Benefits of Cellular Telecommunication and Smart Card Data for Travel Behaviour Analysis
From a cross-sectional to a dynamic approach

Patrick Bonnel,
Université de Lyon, ENTPE, LAET, Lyon
Research objectives

• Traditional travel surveys offer rich semantic data, but only one or few travel day every 5-10 years

• Road-side, cordon, origin-destination surveys might be more frequent but one trip only with limited semantic

• Long term aggregate dynamic possible using panel or pseudopanel analysis

• But short term dynamic analysis or intrapersonal variability analysis very limited

Continuous passive (big) data mean new perspectives for dynamic behavioral analysis: potentials and limits
Mobile phone probe data exploration

- Huge « passive » database with spatio-temporal information
- Possibility to identify individuals’ presence in time and space
- ORANGE probe data: richer than CDR (call detail record), less dependent to individual communication behavior (handover, Location area update, attach/detach events)
Mobile phone data filtering and expansion

- Maximum Inter-event Time (MIT) \(\leq 180\) minutes (a mobile switch on should have at least one LAU every 3 hours)
- Entropy (H) \(\leq 0.9\) (avoid machine-to-machine devices)
- Number of observations (NO) \(\geq 4\) (for trip identification)
- Household home location identification (for expansion to whole population, expansion factors 3.5-10)

Data base represents 50% of initial mobile phone devices
Mobile phone “ground truth” validation

Comparison with Rhône-Alpes travel survey

For 30 min:  \[ y_{ij} = 0.85 \times x_{ij} + 877 \]  \( R^2 = 0.95 \)

For 40 min:  \[ y_{ij} = 0.80 \times x_{ij} + 788 \]  \( R^2 = 0.95 \)

Slope close to 1 and constant to 0
Mobile phone/HTS temporal profile

Mobile phone temporal profile for whole Rhône-Alpes needs to be corrected (smaller peak especially in morning)

Initial profile

After debiasing process
Mobile phone – various temporal profile

- Rhône-Alpes spatial clustering (77 zones) based on departure time distribution
- Profile based on origin (without intrazonal trips)

![Graphs showing temporal demand in different areas: rural, urban, mix, and very dense urban areas.](image-url)
Smart card data for Lyon conurbation

- Lyon conurbation (1.3 million inhabitants) transit network transaction only at vehicle boarding (including transfer)
  - Smart card (80% of validation, same Id over a long period)
  - Magnetic paper ticket (20% of validation, without Id)
- AVL (Automatic vehicle location)
- Automated passenger counting system (bus, tramway, subway)
- Origin-destination surveys (on board, all routes at least every 5 years)
- Household travel survey (every 10 years, nearly 1% stratified sampling, face-to-face)
Smart card data processing and expansion

Data correction and imputation

- Missing data imputation + deduplication
- Transfer identification to transform trip-legs into trips (rules from literature)
- Destination inference rules only for smart card data (same Id): 80.8% success
- Fraud (or non-validation) represents 21% of total transit trips

Data expansion

- Transit trips with alighting location: ≈ 50% of total transit trips
- Expansion with non uniform scaling factors based on route/subway station passenger counting (≈ 50k scaling factors)
## Smart card data “ground truth” validation

<table>
<thead>
<tr>
<th></th>
<th>Smart card data</th>
<th>Public transport origin-destination survey</th>
<th>Household travel survey (HTS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trip legs (million)</td>
<td>1.56</td>
<td>1.51</td>
<td>1.11</td>
</tr>
<tr>
<td>Trips (million)</td>
<td>1.10</td>
<td>1.16</td>
<td>0.80</td>
</tr>
<tr>
<td>Bus trip legs (%)</td>
<td>41</td>
<td>39</td>
<td>43</td>
</tr>
<tr>
<td>Tramway trip legs (%)</td>
<td>23</td>
<td>22</td>
<td>21</td>
</tr>
<tr>
<td>Subway trip legs (%)</td>
<td>37</td>
<td>39</td>
<td>36</td>
</tr>
</tbody>
</table>

- Much less trip-legs and trips in household survey compared to smart card data and O-D survey which appears much more coherent.
Smart card data, spatial “ground truth” validation

- Spatial comparison at O-D level (18 zones)
- Smart card data are much more coherent with transit O-D survey, than with household travel survey
Smart card analysis over 6 months

• Clustering on vector day with/without trip

• We can identify 3 groups:
  • Consistent transit users (regular user, top dendrogram) 45% of users, 69% of trips
  • Low frequency users (middle dendrogram) 14% - 1%
  • Intermittent transit users (bottom of dendrogram) 41% - 30%

• Day-to-day regularity does not mean individual regularity
Smart card analysis over 6 months – among regular users

- Cluster 1 very high transit use even WE, no calendar effect
- Cluster 2 high use in week-days, lower WE, no calendar
- Cluster 3 WE + calendar (holidays) effects
- Cluster 4 regular use without clear effect of WE and calendar
- Cluster 5 calendar and WE effects with lower use (than C3)
- Cluster 6 sparse use but somewhat regular without clear structure
Big data base are not error free

• Mobile phone data need filtering process for example to suppress machine-to-machine devices, or devices with too few data

• Smart card data need correction for example deduplication

• Data imputation is often required for missing information
Big data base need expansion factors

• Even if big, data base are not exhaustive and do not represent whole population
• Individuals might have no/several devices

• Expansion factors are required with spatio-temporal scaling factors
• External sources improve scaling quality
Big data base require « ground truth » validation

• Passive big data sources evolve continuously
• Passive big data might be context-dependant

• Regular ground truth validation is recommended using external information not used in data processing
But big data present a high potential for dynamic analysis even at disaggregate level

After correction, debiasing process, expansion and validation continuous big volume data is available allowing:

• Aggregate and disaggregate regularity/variability analysis

• Detailed spatio-temporal analysis including O-D matrix analysis at fine grain level

• Intrapersonal variability when same Id is available over time
### Mobile phone data

- 2G et 3G signaling data collected during June 2017
- From « Orange » mobile phone operator
- For all Rhone-Alps Region
- More than 2 millions users and 300M data per day
- Each trace is anonymized and with timestamp. Mobile ID is changed every day

<table>
<thead>
<tr>
<th>Timestamp</th>
<th>IMSI (ID)</th>
<th>LAC</th>
<th>ID cellule</th>
<th>évènement</th>
</tr>
</thead>
<tbody>
<tr>
<td>2017-06-01</td>
<td>201803567834</td>
<td>104</td>
<td>20865</td>
<td>CALL</td>
</tr>
</tbody>
</table>

Localisation = Location Area Code (LAC) + ID cellule
EDR Rhône-Alpes data

- Regional travel survey (EDR)
- Conducted between 2012 and 2015 on all Rhone-Alps region
- 38,000 individuals above 11 years old, 143,000 trips
- Territory zoning: 77 sampling sectors, aggregated in 14 macro zones
Cell phone activity: Indicators-based filtering

- Maximum Inter-event Time (MIT) [7am-10pm] ≤ 180min (presence)
- Entropy (H) ≤ 0.9 ; \( H(X) = -\sum_{i=1}^{n} p(x_i) \log(p(x_i)) \) (uniformity)
- Number of observations (NO) ≥ 4 (frequency should be ≥ 8 with LAU)
- Filtering of outlier, uniform and machine-generated behaviors
Stationary threshold choice

<table>
<thead>
<tr>
<th>Stationary activity time threshold</th>
<th>60 minutes</th>
<th>50 minutes</th>
<th>40 minutes</th>
<th>30 minutes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trips number EDR (in thousand)</td>
<td>2,211</td>
<td>2,260</td>
<td>2,344</td>
<td>2,448</td>
</tr>
<tr>
<td>Trips number Orange (in thousand)</td>
<td>1,607</td>
<td>1,743</td>
<td>1,905</td>
<td>2,108</td>
</tr>
</tbody>
</table>

Probability distribution of total trips per user with a threshold 30min and 60min

We keep 30-40 minutes threshold which seems reasonable regarding sector size and gives the best results in comparison to EDR
O-D expansion: expansion factor (2)

• **Expansion Factor** definition on sector level (77 sectors):

\[ F_{\text{exp}}(\text{sector}_i) = \frac{\text{Population of sector}_i \text{ (over 11 years)}}{\text{Nb of home locations detected in sector}_i} \]

Expansion factor probability distribution

Spatial distribution of sector expansion factors in the Rhône-Alpes region after user filtering
EDR – Orange comparison (2)

For 30 min:  \( y_{ij} = 0.70 \times x_{ij} + 2,193 \)  \( R^2 = 0.96 \)
For 40 min:  \( y_{ij} = 0.66 \times x_{ij} + 1,964 \)  \( R^2 = 0.96 \)

O-D pairs between the two Lyon sectors are badly estimated and strongly impact slope.
Signaling mobile phone data

• Entropy formula

\[ H(X) = - \sum_{i=1}^{n} p(x_i) \log(p(x_i)) \]

The entropy measures the randomness of a system or on the opposite its regularity. High measure of entropy corresponds to very regular signals like those generated by machine-to-machine communications or IOT (Internet of Things).