jn})^2 + w_2 \sum_{i_m,j_n,r} (p_r^{im,j} - \bar{p}_r^{im,j})^2 \\
\text{s.t.} & \quad q_r^{im,jn} = \sum_{r=1}^{R(i,j)} q_r^{im,jn} \quad \forall i_m, j_n \\
& \quad q_r^{im,jn} = q_r^{im,j} \cdot p_r^{im,j} \cdot \mu_r^{im,jn} \quad \forall i_m, j, r \\
& \quad p_r^{im,jn} \text{ satisfies the ALC of MNL} \quad \forall i_m, j_n, r \\
& \quad \sum_r p_r^{im,j} = 1 \quad \forall i_m, j \\
& \quad 0 \leq p_r^{im,j} \leq 1 \quad \forall i_m, j \\
& \quad q_r^{im,jn} \geq 0 \quad \forall i_m, j_n, r
\end{align*}
\]

- Sub-problem 1(b)

\[
\max_\beta \sum_{i_m,j} q_{im,j} \sum_{r=1}^{R(i,j)} p_r^{im,j} \cdot \log \frac{\exp(\beta x_{r,m})}{\sum_{r'=1}^{R(i,j)} \exp(\beta x_{r',m})}
\]

- Sub-problem 2

\[
\mu_r^{im,jn} = \text{Network Loading } (\beta, q_r^{im,j})
\]
Preliminary Test (Small Network)

OD entry-exit flow difference (Obj function)