Extracting mobility behavior from cell phone data
DATA SIM Summer School 2013

PETER WIDHALM
Mobility Department
Dynamic Transportation Systems
T +43(0) 50550-6655 | F +43(0) 50550-6439
peter.widhalm@ait.ac.at | http://www.ait.ac.at

MARKUS PIFF
Mobility Department
Dynamic Transportation Systems
T +43(0) 50550-6021 | F +43(0) 50550-6439
markus.piff@ait.ac.at | http://www.ait.ac.at
AIT Austrian Institute of Technology
Research Focus

- EFFICIENT
- SAFE
- GREEN
Research Focus

- EFFICIENT
- SAFE
- GREEN

Dynamic Transportation Systems:
- Multimodal transport
- Traffic Flows
- Human Factors
- Data acquisition & analysis
- Simulation & optimization
Goals of our research

- Aid travel demand modeling by providing information about mobility behavior extracted from cell phone data
- Integrate with PTV VISUM
- Validate results
Background and Motivation

- Travel demand modeling
  - estimation of demand flows
  - **Demand flow**: aggregation of trips
    - from an origin (zone)
    - to a destination (zone)
    - within a time interval
    - using a specific mode of transport

→ OD matrices
Background and Motivation

- **Trips result from choices**
  - frequency and timing
  - destination
  - mode and route

  \[ \rightarrow \text{System of submodels} \]
  (typically 4-step model)

- **In each step either**
  - **“snapshot“ of current situation or prior knowledge**
    - no model, no assumptions, no predictions
    - some sort of extrapolation
  - **prediction models**
    - behavior explained by socioeconomic attributes, activity and transport supply
    - e.g. Random Utility Models
Background and Motivation

- **Required data:**
  - *average number of trips per person for given origin and purpose*

- **Model assumption:**
  - Activity needs governed by socioeconomic attributes → behavioral groups w.r.t. trip purpose
  - Travel needs governed by activity supply

  → (observed) behavior (clusters) explained by socioeconomic attributes and attributes of trip origin
Background and Motivation

- Required data:
  - distribution of trips starting at given origin and with given purpose over destinations

- Model assumption:
  - Number of trips attracted by a destination is governed by
    - Costs to reach destination (from given origin): travel time, monetary costs, …
    - Activity supply of destination (w.r.t. given trip purpose)

  → (observed) behavior explained by supply attributes of trip destination and travel „costs“ between origin and destination
Background and Motivation

- Required data:
  - distribution of trips with given origin, purpose, destinations over modes of transport

- Model assumption:
  - Mode choice is governed by purpose, trip length, transportation supply and socioeconomic attributes (age, income, car ownership, …)
  - (observed) behavior explained by mode specific travel „costs“ (transport supply) and socioeconomic attributes (and mode specific „preference constants“)
Background and Motivation

- Required data:
  - level-of-service (travel times, delays) of route alternatives depending on flow volume

- Model assumption:
  - Route choice is governed by level-of-service (travel times, number of stops, monetary costs, …).
  - Level-of-service can depend on flow volumes (street networks)

  → (observed) behavior explained by level-of-service (→ Volume Delay functions)
Background and Motivation

- Required data:
  - Level of service (travel times, delays) of route alternatives depending on flow volume
- Model assumption:
  - Route choice is governed by level of service (travel times, number of stops, monetary costs, …).
  - Level of service can depend on flow volumes (street networks).
  - ▸ (observed) behavior explained by level-of-service (Volume Delay functions)
Background and Motivation

- **Traditional surveys**
  - expensive
  - limited sample size
  - no up-to-date data

- → **new technologically aided methods** are needed!
  - e.g.
    - GPS tracker
    - cell phones
Using cell phone data for travel surveys

Active techniques (GPS, sensor data)

- requires Smartphone app
- burden on participant
- costs
- → sample size still limited

Passive techniques (telecom network traffic)

- infrastructure already in place
- no recruiting required
- large sample size!
Cell phone data – Passive techniques

- **Call Detail Records (CDRs)**
  - billing data
  - better availability
  - depend on phone usage behavior

- **Cellular network protocol events**
  - better reconstruction of actual trajectory:
    - **motion-triggered records** (Location/Routing Area Updates)
      - whenever device crosses Location/Routing Area border
    - **time-triggered records** (periodic location updates, ~3 or 6 hours)
      - guaranteed minimum location update frequency
Approach – passive technique

- Reconstruction of stops and trips
  - inference of arrival / departure time (→ duration) + uncertainties
    - based on min. travel times
  - location clustering
    - location labels within day trajectory, e.g.: A-B-C-A-D-A
Approach – passive technique

- Stop patterns can be clustered to find daily mobility patterns: „Motifs“

Approach – passive technique

- Assign attributes to stops and trips

- earliest/latest arrival/departure
- min/max duration
- traffic zone
  - land use shares + POI categories
  - sociodemographics
- earliest/latest departure
- min/max duration
- distance / length
Approach – passive technique

- Inference of activities
  - based on
    - probability of activity pattern \( m = (a_{m,1}, ..., a_{m,n}) \)
      - prior \( p(m) \)
      - location sequence \( s \) \( p(s|m) \)
      - probability of activities \( a_{m,i} \)
        - time \( t_i \) \( p(t_i|a_{m,i}) \)
        - duration \( d_i \) \( p(d_i|a_{m,i}) \)
        - land use / POIs \( l_i \) \( p(l_i|a_{m,i}) \)

\[
p(m|s) = \frac{p(m)p(s|m)}{\sum_{m'\in M} p(s|m')p(m')}
\]

\[
p(a_{m,i}|t_i, d_i, l_i) = \frac{p(t_i|a_{m,i})p(d_i|a_{m,i})p(l_i|a_{m,i})}{\sum_{a'\in A} p(t_i|a')p(d_i|a')p(l_i|a')}
\]

\[
p(m|s, t, d, l) = p(m|s) \prod_i p(a_{m,i}|t_i, d_i, l_i)
\]
Approach – passive technique

- Inference of activities
  - based on
    - probability of activity pattern $m$
      - prior $p(m)$
      - location sequence $s$ $p(s|m)$
      - probability of activities $a_{m,i}$
        - time $t_i$ $p(t_i|a_{m,i})$
        - duration $d_i$ $p(d_i|a_{m,i})$
        - land use / POIs $l_i$ $p(l_i|a_{m,i})$

$$p(m|s) = \frac{p(m)p(s|m)}{\sum_{m' \in M} p(s|m')p(m')}$$

$$p(a_{m,i}|t_i, d_i, l_i) = \frac{p(t_i|a_{m,i})p(d_i|a_{m,i})p(l_i|a_{m,i})}{\sum_{a' \in A} p(t_i|a')p(d_i|a')p(l_i|a')}$$

$$p(m|s, t, d, l) = p(m|s) \prod_{i} p(a_{m,i}|t_i, d_i, l_i)$$

- $f$requency of activity patterns, e.g.
  - Home-Work-Home 18.6%
  - Home-Leisure-Home 11.4%
  - Home-Shop-Home 11.0%
  - Home-Work-Shop-Home 1.0%
Approach – passive technique

- Inference of activities
  - based on
    - probability of activity pattern
      - prior
      - location sequence $s$
      - probability of activities $a_{m,i}$
        » time $t_i$
        » duration $d_i$
        » land use / POIs $l_i$

\[
p(m|s) = \frac{p(m)p(s|m)}{\sum_{m' \in M} p(s|m')p(m')}
\]

\[
p(a_{m,i}|t_i, d_i, l_i) = \frac{p(t_i|a_{m,i})p(d_i|a_{m,i})p(l_i|a_{m,i})}{\sum_{a' \in A} p(t_i|a')p(d_i|a')p(l_i|a')}
\]

\[
p(m|s, t, d, l) = p(m|s) \prod_{i} p(a_{m,i}|t_i, d_i, l_i)
\]

$A$-$B$-$C$-$A \neq \text{Home-Work-Home}$

$\approx \text{Home-Work-Shop-Home}$

$\approx \text{Home-Leisure-Leisure-Home}$
Approach – passive technique

- Inference of activities
  - based on
    - probability of activity $m_i$ pattern
      - prior $p(m)$
      - location sequence $s$
      - probability of activities $a_{m,i}$
        » time $t_i$ $p(t_i|a_{m,i})$
        » duration $d_i$ $p(d_i|a_{m,i})$
        » land use / POIs $l_i$ $p(l_i|a_{m,i})$

\[
p(m|s) = \frac{p(m)p(s|m)}{\sum_{m' \in M} p(s|m')p(m')}
\]

\[
p(a_{m,i}|t_i, d_i, l_i) = \frac{p(t_i|a_{m,i})p(d_i|a_{m,i})p(l_i|a_{m,i})}{\sum_{a' \in A} p(t_i|a')p(d_i|a')p(l_i|a')}
\]

\[
p(m|s, t, d, l) = p(m|s) \prod_{i} p(a_{m,i}|t_i, d_i, l_i)
\]
Approach – passive

- Inference of activities
  - based on
    - probability of activity pattern
      - prior
      - location sequence
      - probability of activities
        - time
        - duration
        - land use / POIs
  - \( p(l_i|a_{m,i}) \)

\[
p(m|s) = \frac{p(m)p(s|m)}{\sum_{m' \in M} p(s|m')p(m')}
\]

\[
p(a_{m,i}|t_i, d_i, l_i) = \frac{p(t_i|a_{m,i})p(d_i|a_{m,i})p(l_i|a_{m,i})}{\sum_{a' \in A} p(t_i|a')p(d_i|a')p(l_i|a')}
\]

\[
p(m|s, t, d, l) = p(m|s) \prod_i p(a_{m,i}|t_i, d_i, l_i)
\]
Approach – passive technique

- Inference of activities
  - based on
    - probability of activity pattern \( m = (a_{m,1}, \ldots, a_{m,n}) \)
      - prior \( p(m) \)
      - location sequence \( s \) \( p(s|m) \)
      - probability of activities \( a_{m,i} \)
        - time \( t_i \) \( p(t_i|a_{m,i}) \)
        - duration \( d_i \) \( p(d_i|a_{m,i}) \)
        - land use / POIs \( l_i \) \( p(l_i|a_{m,i}) \)

\[
p(m|s) = \frac{p(m)p(s|m)}{\sum_{m' \in M} p(s|m')p(m')}
\]

\[
p(a_{m,i}|t_i, d_i, l_i) = \frac{p(t_i|a_{m,i})p(d_i|a_{m,i})p(l_i|a_{m,i})}{\sum_{a' \in A} p(t_i|a')p(d_i|a')p(l_i|a')}
\]

\[
p(m|s, t, d, l) = p(m|s) \prod_i p(a_{m,i}|t_i, d_i, l_i)
\]
Approach – passive technique

- Inference of activities

  - **iterative EM-like clustering**
    - alternation of probabilistic assignment and parameter update

  - **initialization:** prior beliefs
    - previous surveys
    - active surveys with small sample size
    - literature
    - „rule-of-thumb“
Estimation of Travel Demand Flows (OD matrices)

- Simple approach: “snapshot” of current demand flows
  - Infer home location
  - Extrapolate number of OD-trips with known average number of trips produced in home location (based on number of residents and trip frequency) and known modal share
  - Implicit assumption: residents in same traffic zone have similar travel behavior

Estimation of Travel Demand Flows (OD matrices)

Validation – passive technique

- Comparison of **travel times** and **flow volumes** in the street network

**measurements:**

- Floating Car Data
- Traffic Detectors

**model output:**

Floating Car Data

Traffic Detectors
Validation – passive technique

- Comparison of **trajectories**, **trips**, **stops**, **activities**

Network Traffic

**passive**

**active**

Smartphone App:
- GPS track
- annotated activities
Approach – active technique

- Mode & path choice:

Approach – active technique

- Mode & path choice:

  Random subspace classifier ensemble

  $$\sum \text{averaged posterior class probabilities}$$

  HMM

Validation:

<table>
<thead>
<tr>
<th>Mode</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bus</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>Car</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>Bike</td>
<td>0.7</td>
<td>0.7</td>
</tr>
<tr>
<td>Tram</td>
<td>0.6</td>
<td>0.6</td>
</tr>
<tr>
<td>Train</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>Subway</td>
<td>0.4</td>
<td>0.4</td>
</tr>
<tr>
<td>Walk</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>Motorcycle</td>
<td>0.2</td>
<td>0.2</td>
</tr>
</tbody>
</table>

8:00 am
8:12 am
8:24 am
8:29 am
Summary

Network Traffic

passive

active

Smartphone App

![Graph showing trips and modes]

Trip Production → Trip Distribution → Mode Choice → Route Assignment

8:00 am
8:12 am
8:24 am
8:29 am
Extracting mobility behavior from cell phone data
DATA SIM Summer School 2013

PETER WIDHALM
Mobility Department
Dynamic Transportation Systems
T +43(0) 50550-6655 | F +43(0) 50550-6439
peter.widhalm@ait.ac.at | http://www.ait.ac.at

MARKUS PIFF
Mobility Department
Dynamic Transportation Systems
T +43(0) 50550-6021 | F +43(0) 50550-6439
markus.piff@ait.ac.at | http://www.ait.ac.at