# An Adaptation of Bayesian Theorem for Passenger Destination Estimation

승객 목적지 추정을 위한 베이지안 이론 적용

Sanghyuck Lee
tkdgur658@gmail.com

2020. 12. 30.

# Outline

- Introduction
- Related Works
- Proposed Method
- Further Research

# Introduction

- Intelligent Transportation Systems (ITS) is an innovative idea in the field of transportation and the transit Origin-Destination matrix (OD matrix) is an essential input for some important ITS applications

- The powerful source data for generating the OD matrix is Automated Fare Collection (AFC) data. However, many AFC systems around the world are entry-only systems and do not include destination information so that essentially we have to go through the Passenger Destination Estimation (PDE) process..
  - Moreover, in the real world, even within the datasets of AFC system that contains alighting information, some of them is often omitted due to reasons such as passengers leaving without tagging.

- Therefore, improving the PDE algorithm based on AFC data can be a great contribution to the entire ITS field.

# Related Works

- For AFC data including only boarding information,
  - From the past, until some time ago, there were mainly algorithmic proposals without good validation.
- For AFC data including both boarding and alighting information,
  - some researchers tried to perform algorithm verification or infer passenger destinations using machine learning technique.

| Card ID | Card style | Route | Stop | Veh. ID | Time | Date | Fare |
|---|---|---|---|---|---|---|---|
| -1760518818 | Monthly card | 115 | 156 | 2798 | 06:27:16 | 2009-07-07 | - |
| -1760518818 | Monthly card | 115 | 3 | 2346 | 18:47:21 | 2009-07-07 | - |
| -1760517522 | Common card | 115 | 19 | 2810 | 08:34:53 | 2009-07-07 | 0.9 |
| -1760517522 | Common card | 115 | 151 | 2808 | 17:47:31 | 2009-07-07 | 0.9 |
| -1760515746 | Monthly card | 115 | 133 | 2794 | 11:56:22 | 2009-07-07 | - |
| -1760515746 | Monthly card | 115 | 133 | 2814 | 18:52:29 | 2009-07-07 | - |
| -1760515714 | Monthly card | 115 | 23 | 2794 | 17:32:51 | 2009-07-07 | - |

AFC data NOT including alighting information

| Transaction Number | Boarding Time (yyyy/mm/dd/hh/mm/ss) | ID of Boarding Location | Alighting Time (yyyy/mm/dd/hh/mm/ss) | ID of Alighting Location |
|---|---|---|---|---|
| 1 | 20150318092455 | 4196066 | 20150318092714 | 4196117 |
| 2 | 20150318214734 | 4196061 | 20150318215035 | 4196065 |
| 3 | 20150318225205 | 12174 | 20150318225855 | 8001017 |
| 4 | 20150318075726 | 70647 | 20150318082805 | 10585 |
| 5 | 20150318070948 | 216 | 20150318084718 | 4101317 |
| ... | ... | ... | ... | ... |

AFC data including alighting information

# Proposed Method

An Adaptation of Bayesian Theorem for Passenger Destination Estimation

- Motivation:
  - A system that can access not only boarding but also **alighting information** has been developed, so that the **posterior probability** can be calculated through Bayesian theory, which shows successful results in many inference problems.

$$p(y|x) = \frac{p(x|y)\,p(y)}{p(x)} \Rightarrow p(s|c) = \frac{p(c|s)\,p(s)}{p(c)} \qquad (1)$$

| Terms | Meaning |
|---|---|
| $C$ | The set of conditions for determining passenger destinations, $C = \{c_1, \ldots, c_n\}$, where $n$ is the number of all the conditions. |
| $S$ | The set of destination candidate stations, $S = \{s_1, \ldots, s_m\}$, where $m$ is the number of all the transit station. |

# Proposed Method

$$p(s|c) = \frac{p(c|s)\,p(s)}{p(c)} = \frac{p(c,s)}{p(c)} \tag{1}$$

$$\mathrm{h}(c) = \operatorname*{argmax}_{s \in S} p(s|c) \ = \operatorname*{argmax}_{s \in S} \frac{p(c,s)}{p(c)} \tag{2}$$

$$= \operatorname*{argmax}_{s \in S} p(c,s) \ = \operatorname*{argmax}_{s \in S} \ p(c_1, \cdots, c_n, s) \tag{3}$$

Using chain rule: $p(c_1, \cdots, c_n, s)$

$$= p(s)p(c_1|y)p(c_2, \cdots, c_n|s, c_1)$$

$$= p(s)p(c_1|s)p(c_2|s)p(c_3, \cdots, c_n|s, c_1, c_2)$$

$$= p(s)p(c_1|s)p(c_2|s) \cdots p(c_n|s, c_1, c_2, c_3, \cdots, c_{n-1})$$

Adapt Naïve assumption:

$$= p(s)\,p(c_1|s)\ p(c_2|s)\ p(c_3|s) \cdots$$

$$= p(s)\ \prod_{i=1}^{n} p(c_i|s) \tag{4}$$

# Proposed Method

By (3) and (4):

$$\begin{aligned} \mathrm{h}(c) &= \underset{s \in S}{\operatorname{argmax}}\, p(s|c) \\ &= \underset{s \in S}{\operatorname{argmax}}\, p(s) \prod_{i=1}^{n} p(c_i|s) \end{aligned} \tag{5}$$

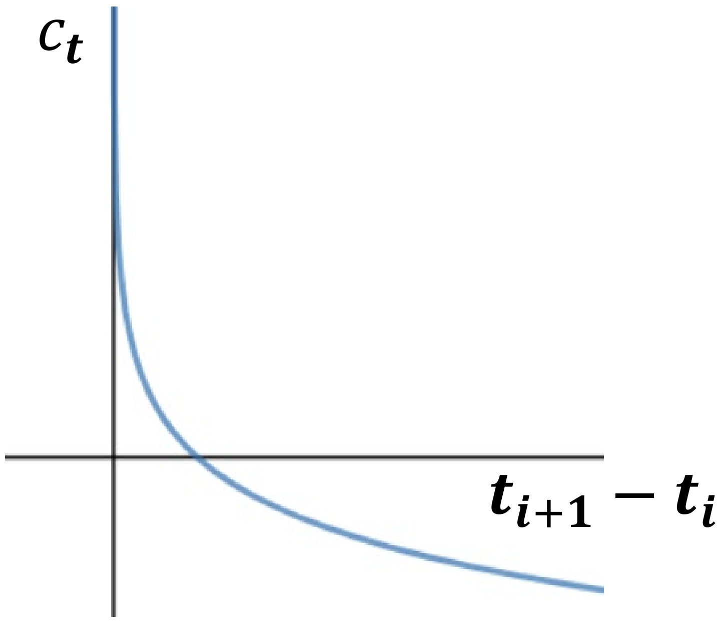
# Proposed Method

- Some conditions are

$$c_t = -log\,(t_{i+1} - t_i) \tag{6}$$

- where $t_i$ is the boarding time of the $i$-th transaction, and

$$c_d = -log\,d_i \tag{7}$$



# Proposed Method

- Some conditions are

$$c_t = -log\,(t_{i+1} - t_i) \tag{6}$$

- where $t_i$ is the boarding time of the $i$-th transaction, and

$$c_d = -log\, d_i \tag{7}$$

- where $d_i$ is the distance between the transit route of the $i$-th transaction and candidate station.

| Station 1 | Station 2 | Station 3 | Station 4 | Station 5 | Station 6 | Station 7 | Station 8 | Station 9 |
|---|---|---|---|---|---|---|---|---|
| a % | b % | c % | d % | e % | f % | g % | h % | i % |

- Checking the variation in the probability distribution of destination for each transaction

# Further Research

- Implement the comparison algorithm

- Ideas for some conditions

- AFC data process

# Appendix

# Preliminary: a simple PDE example

- Use 2 simple assumptions:
  1) The destination of all the trips is the origin of the next trip.
  2) The last destination is the first origin.

| Terms | Meaning |
|---|---|
| $X$ | The AFC dataset consisting of card ID, boarding station and boarding time, $X = \{x_1, \dots, x_n\}$, $x$=$(x^c, x^s, x^t)$, where $n$ is the number of all the transactions. |
| $S$ | The set of all the transit station in $X$, $S = \{s_1, \dots, s_m\}$, where $m$ is the number of different transit station in $X$. |
| $C$ | The set of all the unique card ID in $X$, $C = \{c_1, \dots, c_l\}$, where $l$ is the number of unique card ID in $X$. |
| $x^c, x^s, x^t$ | Card ID, boarding station and boarding time of the transaction $x$ respectively, $\mathrm{x}^c \in C$, $\mathrm{x}^s \in S$, $\mathrm{x}^\mathrm{t} = [0{,}1440]$. |
| $T$ | The set of all the trips that should be estimated from $X$, $T = \{\mathrm{t}_1, \dots, \mathrm{t}_\mathrm{n}\}$, $t = (o, d)$. |
| $o$ | The origin of the trip that should be estimated at $t$ of T. |
| $d$ | The destination of the trip that should be estimated at $t$ of T. |
| $X^c$ | The subset with $x$'s in $X$ that have the same card ID $c$, $X^c = \{x \mid x^c = c\}$. |
| $T^c$ | The set of the trips that will be estimated from $X^c$, that should consist of trips derived from the single card ID. |

# Preliminary: a simple PDE example

**Algorithm 1** A simple PDE algorithm

**Input:** $X, S$;
**Output:** $T$;
Initializing $T$={};
Generate $C$ containing unique card ID from $X$;
**for** c **in** $C$ **do**
  $T^c$ = Estimate($X^c$);
  concat($T$, $T^c$);
**end for**
**return T**

**Algorithm 2** Procedure of the Estimate function

**Input:** $X^c$;
**Output**: $T^c$;
Ascend $X^c$ according to boarding time $x^t$ ;
**for** $(i = 1;\ i < |X^c|;\ i++)$ **do**
  $o_i \leftarrow x_i^s;\ \ d_i \leftarrow x_{i+1}^s$;
**end for**
$o_{|D^c|} \leftarrow x_{|D^c|}^s;\ d_{|D^c|} \leftarrow x_1^s$;
return $T^c$

| | Meaning |
|---|---|
| $X$ | The AFC dataset consisting, $X = \{x_1, \dots, x_n\}$, $x=(x^c, x^s, x^t)$. |
| $S$ | The set of all the transit station in $X$, $S = \{s_1, \dots, s_m\}$, where $m$ is the number of different transit station in $X$. |
| $C$ | The set of all the unique card ID in $X$, $C = \{c_1, \dots, c_l\}$, where $l$ is the number of unique card ID in $X$. |
| $x^c$ | Card ID of $x$, $x^c \in C$. |
| $x^s$ | boarding station of $x$, $x^s \in S$. |
| $x^t$ | boarding time of $x$, $x^t = [0,1440]$. |
| $T$ | The set of all the trips that should be estimated from $X$, $T = \{t_1, \dots, t_n\}$, $t = (o, d)$. |
| $o$ | The origin of the trip |
| $d$ | The destination of the trip |
| $X^c$ | The same card ID subset of $X$, $X^c = \{x \mid x^c = c\}$. |
| $T^c$ | The set of the same card ID trips that will be estimated from $X^c$. |