Estimating Passengers' Path Choice Using Automated Data in Urban Rail Systems

Zhenliang (Mike) Ma[a,b], Haris N. Koutsopoulos[b], Yiwen Zhu[c] , Yunqing Chen[a]

[a] *Department of Civil Engineering, Monash University, Melbourne, Australia*

[b] *Department of Civil and Environmental Engineering, Northeastern University, Boston, United States*

[c] *Microsoft Bay Area, San Francisco, United States*

Abstract

The densification of urban areas has led to a rise in transit demand outpacing the urban rail system capacity in many major cities, which brings issues such as near capacity operations, crowding and safety. Better understanding of passengers' path choice behavior is the prerequisite for effective operations planning. Traditional survey methods are constrained by data coverage, collection cost and reporting accuracy, while Automated Fare Collection (AFC) and Automatic Vehicle Location (AVL) data provides new opportunities for behavior analysis. Though several studies have been estimating path choice using AFC data, they are either limited in applications by estimating path choice fractions (rather than individual choice), or biased in modelling link costs by ignoring the fact that under congested conditions passengers may experience denied boarding at major stations due to capacity constraints. In this paper, we extend the passenger to train assignment model (PTAM) in Zhu, et al. [1] to estimate the individual path choice behaviour by incorporating discrete choice model. The approach is data-driven and requires fare transactions at both entry and exit gates and train operation data at stations. The paper validates the model and demonstrates its applicability using synthetic data. Case studies on actual data are performed to understand passengers' path choice patterns in space and time.

Introduction

Most discrete choice models are based on the random utility maximization (RUM) principle [2]. For estimation, they mainly are revealed and stated preference data [3-6] or data from passive data collection technologies, such as GPS sensors [7] and mobile phone [8]. Surveys are a powerful tool to facilitate behavior analysis. However, they are constrained by high costs, reporting accuracy, and survey coverage, etc.

Automated Fare Collection (AFC) and Automatic Vehicle Location (AVL) data provides opportunities for analysis in areas such as travel behavior, demand modelling, transit operations planning, etc. [9-11]. In addition to aggregate trends of when and where passengers travel, AFC data provides detailed information on the travel patterns of individuals and/or specific groups [12, 13]. Several studies have used AFC data to estimate passengers' path choice probabilities [14-17]. They provide useful insights on the aggregated choice behavior under existing conditions. However, without modelling the individual path choice behavior, they are not useful for operations planning applications, such as timetable design, network expansion, operating strategies and policy interventions, etc.

This study focuses on the estimation of path choice models that are sensitive to performance attributes of the alternatives using AFC and AVL data. Relevant to this context, Sun, et al. [18] developed an integrated Bayesian approach to infer network attributes and passenger route choice behavior using AFC data in a urban rail system. The framework can incorporate various RUM-based discrete choice models. Zhang, et al. [19] developed a data fusion model to estimate individual path choices by combining RP data and AFC data, and modelled the heterogeneous risk attitudes of passengers. However, both studies imposed a strong assumption on link travel times (independent normal distribution) ignoring the fact that under congested conditions passengers may experience denied boarding at major stations due to capacity constraints. During peak periods, a (usually) shorter route may have passengers who are left behind. However, the models above cannot distinguish between a passenger having a longer journey time because they chose a longer route or because they were left behind multiple times on a shorter route [20].

Zhu, et al. [1] proposed a passenger-to-train assignment model (PTAM) by decomposing the journey time into access, wait, in-vehicle, and egress times, and considering the dynamics of being left behind at origin stations explicitly. The model was applied to estimate the left behind at key stations for non-transfer trips with capacity constraints and validated using both synthetic and actual data [21]. Hörcher, et al. [22] extended the PTAM to the case with transfers and presented a discrete choice model to estimate the user cost of crowding in urban rail systems. The 'actual' path that a passenger chose was identified using the probabilistic PTAM results (based on access, egress, and transfer times). The hard identification rule used (e.g. path selected if probability larger than a certain threshold) may bias the choice observations, and eventually impacted the estimation results of the choice model.

Methodology

We consider a closed AFC system where tap-in and tap-out data are available, and train arrival and departure at stations are available from AVL system. Define a passenger trip $x = (o, d, t_b, t_e)$, representing origin, destination, begin and end times. Train movement $y = \left(l, s, t_a, t_g\right)$, representing line, station, arrival and departure times. The day is divided into fixed time intervals Δ , e.g. 15 minutes.

Let $\mathcal{R} = \{R^1, R^2, \cdots, R^M\}$ be the effective path set of an OD pair. Assume that the probability of each path being chosen is stable given a time period h and follows $\boldsymbol{\pi}_h = \{\pi_h^1, \pi_h^2, \cdots, \pi_h^M\}$, where $\sum_{m=1}^{M} \pi_h^m = 1$. Given the path attribute vector $\boldsymbol{z}_h^m$ for path m (e.g. travel time, transfers, crowding, etc.), the path chosen probability π_h^m can be calculated using RUM-based choice model, such as Multinomial Logit Model (MNL).

$$\pi_h^m = \frac{exp(\boldsymbol{\beta}\boldsymbol{z}_h^m)}{\sum_{m'=1}^{M} exp(\boldsymbol{\beta}\boldsymbol{z}_h^{m'})} \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_M)$ are the unknown parameters to be estimated.

Given passenger trips $\mathbf{X} = \{X_{od} | \forall o \in O, \forall d \in D\}$ for an OD pair at time τ_h, where $X_{od} = \{x^1, x^2, \cdots, x^Q\}$, and train operation table $\mathrm{Y} = \{y_{ls} | \forall l \in L, \forall s \in S\}$, the objective is to estimate path choice model parameters $\boldsymbol{\beta}$, by maximizing the likelihood of observing all trips for all OD pairs with path choice sets.

$$L(\mathbf{X}, \mathrm{Y}, \boldsymbol{\beta}) = \prod_{X_{od} \in \boldsymbol{X}} L(X_{od}, \mathrm{Y}, \boldsymbol{\beta}) \tag{2}$$

The likelihood of observing all trips of an OD pair is calculated as:

$$L(X_{od}, \mathrm{Y}, \boldsymbol{\beta}) = \prod_{x^q \in X_{od}} \left(\sum_{R^m \in \mathcal{R}_{od}} \left(\pi_h^m \times \Pr(x^q.t_e | x^q.t_b, R^m, \mathrm{Y})\right) \right) \tag{3}$$

where $\Pr(x^q.t_e | x^q.t_b, R^m, Y)$ is the possibility that a passenger x^q exits the destination station at time $x^q.t_e$ given he/she enters the origin station at $x^q.t_b$, chooses route R^m.

Considering possible transfers, a path is represented as a sequence of segments ordered by the chronical order $R = (r^1 \rightarrow r^2 \dashrightarrow r^J)$, where path segment $r = (l, s, s')$, is determined by the line, and the starting and ending stations of a segment. Given the path segment, passengers may board different trains. Figure 1 shows the possible movements of a passenger who enters the system at tap-in and exits at tap-out and involves one transfer (two segments). Depending on the walking time and left behind on different segments, the passenger can board trains Line1_Train1, Line1_Train2 or Line1_Train3 for the first segment, and trains Line2_Train1, Line2_Train2 or Line2_Train3 for the second segment. Itineraries represent different combinations of trains for the two segments. The feasible itinerary are the itineraries whose tap-in time plus minimum access time is earlier than the train departure time at the first segment, transfer time is larger than the minimum transfer time requirement, and tap-out time is later than the train arrival time plus the minimum egress time at the last segment. A feasible itinerary is the set of trains that a passenger successfully boarded on each segment $\Omega^i = \left(\Lambda_{r_1} \rightarrow \Lambda_{r_2} \dashrightarrow \Lambda_{r_J}\right)$.

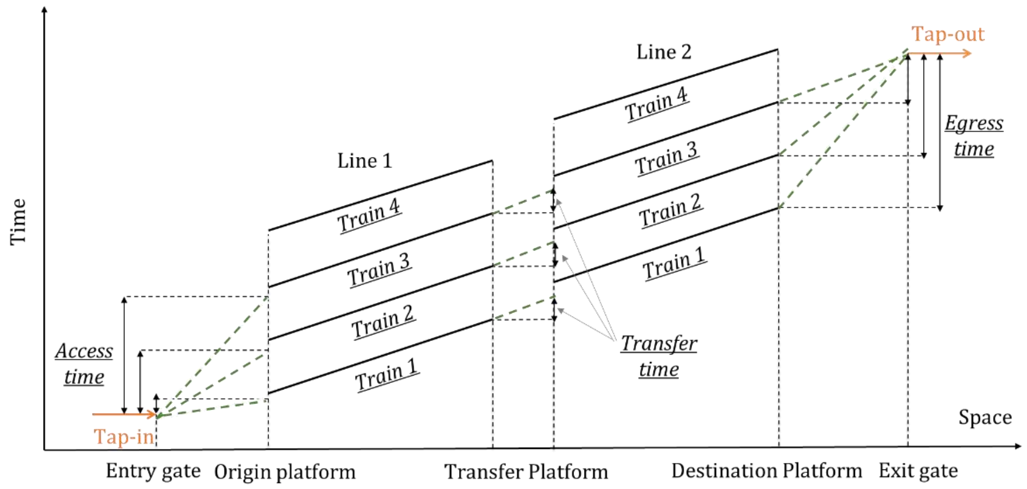


Figure 1: Time-space diagram for a journey involving one transfer

Given a feasible itinerary set $\mathbf{\Omega} = \{\Omega^1, \Omega^2, \cdots, \Omega^I\}$ for a specific path of an OD pair, the probability of observing a trip tap-in and tap-out is:

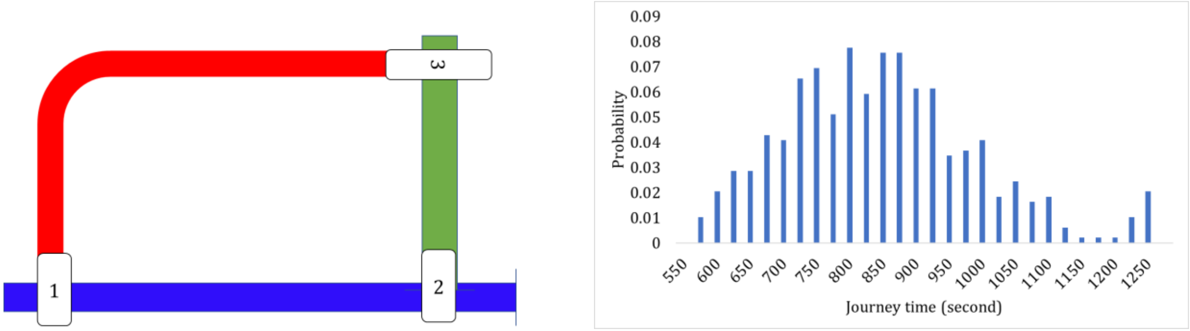
$$\Pr(x^q.t_e | x^q.t_b, R^m, Y) = \sum_{\Omega^i \in \mathbf{\Omega}} \Pr\left(x^q.t_e | \Omega^i, x^q.t_b, R^m, Y\right) \times \Pr\left(\Omega^i\right) \quad (4)$$

Let $f_a(t), f_{tr}(t), f_e(t)$ be the distributions of access time, transfer time and egress time, respectively. Let $\boldsymbol{\theta}_r = \left(\theta_r(1), \theta_r(2), \cdots, \theta_r(N)\right)$ be the probability of left behind different times at path segment r, where N is the maximum left behind times, e.g. 5. The walking time distributions and left behind probabilities can be estimated using the approaches in [20]. We further assume that the left behind probabilities for transfer and new tap-in passengers are similar. Given the known itinerary (train arrival time at the last segment) and tap-out time, $\Pr\left(x^q.t_e|\Omega^i, x^q.t_b, R^m, Y\right)$ can be calculated from $f_e(t)$. $\Pr\left(\Omega^i\right)$ can be derived from $f_a(t), f_{tr}(t)$ and $\boldsymbol{\theta}_r$ using the PTAM formulation in [1].

Case Study

To validate the proposed methodology, synthetic data was generated using the tap-in times and the train movement data from a major urban rail system during the peak period. Note that the proposed methodology can also be used to estimate the path fractions of an OD pair by maximizing the Equation (3) with respect to $\boldsymbol{\pi}_h$. Without loss of generality, the following example was used to illustrate the methodology for fraction estimation.

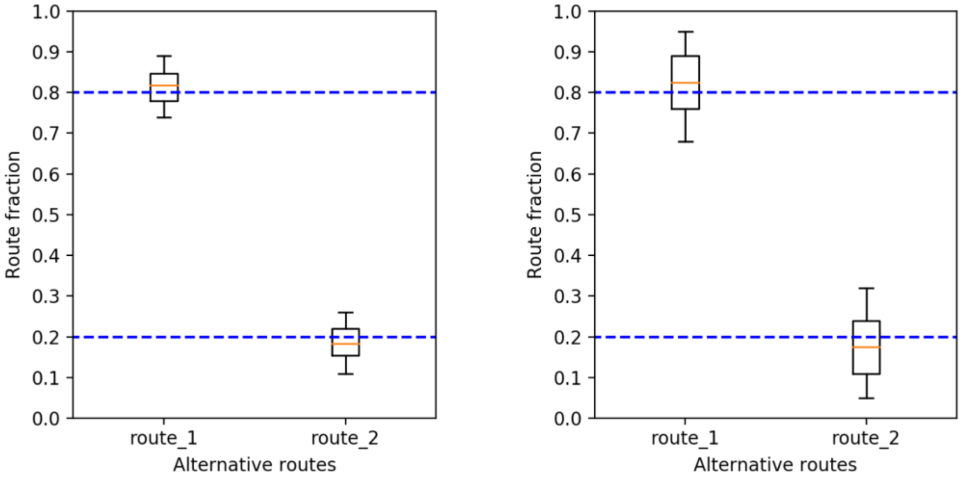
Figure 2a shows the configuration of the network where station 1 is crowded with serious left behind. The tap-in and tap-out data was generated for OD pair 1->3. Passengers were assigned to a path (1-3 or 1-2-3) according to pre-defined fractions (80%, 20%). At station 1, red line, up direction, the probabilities to be left behind 0 time, once, and twice are set to be 20%, 50%, 30%. All the other platforms have no left behind. The walking speed follows lognormal distribution with mean speed 1.2 m/s and standard deviation 0.25 m/s. The walking distance for access and egress is 150m and transfer distance 50m. Figure 2b shows the journey time distribution of the synthetic data.



(a) (b)

Figure 2: Case study settings. a) Network with transfers and route choice; b) Journey time distribution of OD pair 1->3

Figure 3 shows the path fraction estimation results and sensitivity analysis with respect to left behind probabilities (Figure 3a) and walking speed distribution (Figure 3b). For the sensitivity analysis, we randomly add error (0-5%) to the no left behind probability (re-scale others to sum to 1) and walking speed distribution mean and run the experiments for 50 times to draw the boxplot. The results show that the median of the path fraction is close to the actual ones (blue dash line), and the estimation is more sensitive to the accuracy of the walking speed distribution compared to that of the left behind estimations.



(a) (b)

Figure 3: Estimation results and sensitivity with respect to, a) left behind; b) walk speed

The performance of estimating path choice parameters $\boldsymbol{\beta}$ will be examined using the synthetic data. Then, case studies using actual AFC and AVL data will be performed to explore how passengers will change their path choice behavior in response to crowding in both space (OD pairs with different distance) and time (time period of day).

Reference

[1] Y. Zhu, H. N. Koutsopoulos, and N. H. M. Wilson, "A probabilistic Passenger-to-Train Assignment Model based on automated data," *Transportation Research Part B: Methodological,* 2017.

[2] M. E. Ben-Akiva, S. R. Lerman, and S. R. Lerman, *Discrete choice analysis: theory and application to travel demand*. MIT press, 1985.

[3] S. Raveau, J. C. Muñoz, and L. de Grange, "A topological route choice model for metro," *Transportation Research Part A: Policy and Practice,* vol. 45, no. 2, pp. 138-147, 2011/02/01/ 2011.

[4] S. Raveau, Z. Guo, J. C. Muñoz, and N. H. M. Wilson, "A behavioural comparison of route choice on metro networks: Time, transfers, crowding, topology and socio-demographics," *Transportation Research Part A: Policy and Practice,* vol. 66, pp. 185-195, 2014/08/01/ 2014.

[5] F. Jin, E. Yao, Y. Zhang, and S. Liu, "Metro passengers' route choice model and its application considering perceived transfer threshold," *PLOS ONE,* vol. 12, no. 9, p. e0185349, 2017.

[6] Y. Zhang, E. Yao, H. Wei, and K. Zheng, "A constrained multinomial Probit route choice model in the metro network: Formulation, estimation and application," *PloS one,* vol. 12, no. 6, p. e0178789, 2017.

[7] M. Zimmermann, T. Mai, and E. Frejinger, "Bike route choice modeling using GPS data without choice sets of paths," *Transportation Research Part C: Emerging Technologies,* vol. 75, pp. 183-196, 2017/02/01/ 2017.

[8] Z. Wang, S. Y. He, and Y. Leung, "Applying mobile phone data to travel behaviour research: A literature review," *Travel Behaviour and Society,* vol. 11, pp. 141-155, 2018/04/01/ 2018.

[9] M.-P. Pelletier, M. Trépanier, and C. Morency, "Smart card data use in public transit: A literature review," *Transportation Research Part C: Emerging Technologies,* vol. 19, no. 4, pp. 557-568, 2011/08/01/ 2011.

[10] M. Bagchi and P. R. White, "The potential of public transport smart card data," *Transport Policy,* vol. 12, no. 5, pp. 464-474, 2005/09/01/ 2005.

[11] H. N. Koutsopoulos, Z. Ma, P. Noursalehi, and Y. Zhu, "Transit Data Analytics for Planning, Monitoring, Control and Information," in *Mobility Patterns, Big Data and Transportation Analytics*: Elsevier, 2018.

[12] G. Goulet-Langlois, H. N. Koutsopoulos, and J. Zhao, "Inferring patterns in the multi-week activity sequences of public transport users," *Transportation Research Part C: Emerging Technologies,* vol. 64, pp. 1-16, 2016/03/01/ 2016.

[13] A.-S. Briand, E. Côme, M. Trépanier, and L. Oukhellou, "Analyzing year-to-year changes in public transport passenger behaviour using smart card data," *Transportation Research Part C: Emerging Technologies,* vol. 79, pp. 274-289, 6// 2017.

[14] Y. Sun and R. Xu, "Rail Transit Travel Time Reliability and Estimation of Passenger Route Choice Behavior: Analysis Using Automatic Fare Collection Data," *Transportation Research Record,* vol. 2275, no. 1, pp. 58-67, 2012/01/01 2012.

[15] J. Zhao *et al.*, "Estimation of Passenger Route Choice Pattern Using Smart Card Data for Complex Metro Systems," *IEEE Transactions on Intelligent Transportation Systems,* vol. 18, no. 4, pp. 790-801, 2017.

[16] Y. Sun and P. M. Schonfeld, "Schedule-Based Rail Transit Path-Choice Estimation using Automatic Fare Collection Data," *Journal of Transportation Engineering,* vol. 142, no. 1, p. 04015037, 2016.

[17] F. Zhou, J.-g. Shi, and R.-h. Xu, "Estimation method of path-selecting proportion for urban rail transit based on AFC data," *Mathematical Problems in Engineering,* vol. 2015, 2015.

[18] L. Sun, Y. Lu, J. G. Jin, D.-H. Lee, and K. W. Axhausen, "An integrated Bayesian approach for passenger flow assignment in metro networks," *Transportation Research Part C: Emerging Technologies,* vol. 52, pp. 116-131, 2015/03/01/ 2015.

[19] Y. Zhang, E. Yao, J. Zhang, and K. Zheng, "Estimating metro passengers' path choices by combining self-reported revealed preference and smart card data," *Transportation Research Part C: Emerging Technologies,* vol. 92, pp. 76-89, 2018/07/01/ 2018.

[20] Y. Zhu, "Passenger-to-itinerary assignment model based on automated data," Northeastern University, 2017.

[21] Y. Zhu, H. N. Koutsopoulos, and N. H. M. Wilson, "Inferring left behind passengers in congested metro systems from automated data," *Transportation Research Part C: Emerging Technologies,* 2017/11/10/ 2017.

[22] D. Hörcher, D. J. Graham, and R. J. Anderson, "Crowding cost estimation with large scale smart card and vehicle location data," *Transportation Research Part B: Methodological,* vol. 95, pp. 105-125, 1// 2017.