ESTIMATION OF BUS PASSENGER OD PATTERNS BASED ON AVL DATA

Wenzhe Sun (Kyoto University, Japan)
Jan-Dirk Schmöcker (Kyoto University, Japan)
Koji Fukuda (Hitachi, Japan)
One-leg bus passenger OD matrix

**AFC data**

- Tap-in and tap-out
  - Origin (boarding stop) × destination (alighting stop) × time series (bus run)
- Tap-in only
  - Origin (boarding stop) × time series (bus run)
- Tap-out only
  - Destination (alighting stop) × time series (bus run)

**APC data**

- Boarding counter
  - Origin (boarding stop) × time series (bus run)
- Alighting counter
  - Destination (alighting stop) × time series (bus run)
The gap between AVL data and complete AFC data

AVL data

- bus location (bus stop) × time series (bus run)

- Replicating bus trajectory using AFC data

- Deriving stop-level OD, and in-vehicle passenger dynamics using AVL data

Complete AFC data

- origin (boarding stop) × destination (alighting stop) × time series (bus run)

For each bus run $k$

- Few existing literature

- 2D data (3D if latitude and longitude are of concern)

- Informative for deriving bus trajectories and evaluating service performance regarding punctuality and regularity

- Not designed for collecting passenger information
Motivation and objective

**Motivation**

- To allow operators without AFC data to obtain OD flow estimates
  - Reasonable route and frequency planning
  - Decision making on real-time bus control, especially boarding limitation

**Objective**

Use AVL data to
- Estimate stop-level passenger OD
- Estimate the number of boarding, alighting and on-board passenger for each stop
- Forecast passenger dynamics along the route in real time
**OD estimation methods**

- **AFC data**
  - Incomplete AFC data
  - Origin/Destination inference
    - Assumptions
    - Other data sources
      - Expansion methods (Cui 2006; Frumin 2010; Gordon et al. 2013; Lianfu et al. 2007; Park et al. 2008; Li et al. 2011; Zhao 2004; Zhao et al. 2007; ...)

- **AVL data**
  - Assumptions
  - Seed OD matrix
    - Assumptions
    - Other data sources
      - Estimated true OD matrix

**Origin/Destination inference**
- Barry et al. 2002; Farzin 2008; Ma and Wang 2014; He et al. 2015; Munizaga and Palma 2012;
- Trépanier et al. 2007; Wang et al. 2011; Zhao 2004; Zhao et al. 2007; ...

**Other data sources**

**Assumptions**

The results might be biased if overlooking that:
- Proportion of passengers using smart card may vary in different time periods
- Representative bias may exist if card-users are from specific groups of the PT user universe
- OD matrix has time-dependency
Methodology: framework

Passenger activity time $W = \max(T^A, T^B)$

or $W = T^A + T^B$

$(D - c) \sim N(W, \rho_W)$

c = deceleration + door open + door close + acceleration

$T^A_{l,k} = t^a A_{l,k}$

$T^B_{l,k} = t^b B_{l,k}$

$B_{l,k} = \sum_{i=1}^{N} Q_{l,i,k}$

$A_{l,k} = \sum_{j=1}^{l-1} Q_{j,i,k}$

$Q_{l,i,k} \sim Poisson(a_{l,j} \Delta_{l,k})$

Dwell time $D_{l,k}$

Alighting time $T^{A}_{l,k}$

Average time per person $t^a$

Boarding time $T^{B}_{l,k}$

Average time per person $t^b$

Arrival rate $a_{l,j}$

Headway $\Delta_{l,k}$

For alighting & boarding time per person $t^a$ & $t^b$

- They might increase if on-board number exceeds a threshold
- They also might vary by payment method
Methodology: estimate OD flows to fit AVL data

Dwell time $D_{i,k}$

$D_{i,k} \sim N(W_{i,k} + c, \rho_w)$

Dwell time consists of passenger activity time and vehicle activity time

Passenger activity time

$W_{i,k} = t^b B_{i,k} + t^a A_{i,k}$

or max$(t^b B_{i,k}, t^a A_{i,k})$

Vehicle activity time

$c = \text{deceleration} + \text{door open} + \text{door close} + \text{acceleration}$

• OD pattern is supposed to recur for bus runs over a time period (morning peak, evening peak, off-peak).

• Assume an arrival rate $a_{i,j}$ for each OD pair over time period $T$

$Q_{i,j,k} \sim \text{Poisson}(a_{i,j} \Delta_{i,k})$

$B_{i,k} = \sum_{j=i+1}^{N} Q_{i,j,k}$

$A_{i,k} = \sum_{j=1}^{i-1} Q_{j,i,k}$
Methodology: sampling dimensionality reduction

Arrival rate matrix

<table>
<thead>
<tr>
<th>Stop</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>...</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>0</td>
<td></td>
<td>0</td>
<td>0</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>0</td>
<td></td>
<td></td>
<td>0</td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

Gravity model (GM)

\[ a_{i,j} = \begin{cases} 
\frac{g_i}{(j-i)} + \frac{\alpha}{j-i}, & i < j \\
0, & \text{otherwise}
\end{cases} \]

Matrix factorization (MF)

Find latent matrix \( U \in \mathbb{R}^{N \times M} \) and \( V \in \mathbb{R}^{M \times N} \).

- Approximation method (instead of LU, QR): PMF

\( 2NM \)

Approximation: \( N(N-1)/2 \)
Model testing and verification

Step 1. Collect AVL and AFC data of a same city
Data period: 11 workdays during June 1-15, 2016

Step 2. Extract dwell time and headway from bus AVL data as the model input

Step 3. MCMC using Stan as the sampler (based on Hamiltonian Monte Carlo)

Step 4. Verify estimation results with AFC data

However, AVL data of this city does not record bus departure time. Departure time is thus created by using AFC data. The randomness due to vehicle activity time is reduced.
### MCMC settings and running times

- **Model input**
  - AVL data: \( D_{i,k}, \Delta_{i,k} \)
  - Survey data: \( t^a, t^b, c \)

- **Model output**
  - Parameters: \( \rho_{W}, g_i, \alpha \)
  - Transformed parameters: \( a_{i,j}, Q_{i,j,k}, B_{i,k}, A_{i,k}, O_{i,k}, W_{i,k} \)

<table>
<thead>
<tr>
<th>Time</th>
<th>Stop</th>
<th>Bus run</th>
<th>Chain</th>
<th>Warm-up</th>
<th>Sampling</th>
<th>#Para</th>
<th>T_warm (sec)</th>
<th>T_samp (sec)</th>
<th>T_total (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>7:00-8:00</td>
<td>24</td>
<td>50 (55)</td>
<td>3</td>
<td>1000</td>
<td>3000</td>
<td>24+1+1</td>
<td>132.521</td>
<td>332.031</td>
<td>464.552</td>
</tr>
<tr>
<td>8:00-14:00</td>
<td>24</td>
<td>50 (249)</td>
<td>3</td>
<td>1000</td>
<td>3000</td>
<td>24+1+1</td>
<td>127.549</td>
<td>258.41</td>
<td>385.959</td>
</tr>
<tr>
<td>14:00-18:00</td>
<td>24</td>
<td>50 (114)</td>
<td>3</td>
<td>1000</td>
<td>3000</td>
<td>24+1+1</td>
<td>148.987</td>
<td>380.641</td>
<td>529.628</td>
</tr>
<tr>
<td>18:00-19:00</td>
<td>24</td>
<td>22 (22)</td>
<td>3</td>
<td>1000</td>
<td>3000</td>
<td>24+1+1</td>
<td>41.387</td>
<td>95.175</td>
<td>136.562</td>
</tr>
</tbody>
</table>

\( t^a = 2 \text{sec} \)
\( t^b = 1.5 \text{sec} \)
\( c = 0 \)

\[
W_{i,k} = \max(t^b B_{i,k}, t^a A_{i,k})
\]

\[
\ln(e^{\beta t^b B_{i,k}} + e^{\beta t^a A_{i,k}}) / \beta
\]

2.5 GHz Intel Core i7
16 GB 2133 MHz LPDDR3
CmdStan 2.17.1
Parameter estimation ($\alpha$)

\[ \text{cost} = (j - i) + \frac{\alpha}{j - i} \]

$\bar{\alpha} = 284744$
Parameter estimation \((g_i)\)
Estimation results in the morning peak 7am-8am (11-day aggregated)
Boarding patterns during different time periods

- 7:00-8:00
- 8:00-14:00
- 14:00-18:00
- 18:00-19:00
Alighting patterns during different time periods
Passenger loads during different time periods
OD estimation results 7am-8am

RMSE = 7.2464
OD estimation results (hourly average)

- RMSE1 = 7.2464
- RMSE2 = 14.4273
- RMSE3 = 14.1206
- RMSE4 = 6.9269
Conclusions

Methodological
• A novel methodology using AVL data as the main data source to estimate the bus passenger OD patterns is developed.
• MCMC sampling is used, thus avoid assumption of specific distribution.
• A modified gravity model performs effectively in reducing sampling dimensionality.

Case study findings and practical implications
• Model fit appears fairly acceptable, estimation during peak hour better than during off-peak since constant passenger arrival rate is assumed (which is not the case with low frequency services)
• Offline and real-time application possible.
• Real-time application: Obtain expected downstream dwell times, capacity bottlenecks, basis for service control.
Further work: testing different dwell time functions

1. Boarding or alighting only
2. Simultaneous case
3. Sequential case because of in-vehicle overcrowding
4. Sequential case because of bus bunching, lane congestion, …

- If it is overcrowded in the bus, boarding passenger have to wait for some or all of alighting passengers getting off the bus. Overlapping or seamless sequential case might happen.
- When buses are bunched at the stop, and up to one bus is allowed to load passenger, large gap is likely to be observed between the end of alighting and the start of boarding.

Further work: application in Kyoto data

\[ W_{i,k} = \max(t^b B_{i,k}, t^a A_{i,k}) \text{ may not hold} \]