Inferring trip destinations in transit smart card data using a probabilistic topic model

Zhanhong Cheng\textsuperscript{1,3}, Martin Trépanier\textsuperscript{2,3}, Lijun Sun\textsuperscript{1,3}*

\textsuperscript{1}McGill University
\textsuperscript{2}Polytechnique Montréal
\textsuperscript{3}CIRRELT

Transitdata2019
July 8, 2019
1. Introduction

2. Methodology

3. Guangzhou metro case study

4. Result Analysis

5. Summary
Origin and destination in transit system

Methods for obtaining transit OD matrix

- **OD survey**
  - Time-consuming and costly.
  - Small samples.

- **Smart card data**
  - High sample rate (for most systems).
  - **No destination** information (for most systems).
Existing destination inference methods by smart card data

Rule-based model\textsuperscript{1} (80±% trips)

- Consecutive trips are connected.
- The last destination in a day = first origin of that day.
- The last destination in a day = first origin the next day.

Using consecutive trips to infer a destination.

\textsuperscript{1}Barry et al. 2002; Trépanier et al. 2007.

\textsuperscript{2}He and Trépanier 2015.
Existing destination inference methods by smart card data

Rule-based model\(^1\) (80±% trips)

- Consecutive trips are connected.
- The last destination in a day = first origin of that day.
- The last destination in a day = first origin the next day.

Using consecutive trips to infer a destination.

Individual’s trip-history-based model\(^2\) (an extra 10±% trips)

- Using historical trips with similar origin and departure time.

---

\(^1\)Barry et al. 2002; Trépanier et al. 2007.
\(^2\)He and Trépanier 2015.
Motivation

Current methods
- Individual-based.
- Not applicable to isolated, unseen trips.

Motivation of our work
- Borrow information from not only individual, but also similar travellers.
- Fully utilize spatial and temporal information.
1. Introduction

2. Methodology

3. Guangzhou metro case study

4. Result Analysis

5. Summary
Topic model

Topic model in Natural Language Processing

• Each document has a **topic distribution**. (such as a document is 90% about sport and 10% about tech)
• Each topic has a **topic-word** distributions.

---

3Blei 2012.

Zhanhong Cheng  (McGill University)
Topic model for smart card data

- All trips of one individual \( \Rightarrow \) a document.
- Origin, destination, and departure time \( \Rightarrow \) three types of words.

**Topic-“word” distribution**

- **Time distribution under time topic \( j \)**
  
- **Origin distribution under origin topic \( k \)**

- **Destination distribution under destination topic \( l \)**

\[
p(w^t|z^t_j) \sim \text{Multinomial}(\varphi_{z^t_j}) \\
p(w^o|z^o_k) \sim \text{Multinomial}(\psi_{z^o_k}) \\
p(w^d|z^d_l) \sim \text{Multinomial}(\omega_{z^d_l})
\]
Topic model for smart card data

- **Topic distribution.** Let \( z_{j,k,l} \) denote \( z^t_j, z^o_k, z^d_l \); 
  \( p(z) \sim \text{Multinomial}_{J \times K \times L}(\theta_u) \).
• **Topic distribution.** Let \( z_{j,k,l} \) denote \( z^t_j, z^o_k, z^d_l \); 
\[ p(z) \sim \text{Multinomial}_{J \times K \times L}(\theta_u). \]

• **Topic-“word” distribution:** Multinomial distribution.
Topic model for smart card data

- **Topic distribution.** Let $z_{j,k,l}$ denote $z_{j}^{t}$, $z_{k}^{o}$, $z_{l}^{d}$; $p(z) \sim \text{Multinomial}_{J \times K \times L}(\theta_u)$.

- **Topic-“word” distribution:** multinomial distribution.

- **Probability for a trip:**

  $$p(t, o, d) = \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{l=1}^{L} p(w^t = t | z_{j}^{t}) \ p(w^o = o | z_{k}^{o}) \ p(w^d = d | z_{l}^{d}) \ p(z_{j}^{t}, z_{k}^{o}, z_{l}^{d})$$
Priors:
\[ \theta_u \sim \text{Dirichlet}(\alpha) \]
\[ \varphi \sim \text{Dirichlet}(\beta) \]
\[ \psi \sim \text{Dirichlet}(\gamma) \]
\[ \omega \sim \text{Dirichlet}(\eta) \]

Topic:
\[ p(z) \sim \text{Multinomial}(\theta_u) \]

Topic-“word”:
\[ w^t \sim \text{Multinomial}(\psi_{z^t_j}) \]
\[ w^o \sim \text{Multinomial}(\varphi_{z^o_k}) \]
\[ w^d \sim \text{Multinomial}(\omega_{z^d_l}) \]
The topic model — inference and learning

Destination inference

- Learning topic distribution for passenger $u$.
- Given $o$, $t$, sum over topic distribution:

\[
P (d|o,t; u) \propto P (d, o, t; u)
\]

\[
= \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{l=1}^{L} p (t|z_j^t) p (o|z_k^o) p (d|z_l^d) p (z_j^t, z_k^o, z_l^d; u)
\]  (1)

Model learning

- Based on known $(t, o, d)$ pair (could from survey or rule based model), use Gibbs sampling to obtain the parameters.
Guangzhou metro case study

Data & processing

- Three months, July 1\textsuperscript{st} – Sep 30\textsuperscript{th} 2017.
- 159 stations, both boarding and alighting registered.
- Randomly select 3000 passengers (with 20+ trips).
- Total 200,670 trips.
- Use 70% for training, 30% for testing (wipe out destinations).

The distribution of the number of trips per person
Comparing with trip-history-based models

For user $u_i$ with $o_{ij}$ and $t_{ij}$, predict $\hat{d}_{ij}$ as the most frequent $d$ in historical similar trips. Four rules to define a similar trip:

1. $o = o_{ij}$;
2. $t = t_{ij}$;
3. $o = o_{ij}, t = t_{ij}$, then $o = o_{ij}$;
4. $o = o_{ij}, t = t_{ij}$, then $t = t_{ij}$.

(Undetermined destinations are inferred by the most frequent destination of user $u_i$.)

## Initial result

### Our model

The prediction accuracy under different number of spatial latent topics when $J = 5$

<table>
<thead>
<tr>
<th># origin topics $K$</th>
<th>10</th>
<th>30</th>
<th>50</th>
<th>80</th>
<th>100</th>
<th>130</th>
<th>150</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>12.1%</td>
<td>26.6%</td>
<td>37.4%</td>
<td>51.4%</td>
<td>56.0%</td>
<td>61.5%</td>
<td>62.2%</td>
</tr>
<tr>
<td>30</td>
<td>12.7%</td>
<td>27.0%</td>
<td>38.2%</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>50</td>
<td>12.9%</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>80</td>
<td>11.9%</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

### Four trip-history-based models

The best model with accuracy $68.52\%$
Analysis of initial result

- High dimensions for destination topic, low accuracy.
- Between passengers, discrepancy is more significant than similarity.
- Individual’s trips have very high regularity.
• Forecast the **rank of the destination**.
• Each person has a rank-to-station dictionary.

![Count of a passenger's visited stations](image)
1. Introduction

2. Methodology

3. Guangzhou metro case study

4. Result Analysis

5. Summary
Improved result

<table>
<thead>
<tr>
<th># Topics</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 3 3</td>
<td>66.5%</td>
</tr>
<tr>
<td>4 3 3</td>
<td>66.6%</td>
</tr>
<tr>
<td>5 3 3</td>
<td>66.6%</td>
</tr>
<tr>
<td>4 4 4</td>
<td>65.7%</td>
</tr>
</tbody>
</table>

- Significantly less topics.
- Comparable accuracy to trip-history-based method.
When not using the ground truth for training

- In reality, no ground-truth destinations.
- Using rule-based *estimated destinations* for training.
When not using the ground truth for training

- In reality, no ground-truth destinations.
- Using rule-based **estimated destinations** for training.
Topic distribution

Time topic 1

Time topic 2

Time topic 3

Time topic 4
Passenger clustering

Hierarchical clustering of 500 passengers based on their topic distributions
1. Introduction

2. Methodology

3. Guangzhou metro case study

4. Result Analysis

5. Summary
Conclusions

- Topic model for smart card data with \((t, o, d)\) as a word.
- Predicting by the rank of the destination.
- Explainable latent topics.
- A passenger clustering method characterizing the spatial and temporal similarity.

Problems and future directions

- Cold start problems.
- Other than rank, better representation?
- Apply to real data set without destinations.
We thank **exo** (Réseau de transport métropolitain) and **Mitacs** (Grant IT13223) for funding this research.
Thank you.
Any questions?
Barry, James et al. (2002). “Origin and Destination Estimation in New York City with Automated Fare System Data”. In: Transportation Research Record: Journal of the Transportation Research Board 1817, pp. 183–187.

