Passive Data for Transportation Modelling:
imputing mode and purpose information from cellphone traces

UT-ITE Seminar
February 8, 2019

Ahmadreza Faghih Imani, Ph.D.
Postdoctoral Fellow
Dept. of Civil & Mineral Engineering
University of Toronto
Passive Data for Transportation Modelling: imputing mode and purpose information from cellphone traces

- **Acknowledgements:**
  - James Vaughan, Bilal Yusuf, Prof. Eric Miller
  - This study was part of iCity-South project funded by CAF.
  - CAF: Nicolas Estupinan and Andres Alcala
  - Diego Hernandez, Universidad Catolica del Uruguay
  - Antonio Mauttone, Universidad de la Republica de Uruguay
  - Antel: Juan Pablo Pignataro, Julio Ciambelli and Sebastian Lamas
  - Intendencia de Montevideo: Richard Delgado, Nestor Sosa and Justo Onandi
Introduction

- What is passive data?
- Respondents don’t need to actively participate and report the information.
  - Typically in nature present “revealed preference”

- Examples:
- Public transit smartcard transactions or cellular network data
Cellular Data

- It is possible to identify people movements using cellular network data from the base transceiver stations (BTSs).
Cellular Data

- Several cell to infrastructure connection events:
  - handovers (HO)
  - call detail records (CDR)
  - location updates (LU)
    - HO and CDR provide data of communication events such as calls or SMSs and LU notify the cellular network when a cellphone moves from one BTS to another.

→ Results in time-space traces of movements of cellphones
  → can be used to impute travel movements by users.
## Cellular Data

<table>
<thead>
<tr>
<th>Advantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use of existing data</td>
</tr>
<tr>
<td>Large sample size</td>
</tr>
<tr>
<td>Large coverage</td>
</tr>
<tr>
<td>Long time period</td>
</tr>
<tr>
<td>No interaction with users</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>No individual information</td>
</tr>
<tr>
<td>Depends on cell phone market penetration</td>
</tr>
<tr>
<td>Multiple counting of a person with multiple devices</td>
</tr>
<tr>
<td>Lack of spatial precision</td>
</tr>
</tbody>
</table>
Earlier Studies on Cellular Data

- Mobility Patterns

A Tale of One City: Using Cellular Network Data for Urban Planning

Cellular data from call detail records can help urban planners better understand city dynamics. The authors use CDR data to analyze people flow in and out of a suburban city near New York City.
Earlier Studies on Cellular Data

- Meaningful locations:
  - Home
  - Work/School
Earlier Studies on Cellular Data

- OD estimations

Inferring origin-destination trip matrices from aggregate volumes on groups of links: a case study using volumes inferred from mobile phone data

Noelia Caceres, Luis M. Romero and Francisco G. Benitez

Deriving origin-destination data from a mobile phone network

Authors: N. Caceres 1; J.P. Wideberg 1; F.G. Benitez 1

View affiliations

Source: Volume 1, Issue 1, March 2007, p. 15 – 26

DOI: 10.1049/iet-ntn.2006.0020
Print ISSN: 1751-955X, Online ISSN: 1751-9568

Estimating Origin-Destination Flows Using Mobile Phone Location Data

4 Author(s) Francesco Catarrese; Giuseppe Di Lorenzo; Liang Liu; Carlo Ratti

View All Authors

Inferring origin-destination flows using mobile phone data: A case study of Senegal

5 Author(s) Menkebe Getachew Demisse; Francisco Antunes; Carlos Sento; Santiphakaphimukoon; Trirat Suktivibul

View All Authors

Investigating the Mobile Phone Data to Estimate the Origin Destination Flow and Analysis; Case Study: Paris Region

Anahid Nabavi Larjani 1, 2, 3, Ana-Maria Olteanu-Raimond 4, 5, 1, Julien Perret 6, Mathieu Bredif 7, Cecary Zaimidzi 1

International Workshop on Citizen as Sensor Networks

Transportation Planning Based on GSM Traces: A Case Study on Ivory Coast

Authors

Erik Mellegard; Simon Moritz; Mohamed Zahoor

View All Authors

Origin/Destination-estimation Using Cellular Network Data

3 Author(s)
Earlier Studies on Cellular Data

- OD estimations
  - Estimating traffic flows
  - Optimizing public transport network
  - Activity-based travel demand models
Objectives

1. Identifying trips and activities
2. Detecting home and work location
3. Imputing mode of travel  This is new; primary objective!
• Montevideo, Uruguay
  Population:
  City: 1.38 million
  Metro: near 2 million
  Country: 3.4 million
Data

- 2016 Montevideo Household Mobility Survey (MHMS) records.
- Road and transit network data.
- Census data.
- A very large sample of Antel cellphone traces, consisting of 40% of all such traces for four weeks.
- All public transit fare transaction records for the same time period, provided by the Intendencia de Montevideo.
MHMS Data

- August-October 2016 in the Metropolitan Area of Montevideo.
- 0.34% sample of the households in the region:
  - 2,230 households.
  - 5,946 individuals.
  - 12,546 reported trips.
  - Average daily rate of 2.11 trips/person.
- Spatially aggregated to census segments.

<table>
<thead>
<tr>
<th>Travel Mode</th>
<th>Mode Share in %</th>
<th>Number of Trips</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walk</td>
<td>34.0</td>
<td>4265</td>
</tr>
<tr>
<td>Bike</td>
<td>3.5</td>
<td>439</td>
</tr>
<tr>
<td>Auto Passenger</td>
<td>10.0</td>
<td>1251</td>
</tr>
<tr>
<td>Auto Driver</td>
<td>19.2</td>
<td>2410</td>
</tr>
<tr>
<td>Motorcycle</td>
<td>6.1</td>
<td>769</td>
</tr>
<tr>
<td>Bus</td>
<td>25.2</td>
<td>3166</td>
</tr>
<tr>
<td>Other</td>
<td>2.0</td>
<td>246</td>
</tr>
</tbody>
</table>
Network Data

- Computerized representations of the Montevideo road & transit networks were constructed in Emme for modelling purposes. Sources:
  - Roads: OpenStreetMap database.
  - Bus Network: Open data portal of Montevideo government.
Census Data

- Population, Home and Jobs at Census Segment level.
Antel Data (1)

- Antel is the primary cellular telecommunications company in Uruguay.
- Traces of a random sample of 40% of mobile phones within Montevideo and the surrounding metropolitan area.
- May 2\textsuperscript{nd} to May 29\textsuperscript{th}, 2018.

- Raw data processed by Antel:
  - To eliminate as much noise in the data as possible & to preserve user anonymity.
  - Trace data are temporally reported in minutes.
  - Spatially aggregated to 135 zones in the Montevideo region

- 117,862,000 cellphone traces for about 948,600 unique cellphones.
Challenges with Data

- Spatial precision

The Montevideo has 186 TAZs
Antel Data (2)

- **Zone size:**
  - Average: 37.1 km$^2$
  - Min: 0.38 km$^2$, Max: 1,020 km$^2$.

- **Activity:** any time a cellphone user stays more than 30 minutes in a zone.

- **Trips:** when a user moves between two activities with the first activity location identified as the trip origin and the second activity location as the trip destination.
Method - Example

<table>
<thead>
<tr>
<th>Time</th>
<th>7:45</th>
<th>8:00</th>
<th>8:15</th>
<th>8:30</th>
<th>...</th>
<th>12:00</th>
<th>12:15</th>
<th>...</th>
<th>13:15</th>
<th>13:30</th>
<th>...</th>
<th>18:15</th>
<th>18:30</th>
<th>18:45</th>
<th>19:00</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zone</td>
<td>221</td>
<td>230</td>
<td>229</td>
<td>218</td>
<td>...</td>
<td>218</td>
<td>212</td>
<td>...</td>
<td>212</td>
<td>218</td>
<td>...</td>
<td>218</td>
<td>236</td>
<td>230</td>
<td>221</td>
<td>...</td>
</tr>
<tr>
<td>Zone</td>
<td>Home</td>
<td>Travel</td>
<td>Activity 1</td>
<td>Activity 2</td>
<td>Activity 3</td>
<td>Travel</td>
<td>Home</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Diagram for method example:

- Home Zone 221
- Trip
- Work Zone 218
- Trip
- Other Zone 212
- Trip
- Work Zone 218
- Trip
- Home Zone 221
Antel Trips (1)

- The data were cleaned to remove some inconsistent trips such as jumps in locations and trips with zero duration with the same origin and destination.
Antel Trips (2)

- The general pattern of trip-making observed in the Antel data is similar to the MHMS pattern, except the MHMS AM trips are more peaked. This may be due to:
  - Antel traces missing some morning trips (cellphones not turned on when leaving home in the morning?).
  - MHMS missing mid-day trips (which would skew the MHMS distribution).
- Cell traces from weekdays only used to be comparable to household travel survey.
Home Location (1)

- From cell traces, it is possible to infer and make high accuracy estimates of people’s primary locations such as their place of residential location.
- An Antel zone is labeled as the cellphone user’s home zone using different approaches and then compared with the census.
  - Home is the zone with most stayed duration.
  - Home is the zone with most stayed duration during specific periods.
  - Home is the zone with most first trips originated and last trips destined.
  - Home is the zone with most trips destined there.
Home Location (2)

- Most stayed duration in weekends provides the highest fit to the distribution of the homes from census.
  - For some (13.6%), this method could not clearly identify a location for home (for example because of no traces recorded during weekends)
  - For these group, the most stayed location is selected as home zone.

<table>
<thead>
<tr>
<th>Type of Algorithm</th>
<th>% of users home identified</th>
<th>Correlation with census home</th>
</tr>
</thead>
<tbody>
<tr>
<td>most stayed duration</td>
<td>100.0%</td>
<td>0.525</td>
</tr>
<tr>
<td>most stayed duration during night time (19 to 9)</td>
<td>90.0%</td>
<td>0.545</td>
</tr>
<tr>
<td>most stayed duration during night time (21 to 7)</td>
<td>74.9%</td>
<td>0.555</td>
</tr>
<tr>
<td>most stayed duration during weekends</td>
<td>86.4%</td>
<td>0.572</td>
</tr>
<tr>
<td>most first trips originated and lasted trips destined</td>
<td>43.3%</td>
<td>0.390</td>
</tr>
<tr>
<td>most trip destined</td>
<td>76.7%</td>
<td>0.436</td>
</tr>
<tr>
<td>most trip destined during night time (19 to 7)</td>
<td>56.2%</td>
<td>0.488</td>
</tr>
<tr>
<td>most trip destined during weekend</td>
<td>52.8%</td>
<td>0.497</td>
</tr>
</tbody>
</table>
Work Location

- The second most visited location for cell phone users is very likely to be their usual place of work/school.
- Similar approach to home location detection. At the end, 66% of cellphones were given a work/school location.
  - Most stayed location during weekdays daytime (8 to 18) with minimum of 1 hour stay.
    - A threshold of minimum of 1 hour stay in a day in average is also used to remove the locations with frequent visits but with short duration.
    - For those users that work location were not identifiable, it is assumed that those users are not workers/students.

<table>
<thead>
<tr>
<th>Type of Algorithm</th>
<th>% of users home identified</th>
<th>Correlation with census jobs</th>
</tr>
</thead>
<tbody>
<tr>
<td>most stayed duration during day time (8 to 18), minimum of 1 hour stay</td>
<td>68%</td>
<td>0.471</td>
</tr>
<tr>
<td>most stayed duration during day time during weekdays (8 to 18), minimum of 1 hour stay</td>
<td>66%</td>
<td>0.474</td>
</tr>
</tbody>
</table>
Home based trips

- Antel home destinations correspond well to the MHMS distribution.
- Again, AM trip origins from home seem to be under-represented in the Antel data. The rest of the day looks good.
Work Based Trips

- MHMS data may be missing morning/mid-day work trip origins?
- Again, Antel AM work destinations appear to be low.
Home-Work Trips

- Note that these are not tours; i.e., these are one-way trips
- Comparison to MHMS trips is generally good, except, again for AM home-work trips.
Home Density

- Household residential are computed based on the home zones identified for the 40% sample of Antel cellphone users.
- It is individual based and not household based; i.e. homes for individuals in the same household are double-counted.
- Still the spatial distribution resembles the census distribution.
Job Density

- Similarly, employment zone job densities are computed from the 40% Antel user sample.
- It is based on the work/school locations identified, which are basically the second primary location where a user spends time.
- Still the spatial distribution resembles the census distribution.
Transit Data

- The Intendencia de Montevideo provided data for Montevideo’s integrated public transit system (STM) for the same period in May 2018.
- Every transit boarding is available: smartcard and cash payments.
- The transit system is tap-on only, hence destinations must be estimated.
- Intendencia de Montevideo has developed their own transaction data processing methods to estimate alighting stops.
  - 62% alighting stops have been successfully estimated.
Transit Data

- 29,868,716 recorded transactions.
- 82.5% by smartcards.
- 734,569 unique smartcards.
- On average, a smartcard was used for 33.6 transactions (SD=29.3) in May.
Transit Data – Identifying locations

- For the smartcard transactions, it is possible to investigate the frequency of bus stops by time of the day.
- Assuming the first transaction of the day is usually made at the stops near home location, we can label that stop (and the zone) as the home location of the rider.
  - Since some of the smartcard users do not have a repetitive travel pattern (e.g. 7.6% of the users only have 1 or 2 transactions in the month), we keep a threshold of minimum three trips originated from the stop to be considered for home location.
- Similarly, the last transaction of the day is labeled as work location (with the same minimum 3 transactions threshold).
- Overall, 76% of smartcards have assigned a home zone and 65% have a work zone assigned.
Transit Data - Trips

- Hourly distribution of trips with origin/destination at home/work during weekdays.
Transit Data - Home & Work Locations

Home Density
- 0 - 500
- 501 - 1000
- 1001 - 2000
- 2001 - 3000
- 3001 - 5000
- 5001 - 8000
- 8001<

Job Density
- 0 - 1000
- 1001 - 2000
- 2001 - 3000
- 3001 - 5000
- 5001 - 8000
- 8001 - 12000
- 12001<
Transit Data (2)

- Comparing the daily frequency of trips based on the Antel cellphone data and smartcard transaction data.
Imputing **travel mode** for Antel cellphone traces

- Neural Network model.

![Diagram showing travel modes and features](image-url)
Analysis Method

1. Convert MHMS O-D trips into pseudo-cellphone traces.
2. Train a neural net on the MHMS pseudo-traces to identify travel mode.
3. Apply the trained model to the Antel cellphone traces to classify them by mode.
Steps in Constructing the Mode Choice Model

- Model Road & Transit Networks
- Create Artificial Traces
- Train Classifier
- Execute Classifier
- Validate Results
Create the Network Models

- Based on Open Street Map
- Transit lines loaded from GTFS & Open Portal and then manually inspected removing artifacts from the importing process
- Demand is loaded from MHMS for each time period
- Route assignment parameters are taken from GTModel V4.0.2
Creating Artificial Traces

- All of the trips for a person in the MHMS are loaded
- Each trip is then processed assign the distance travelled when crossing an Antel zone boundary assigning that value to a five minute bin
Neural Network Features

- Each artificial trace has two values for each five minute bin of time during the day.
  - The first value contains the distance travelled when crossing boundaries.
  - The second value contains whether or not this time bin occurs during trip to assign.
- Each record contains the **full day**’s worth of trip distance features.

<table>
<thead>
<tr>
<th></th>
<th>Time of Day</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0:00</td>
</tr>
<tr>
<td>Distance</td>
<td>0</td>
</tr>
<tr>
<td>Active</td>
<td>0</td>
</tr>
</tbody>
</table>
Train Classifier

- The neural net was trained using Microsoft’s Cognitive Neural Toolkit (CNTK)
- Artificial traces were split into two datasets, one for estimation and one for testing.
- On the estimation set an accuracy of 0.98 was reached.
- On the testing set an accuracy of 87% was reached.

| Predicted vs. Observed Mode (% of total trips) |
|-----------------|-----------------|-----------------|-----------------|-----------------|
| Predicted       | Observed Mode   |
| Mode            | Auto            | Transit         | Active          | Total           |
| Auto            | 33.65           | 1.08            | 2.54            | 37.27           |
| Transit         | 1.76            | 21.36           | 1.92            | 25.04           |
| Active          | 2.91            | 2.88            | 31.91           | 37.7            |
| Total           | 38.32           | 25.32           | 36.37           | 100             |
Cellphone mode results: share, all day, by trip origin
Cellphone auto mode shares by time of day & trip origin zone
Cellphone transit mode shares by time of day & trip origin zone
Cellphone active mode shares by time of day & trip origin zone
Validation Results (1)

- Aggregate numbers are then compared against the MHMS

<table>
<thead>
<tr>
<th></th>
<th>Auto</th>
<th>Transit</th>
<th>Active</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auto Cell Traces</td>
<td>4083414</td>
<td>5159537</td>
<td>2515295</td>
</tr>
<tr>
<td>Transit</td>
<td>4250</td>
<td>2992</td>
<td>4461</td>
</tr>
<tr>
<td>Active</td>
<td>36.32%</td>
<td>25.57%</td>
<td>38.12%</td>
</tr>
</tbody>
</table>

- It seems that the model has confusion between transit and active transportation however the split between auto and non-auto looks to be good.
Validation Results (2)

- Observed mode share by origin zone vs predicted
  - Exclude intrazonal trips in MHMS.
    - less than 0.16 for auto and transit and about 0.1 for active mode.

<table>
<thead>
<tr>
<th>MODE SHARE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUTO</td>
<td>0.1593</td>
</tr>
<tr>
<td>TRANSIT</td>
<td>0.1597</td>
</tr>
<tr>
<td>ACTIVE</td>
<td>0.1046</td>
</tr>
</tbody>
</table>
Validation Results (3)

- Observed mode share by time of the day vs predicted.

<table>
<thead>
<tr>
<th></th>
<th>AUTO</th>
<th>TRANSIT</th>
<th>ACTIVE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AM</strong></td>
<td>43%</td>
<td>42%</td>
<td>16%</td>
</tr>
<tr>
<td><strong>MD</strong></td>
<td>33%</td>
<td>44%</td>
<td>23%</td>
</tr>
<tr>
<td><strong>PM</strong></td>
<td>41%</td>
<td>40%</td>
<td>20%</td>
</tr>
<tr>
<td><strong>EV</strong></td>
<td>52%</td>
<td>34%</td>
<td>14%</td>
</tr>
<tr>
<td><strong>ON</strong></td>
<td>45%</td>
<td>44%</td>
<td>11%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>AUTO</th>
<th>TRANSIT</th>
<th>ACTIVE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AM</strong></td>
<td>31%</td>
<td>54%</td>
<td>14%</td>
</tr>
<tr>
<td><strong>MD</strong></td>
<td>29%</td>
<td>51%</td>
<td>20%</td>
</tr>
<tr>
<td><strong>PM</strong></td>
<td>35%</td>
<td>42%</td>
<td>24%</td>
</tr>
<tr>
<td><strong>EV</strong></td>
<td>42%</td>
<td>28%</td>
<td>30%</td>
</tr>
<tr>
<td><strong>ON</strong></td>
<td>63%</td>
<td>22%</td>
<td>15%</td>
</tr>
</tbody>
</table>
**Validation Results (4)**

- Observed mode share vs predicted by trip length.

![Graphs showing mode share predictions vs actual data](image)

**Mode Share by Distance Category**

- Short transit trips are over-predicted & short active trips are under-predicted.
Validation Results (5)

- Predicted cellphone transit trip vs smartcard data
  - Average peak hour trips.
  - For expansion of the Antel sample to the population, it is assumed:
    - 40% sample.
    - 60% market penetration.
- Temporal pattern is very good.
- Consistent with other results, it appears that we are over-estimating transit trips in the cellphone data.
Validation Results (5)

Cellphone inferred transit trips vs. smartcard trips, AM Peak

Cellphone inferred transit trips vs. smartcard trips, PM Peak

Cellphone inferred transit trips vs. smartcard trips, Mid-Day
Conclusions

Next Steps?
Primary Study Conclusions

- O-D trip matrices can be constructed at the traffic zone level from both cellphone trace and smartcard data.
- Trip mode can be inferred for cellphone data using a neural network classifier trained on MHMS survey data.
  - This is a “first of the kind” result.
  - Classifier worked well for auto vs non-auto trips
- Home & work locations can be inferred for both cellphone trace and smartcard data, providing that records are available for a sufficient number of days to identify trip patterns.
Caveats & Further Analysis (1): Transit Mode Share

- Transit trips are currently being over-predicted for the cellphone traces; largely due to “confusion” with active trip-making.
  - This is a not unexpected result.
- NEXT STEP: We are now trying to add smartcard data to the analysis to improve transit trip classification.
Caveats & Further Analysis (2): Home & Work Locations

- Home & work locations currently are identified based on regularity of trip patterns within the cellphone traces & smartcard records.
- Not all cases can be identified with certainty.
- NEXT STEP: Use the records for which home (work) locations are identified with high confidence as a labelled training set to develop a neural net classifier that can be applied to the remaining record to improve their inferred home (work) locations.
Caveats & Further Analysis (3): O-D Trip Matrices

- MHMS, cellphone traces & smartcard records all provide O-D trip matrices.
- Each, however, is a sample from the actual population of trips.
- Each has its sampling errors:
  - MHMS is detailed & precise but very small sample (& people may under-report short and/or non-home-based trips).
  - Cellphone traces: Not all people have cellphones; not all people have cellphones on all the time; double counting for people two cellphones.
  - Smartcard records: Only transit trips; destinations need to be inferred.
NEXT STEP: Attempt to combine all data to yield a best estimate of total trip making (by mode, time of day & trip purpose) for Montevideo.
- Need to account for possible double-counting of trips across the datasets.
Thank you.

Questions?

a.faghihimani@utoronto.ca