Combined estimation of activity generation models incorporating unobserved small trips using probe person data

The University of Tokyo
Sohta Itoh

Sep. 22nd, 2013
Contents

- Research background
- Comparison between PT and PP data
- Combined estimation model
- Correcting sampling bias
- Conclusion
Research background

1960s  **Person Trip survey** (Paper-based)
(1955 CATS, 1967 Hiroshima)

1980s  **Activity based model** – disaggregate data

2000s  **Probe Person survey** (GPS-based)
(Zitto and D’este, 1995; Murakami and Wagner, 1999; Asakura and Hato, 2004; Hato et al., 2006; Stopher et al., 2011)

Non-response activities

- Short trips and activities are often underreported
  (Wolf et al., 2001; Bricka and Bhat, 2006; Itsubo and Hato, 2006)

Changes of activity patterns

- Aging society
- Inner-city problems

Non-response bias

Short activities becomes important
Methods of PP survey

### GPS satellite
Tracking trajectory by GPS or base station

### Database

### Workplace

### Base station

### PP Survey
ex) observe 3 activities

#### PP data

- **GPS**
  - Timestamp
  - Latitude
  - Longitude

- **Web**
  - Trip purpose

- **diary**
  - Transportation mode

+ personal information

Legend:
- : location data (trajectory data)
- : trip destination (activity locations)

500m
PP survey data

Walk
Car
Bike
Motorcycle
Bus
Train

2019/7/5, 17:01:52

(PM)
### Comparison between survey data

<table>
<thead>
<tr>
<th>PT survey data</th>
<th>Massive location data</th>
<th>PP survey data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paper-based (Rely on respondents’ memories)</td>
<td>GPS (Automatical but fragmentary)</td>
<td>GPS (Automatical) + Web diary</td>
</tr>
<tr>
<td>Zone-based</td>
<td>Dot data (High-resolution)</td>
<td>Dot-based (High-resolution)</td>
</tr>
<tr>
<td>Large sample</td>
<td>Large sample</td>
<td>Small sample</td>
</tr>
</tbody>
</table>

- Activities within zones are unknown
- Short trips and activities can be observed

**Combined Estimation using both PT and PP data**
- Both data are obtained in Yokohama, Japan
- Respondents are resided in Yokohama

**PT survey**

| Surveillance period | 2008/10 - 2008/11  
| (each respondent answers his/her travel behavior of 1 day in surveillance period) |
| Method                  | Paper questionnaire |
| The number of all trips  | 1,906,032 trips  |
| The number of trips in Yokohama | 253,737 trips |

**PP survey**

| Surveillance period | 35 days (2010/07/05 - 2010/08/08) |
| Survey methods      | Probe Person survey with GPS cell phone + Web diary |
| The number of samples | 40 people |
| The number of Trips  | 3,617 trips |
| The number of location data | 789,074 points |
In almost all of categories, the number of activities of PT data is smaller than that of PP data

<table>
<thead>
<tr>
<th>Age</th>
<th>The number of activities</th>
<th>The sum of activity duration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>t-statistics</td>
</tr>
<tr>
<td></td>
<td>PT</td>
<td>PP</td>
</tr>
<tr>
<td>age 20s</td>
<td>1.26</td>
<td>1.39</td>
</tr>
<tr>
<td>age 30s</td>
<td>1.40</td>
<td>1.60</td>
</tr>
<tr>
<td>age 40s</td>
<td>1.53</td>
<td>1.74</td>
</tr>
<tr>
<td>age 50s</td>
<td>1.55</td>
<td>1.80</td>
</tr>
<tr>
<td>age 60s+</td>
<td>1.56</td>
<td>1.58</td>
</tr>
<tr>
<td>male</td>
<td>1.49</td>
<td>1.78</td>
</tr>
<tr>
<td>female</td>
<td>1.43</td>
<td>1.43</td>
</tr>
<tr>
<td>total</td>
<td>1.46</td>
<td>1.60</td>
</tr>
</tbody>
</table>

* : reject the null hypothesis of no difference between the mean of PT data and that of PP data at 5% significant level
• It is assumed that PP data does not have unreported activities.
• If missing activities have some characteristics in common, sampling bias affects the estimation result.

Detecting the factors influencing the propensities to record activities

Estimation model framework

- Performed activities
  - PP data
  - PT data

Selection model

Activity generation model (estimating possibility of activities (using common variables of PP/PT))

Unreported activities

Weight (correcting non-response bias)
Introducing selection model

Apply Tobit selection model to activity generation and its observation

● Activity generation model

\[ y_{in1}^* = \beta_{1} x_{in1} + \varepsilon_{in1} \]

Latent variable about activity generation of individual \( i \) and zone \( n \)

\( x_{in1} \): explanatory variables of individual \( i \) and zone \( n \)
\( \varepsilon_{in1} \): error term of individual \( i \) and zone \( n \)

\[ \begin{align*}
    y_{in1} &= 1 \quad \text{if} \quad y_{in1}^* > 0 \quad \text{generate} \\
    y_{in1} &= 0 \quad \text{if} \quad y_{in1}^* \leq 0 \quad \text{not generate}
\end{align*} \]

● Selection model

\[ y_{in2} = \beta_{2} x_{in2} + \varepsilon_{in2} \]

if \( y_{in2} > 0 \) \( y_{in1} \) is observed
if \( y_{in2} \leq 0 \) \( y_{in1} \) is not observed

Latent variable about observation of individual \( i \) and zone \( n \)

\( x_{in2} \): explanatory variables of individual \( i \) and zone \( n \)
\( \varepsilon_{in2} \): error term of individual \( i \) and zone \( n \)
\( y_{in2} \): unobserved variable of individual \( i \) and zone \( n \)

\[ \begin{pmatrix}
    \varepsilon_{in1} \\
    \varepsilon_{in2}
\end{pmatrix} \sim N\left( \begin{pmatrix}
    0 \\
    0
\end{pmatrix}, \begin{pmatrix}
    \sigma_1^2 & \rho \sigma_1 \\
    \rho \sigma_1 & 1
\end{pmatrix} \right) \]
Introducing selection model

- Activity generation model
  \[ y_{in1} = \beta_1 x_{in1} + \varepsilon_{in1} \]

- Selection model
  \[ y_{in2} = \beta_2 x_{in2} + \varepsilon_{in2} \]

Expected value of latent variable \( y_{in1} \) after considering selection bias

\[
E(y_{in1} \mid y_{in2} > 0) = \beta_1 x_{in1} + E(\varepsilon_{in1} \mid \varepsilon_{in2} > -\beta_2 x_{in2})
\]

\[
= \beta_1 x_{in1} + \rho \sigma_1 \frac{\phi(\beta_2 x_{in2})}{\Phi(\beta_2 x_{in2})}
\]

Correlation term (apply only for PT data)

- \( \Phi \) : cumulative distribution function of the standard normal distribution
- \( \phi \) : probability density function of the standard normal distribution
### Estimation results

<table>
<thead>
<tr>
<th>Independent variables</th>
<th>The normal activity generation model</th>
<th>The sample selection model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Parameter</td>
<td>t score</td>
</tr>
<tr>
<td>For activity generation model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>-1.902</td>
<td>-76.64 *</td>
</tr>
<tr>
<td>Male</td>
<td>0.091</td>
<td>12.59 *</td>
</tr>
<tr>
<td>Age ( \geq 60 )</td>
<td>-0.116</td>
<td>-15.37 *</td>
</tr>
<tr>
<td>Single-member household</td>
<td>0.090</td>
<td>8.79 *</td>
</tr>
<tr>
<td>Car ownership</td>
<td>-0.003</td>
<td>-0.42</td>
</tr>
<tr>
<td>Distance from home (km)</td>
<td>-0.108</td>
<td>-98.83 *</td>
</tr>
<tr>
<td>Distance from workplace (km)</td>
<td>-0.025</td>
<td>-43.52 *</td>
</tr>
<tr>
<td>Store space (ha) (^1)</td>
<td>0.043</td>
<td>71.31 *</td>
</tr>
<tr>
<td>( \gamma )</td>
<td>0.125</td>
<td>5.09 *</td>
</tr>
<tr>
<td>( \rho )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>For selection model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age 20-39 years</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Age ( \geq 60 )</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Distance from home (km)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Distance from workplace (km)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Stay Duration (min.)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>( \mu )</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

- Observations (PT) 1,780,164
- Observations (PP) 23,000
- Initial log-likelihood \(-1,249,858\)
- Final log-likelihood \(-65,013\)
- Rho-squared \(\rho^2\) 0.948

Following attributes associate with activity under-reporting at the significant level:

- male
- stay duration
- age 20-39 years
- age 60+

\(^1\) : The sum of space about retail stores in the zone

- Not relevant; * Significant at 5% level.
Correcting sampling bias

To correct the bias, the inverse of observation probability is considered the weight as:

\[ w_{in} = \frac{1}{p(y_{in2} > 0 \mid x_{in2})} = \frac{1}{\Phi(\beta^*_2 x_{in2})} \]

\( \beta^* \): the parameter estimated in the model comes from the estimation results

Observation activity data (disaggregate)  Corrected results

activities with attributes \( x \)

\[ \text{multiply the correcting weight} \]

\[ \frac{1}{\Phi(\beta^*_2 x_{in2})} \]
The rate of frequency of weighted PT is similar to PP, which represents the bias of short activities is corrected.
Correcting sampling bias

The rate of discretionary activities is expanded by weighting.

Work 240 min.  Shopping 15 min.  Private 30 min.

Adding activities stochastically
Comparison between PT and PP
We have discussed the advantages of both new GPS-based PP surveys and conventional PT surveys

Combined estimation using PT and PP data
Introducing the selection model, we show several demographic attributes and activity characteristics associate if activities are missed or not and consider the selection bias

Correcting the sampling bias
By multiplying the inverse of probabilities of observation obtained from the selection model, the bias is appropriately assessed and corrected
Thank you for your attention!
References


