Revealed preference data from WiFi traces for pedestrian activity scheduling

Antonin Danalet
TRANSP-OR, EPFL

8th Workshop on Discrete Choice Models
May 31st, 2013
Presentation outline

- Motivation
- Data requirement
- Methodology
- A case study on EPFL campus
- Conclusion
- Future work
MOTIVATION
Lausanne railway station
Lausanne railway station

- 85’000 passengers today,
  170’000 in 2030
- 40% in the metro
- New underpass
- Mixed traffic / only pedestrians in front of the station?
- Léman 2030: 10-year project
Understand pedestrian activities

What we are doing: Campus

What we want to do: Station

Legend
Activities
Weighted shortest path
Pedestrian network

1. Restaurant
2. Cafeteria
3. Classroom
4. Author's office
5. Ticket machine
6. Platform 9
7. First signal
8. Supermarket
9. Kiosk

TRANSP-OR

A. Danalet

TRANSP-OR ENAC EPFL
DATA REQUIREMENT
Data requirement

- **Required**
  - Localization data with full coverage of the facility
  - Pedestrian network

- **Not really required but often available information**
  - Prior aggregated occupancy
Data requirement: Localization

- Data from communication network infrastructure
  - GSM traces (Calabrese et al. 2011, Bekhor et al. 2011)
  - WiFi traces

- Post-processing is needed (Rieser-Schüsseler 2012)
  - Detection of stop points
  - Activity purpose detection through land-use information and spatial matching
Data requirement: Pedestrian network

- We need maps
  - With points of interests and destinations
  - With turn-by-turn directions
  - With multi-floor management

- More and more available in airports, malls, museums, campuses, hospitals
  - Microsoft: 2700 indoor maps
  - Google: > 10’000 indoor maps
  - Start-ups: Wifarer, Meridian, Point Insider, ByteLight
Data requirement: Prior occupancy

- The *a priori* number of people $C(x,t)$ who are performing an activity
  - Capacity (nb of seats), registered students, expected passengers, capacity of a scene, point-of-sale data, …
- At each destination $x$
  - Classroom, platform, scene, …
- At any time $t$
  - Class schedules, train schedules, opening hours, …
- Example: 1500 passengers on platform 4 at 16h04
  32 students in GC B3 31 at 17h15
METHODOLOGY
Methodology

- **Goal**: extract the possible activity-episodes performed by pedestrians from digital traces from communication networks

**Input**
- Localization measurement
- Prior occupancy
- Pedestrian map

**Output**
- set of candidate activity-episodes sequences associated with the probability of being the true one
Methodology

- Probabilistic measurement model: A Bayesian approach
  - Measurement equation
  - Prior
- Generation of activity-episode sequences
  - Episode location
  - Episode start and end times
- Intermediary signals
- Sequence elimination procedure
Definitions / Notations

- Measurement: \( \hat{s} = (\hat{x}, \hat{t}) \)
- Activity-episode: \( a = (x, t^-, t^+) \)
- Episode location, start time and end time
- Activity-episode sequence: \((a_1, \ldots, a_m) = a_{1:m}\)
- Activity: \( A(a) \)
- Activity pattern: \((A_1, \ldots, A_m) = A_{1:m}\)
Probabilistic measurement model

\[ P(a_{1:m} \mid \hat{s}_{1:n}) \propto P(\hat{s}_{1:n} \mid a_{1:m}) \cdot P(a_{1:m}) \]

- Measurement likelihood
- Prior
- Activity model
Probabilistic measurement model

\[
P(\hat{s}_{1:n}|a_{1:m}) = \prod_{j=1}^{m} P(\hat{s}_{i_{j-1}+1:i_{j}}|a_{j}) \\
= \prod_{j=1}^{m} \prod_{i=1}^{n} P(\hat{s}_{i_{j}}|a_{j}) \\
= \prod_{j=1}^{m} \prod_{i=1}^{n} P(\hat{x}_{i_{j}}|x_{j})
\]
Prior

\[ P(a_1:m) = \prod_{j=1}^{m} P(a_j) \quad (1) \]

\[ = \prod_{j=1}^{m} P(x_j, t_j^-, t_j^+) \quad (2) \]

\[ = \prod_{j=1}^{m} \frac{C_{x_j}(t_j^-, t_j^+)}{\sum_{x \in X} C_x(t_j^-, t_j^+)} \quad (3) \]

\[ = \prod_{j=1}^{m} \int_{t_j=t_j^-}^{t_j^+} \frac{C_{x_j}(t_j)}{\sum_{x \in X} C_x(t_j)} dt \quad (4) \]
Generation of activity-episode sequences
Domain of data relevance

\[
floor(\hat{x}) + 1 \\
floor(\hat{x}) \\
floor(\hat{x}) - 1
\]

25 meters

\[
\min(cF, 80 \text{ meters})
\]
Generates of activity-episode sequences

\[
\begin{align*}
t_i^- &\sim U(t_i, t_{i+1} - t_{x_i,x_{i+1}}) \\
t_i^+ &\sim U(t_i^+, t_{x_i,x_{i+1}} + t_{x_{i+1},x_{i+1}})
\end{align*}
\]
Generation of activity-episode sequences

\[ f(t_{i+1}^-) = \frac{1}{\hat{t}_{i+1} - tt_{x_i,x_{i+1}} - \hat{t}_i} \ln \frac{\hat{t}_{i+1} - tt_{x_i,x_{i+1}} - \hat{t}_i}{\hat{t}_{i+1} - t_{i+1}^-} \]

\[ E(t_{i+1}^-) = \frac{\hat{t}_i + tt_{x_i,x_{i+1}}}{4} + \frac{3 \cdot \hat{t}_{i+1}}{4} \]
Generation of activity-episode sequences

\[ t_{x_i, x_{i+1}} = \frac{\text{dist}(x_i, x_{i+1})}{v} \]

- Distance: shortest path in the pedestrian graph
- Speed: 1.34 meters/second (Buchmueller and Weidmann, 2006)
Generation of activity-episode sequences

\[ a_i^1 \rightarrow a_i^2 \rightarrow a_i^3 \rightarrow a_{i+1}^1 \rightarrow a_{i+1}^2 \rightarrow a_{i+1}^3 \]
Intermediary signals

- Eliminate intermediary signal if

\[ E(t^+) - E(t^-) < T_{\text{min}} \]

since we generate an activity episode at each signal.
Sequence elimination

\[ E(t) - E(t - T_{\text{min}}) \]
A CASE STUDY ON EPFL CAMPUS
EPFL data: Localization

- 8 participants for 2 months with known ID
- Non-participants: 46 days, but only 10 with courses
  - 200 students in 6 different classes
  - 317 employees
  - 700 students from University of Lausanne
- For 151 GC students, 152’598 observations
EPFL data: Pedestrian network

- Source: map.epfl.ch
- 56’655 edges
- 4 different levels of path
  - Major (« highway »)
  - Inter-building
  - Intra-building
  - Access to offices
- Shortest path
- All offices, restaurants, classrooms and other points of interest are coded: X
EPFL data: Potential occupancy $C(x,t)$

- Class schedules with
  - Number of students
  - Name of the classroom
- Number of employees per office
  - Name of the office
  - Sum of percent of work (e.g., 3 full times = 300%)
- Number of seats in restaurants
  - Localization
  - Opening hours
- Number of seats in library
Generation of activity-episode sequences

Domain of data relevance for each signal

- Points of interest
- Offices and classrooms

Total

Signals
## Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Truth</th>
<th>Δx</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Arrival time</strong></td>
<td><strong>Departure time</strong></td>
<td><strong>Floor</strong></td>
</tr>
<tr>
<td>8:33-8:33</td>
<td>10:38-10:38</td>
<td>1</td>
</tr>
<tr>
<td>10:40-10:40</td>
<td>11:51-11:51</td>
<td>3</td>
</tr>
<tr>
<td>11:54-11:54</td>
<td>12:47-12:53</td>
<td>1</td>
</tr>
<tr>
<td>12:51-12:58</td>
<td>13:03-13:44</td>
<td>3</td>
</tr>
<tr>
<td>13:06-13:47</td>
<td>13:53-14:02</td>
<td>2</td>
</tr>
<tr>
<td>13:55-14:04</td>
<td>19:45-19:45</td>
<td>3</td>
</tr>
</tbody>
</table>
Results
CONCLUSION
Conclusion

- Prior is needed to **overcome low precision**: make use of non-localization aggregate data, with theoretical or no time dimension in it
- **Localization data brings dynamics** in the model: individual sequences with start and end times
- Our methodology is **merging** these two different types of data
- Robust for **low density data** as well
FUTURE WORK
Further work

- Binary choice model for class attendance
  - No use of the dynamic part of the sequence
- Analysis of the access to campus
  - First and last destination of the sequence
  - Arrival times on campus, departure times
- Based on class attendance and on available time budget: activity scheduling
References