Modeling temporal flow assignment in metro networks using smart card data

Lijun Sun
Future Cities Laboratory
Singapore-ETH Centre
Singapore 138602, Singapore
Email: lijun.sun@ivt.baug.ethz.ch

Jian Gang Jin
School of Naval Architecture, Ocean and Civil Engineering
Shanghai Jiao Tong University
Shanghai 200240, China
Email: jiangang.jin@sjtu.edu.cn

***Abstract*—Understanding passenger flow assignment patterns in a complex metro network is crucial to maintaining service reliability and developing efficient response during disruptions. In reality passengers' perception of different cost attributes may vary with time. This paper focuses on quantifying the temporal variation of passenger route choice behavior and its impact on overall passenger flow assignment. In order to efficiently estimate model parameters, we adapt a previous model to a missing data problem by introducing latent variables on route choice outcomes for each travel time observation. The revised model can be estimated using expectation-maximization (EM) algorithm. We apply the proposed framework on Singapore's metro system and temporal grouped smart card transactions. We find that route choice coefficients vary substantially with time. The relative value of transfer time in terms of in-vehicle time ranges from 2 to 3, being higher at off-peak hours than during morning/evening peaks. The result suggests that passenger care more about total travel time during peak hours, whereas comfort (e.g., less transfer time) is of more concern during off-peaks. The proposed framework is general and can be applied on other networks.**

Keywords*—*Data-driven, Metro network, Smart card data, Route choice, Flow assignment.

I. Introduction

The demand and range of urban mobility have been increasingly dependent on urban transit systems worldwide. As a transit system with dedicated infrastructure, metro system has larger capacity and higher speed than other transit modes, playing important roles in carrying urban transportation demand in large cities. In order to better meet the increasing demand, governments and transportation agencies are devoting more resource to constructing and extending metro networks. As a result, metro networks in large cities—such as Beijing, London and Tokyo—usually contain multiple services with intersections as transfer stations, and in these networks passengers may have more than one alternative for traveling from origin to destination. Despite that such dynamics could be captured using a discrete choice model, passengers' perception on different factors/attributes also varies with time in reality. However, such temporal variation is seldom addressed in the literature [1]. With regard to the increasing complexity of metro networks, understanding passenger flow assignment patterns in a complex metro network and its temporal variation become crucial to the design and operation of urban metro systems. An accurate flow assignment model not only helps operators to maintain service reliability, but also assist transit agencies in developing efficient failure response strategies [2, 3].

The deployment of automated fare collection (AFC) systems in public transit provides us with a new opportunity to study the operation of metro systems by analyzing smart card data [4]. The high spatial and temporal resolution of this data enables us to investigate both passenger behavior and service operation characteristic to a greater extent. Taking advantage of the wealth of smart card transactions, in this paper we focus on characterizing the temporal variation of network attributes and passenger behavioral parameters without using additional information from operators and agencies. In doing so, we improve a former model in [2] by introducing latent variables on route choices, which enables us to apply the expectation-maximization (EM) algorithm to efficiently estimate model parameters. This proposed framework is able to conduct diagnoses on train operation and learn flow assignment patterns simultaneously, without using other inputs at operational levels (e.g, train timetable/schedule and historical operation logs).

The remainder of this paper is organized as follows: Section II gives a brief review of related works. We describe the modeling framework in Section III and present the EM algorithm based solution approach in Section IV. In Section V, we apply the proposed framework and solution algorithm on Singapore's Mass Rapid Transit (MRT) network and travel time observations in different time periods as a case study. We also discuss the temporal variation of parameter inference and its implication. Finally, Section VI concludes our study and summarizes the main findings.

II. Related Works

Most metro systems are closed environments, which only register passenger's tapping-in and tapping-out activities at fare gantries [5]. As a result, we have limited knowledge about passenger behavior in between these two transactions. One of the most important information being missed is passenger route choice. This is particular important when the metro system of interest has a large and complex network, in which passengers may choose different alternatives, resulting in different flow assignment patterns. Therefore, studying passenger flow assignment in a metro network is crucial to maintaining service reliability, identifying critical transfer stations/facilities and developing efficient failure response strategies.

Research on passenger route choice behavior has long been relying on field surveys that collect passenger preference data (e.g., stated/revealed preferences) [1]. Recently, the field has observed an increasing number of studies that utilize the emerging smart card data [4]. Although the smart card system is introduced for the purpose of fare collection, it also provides us with large quantities of data registering passengers' tapping-in and tapping-out activities with both spatial and temporal information. Essentially, a full metro trip transaction can be represented by a tuple with five elements — (card id, origin station, destination station, tapping-in time, tapping-out time). The interval between tapping-in and tapping-out is usually referred to as "travel time".

Given its high spatial-temporal resolution, the data also allows researchers to develop data-driven methods to optimize service operation and to study passenger behavior. For example, Sun et al. [6] used highly-resoluted spatial-temporal demand to design demand-sensitive timetables. Using the same data, Jin et al. [7] aimed at integrating existing bus service with metro systems in order enhance the resilience of metro systems to disruptions. In terms of inferring individual behavior, Kusakabe et al. [8] developed a model to infer the exact train that a passenger took during his/her journey using historical smart card transactions. Zhou and Xu [9] proposed a probabilistic model to infer passenger route choice given his/her entry and exit times. Zhu et al. [10] applied a genetic algorithm based framework to calibrate passenger flow assignment models. In a recent paper, Sun et al. [2] proposed an integrated model to infer network attributes and passenger choice parameters simultaneously. This model takes passenger travel time as observations and characterizes travel time as mixture distribution from all potential routes. The authors applied a variable-at-a-time Metropolis sampling algorithm, which belongs to general Markov chain Monte Carlo (MCMC) algorithm, to infer the posterior distribution of unknown parameters. The estimation of route choice parameter is consistent with previous survey-based studies, suggesting that the value of transfer time is about twice the value of in-vehicle time.

The research presented in this paper is an extension of [2]. We modify the previous model to a missing value problem by introducing latent variables on passenger route choices, which is similar to the work of Zhan and Ukkusuri [11] about link travel time estimation on road networks. The new model can be efficiently estimated by applying the EM algorithm and thus facilitate the estimation of temporal variation.

III. Modeling Framework

In this section, we describe a probabilistic model on metro travel time, which has been used in [2]. The model takes travel time transactions from each OD pair as observations and network attributes and passenger behavioral coefficients as parameters. We next introduce the notations used in this paper.

A. Notation

1) Sets and indices:

- $W = \{w_i : i = 1, \ldots, W^n\}$ be the set of OD pairs and W^n is the total number of observed OD pairs from data.
- $T_i = \left\{t_i^j : j = 1, \ldots, T_i^n\right\}$ is the set of travel time observations on OD pair w_i indexed by j. T_i^n is the total number of travel time observations on OD pair w_i.
- $\boldsymbol{T} = \{T_i : i = 1, \ldots, W^n\}$ represents all travel time observations in the metro system.
- $R_i = \left\{r_i^k : k = 1, \ldots, R_i^n\right\}$ is the set of paths connecting OD pair w_i. R_i^n is the total number of considered paths for w_i. Each element r_i^k represents a set of used links.

2) Parameters:

- $\boldsymbol{c} = \{c_a : a \in U \cup V\}$ characterizes network link cost. U and V represent the sets of in-vehicle links and transfer links, respectively.
- $\boldsymbol{\alpha} = \{\alpha_u, \alpha_v\}$ characterizes link cost variation (coefficients of variation for in-vehicle links and transfer links, respectively).
- $\boldsymbol{\theta}$ characterizes passenger route choice behavior (parameters in utility function).
- m describes extra cost on waiting, access walking, egress walking and failed boarding.

We also denote $\boldsymbol{\Theta} = \{\boldsymbol{c}, \boldsymbol{\alpha}, \boldsymbol{\theta}, m\}$ as the set of all unknown parameters. We next introduce model assumptions.

B. Assumptions

- We assume costs (time) of in-vehicle links follow independent normal distributions with a constant coefficient of variation ($x_u \sim \mathcal{N}(c_u, \alpha_u c_u)$ for each in-vehicle link $u \in U$).
- We assume costs (time) of transfer links follow independent normal distributions with a constant coefficient of variation ($x_v \sim \mathcal{N}(c_v, \alpha_v c_v)$ for each transfer link $v \in V$).
- We assume all extra costs y for all OD pairs, including waiting time at boarding station, access/egress walking time and additional waiting caused by failed boarding, is characterized by a universal normal distribution $y \sim \mathcal{N}\left(m, \sigma_y^2\right)$. In the estimation we assume σ_y is a constant that is known in advance.
- We assume route choice behavior is characterized by an Multinomial Logit (MNL) model. Similar to [2], we assume the representative utility of a path r_i^k is assumed to be a linear combination of different route attributes. In general, we have $V_i^k = f(\boldsymbol{\Theta})$.

The link cost assumptions suggest that services are not punctual to exact timetables owing to various disturbance. Therefore, stochastic travel times will be observed in this model as in reality.

C. Metro travel time model

Based on previous assumptions, in this subsection we derive the model of travel time for individual passengers.

Following the MNL route choice assumption, the probability of choosing choosing alternative r_i^k on OD pair w_i is

$$\pi_i^k(\boldsymbol{\Theta}) = \frac{\exp\left(V_i^k(\boldsymbol{\Theta})\right)}{\sum_{k'=1}^{R_i^n} \exp\left(V_i^{k'}(\boldsymbol{\Theta})\right)}, \tag{1}$$

where $V_i^k(\boldsymbol{\Theta})$ is utility of alternative path r_i^k.

Applying the law of total probability on each OD pair, we can write the likelihood of $\boldsymbol{\Theta}$ given observing all smart card travel time transactions T as:

$$\mathcal{L}(\boldsymbol{\Theta}|\boldsymbol{T}) = \prod_{i=1}^{W^n} \prod_{j=1}^{T_i^n} \sum_{k=1}^{R_i^n} \pi_i^k(\boldsymbol{\Theta})\, h\left(t_i^j | r_i^k, \boldsymbol{\Theta}\right), \tag{2}$$

where $h(t|r,\boldsymbol{\Theta})$ is the probability of observing travel time t on alternative path r.

Note that $h(t|r,\boldsymbol{\Theta})$ is also a normal pdf given the additivity property of independent normal random variables

$$t|r,\boldsymbol{\Theta} \sim \mathcal{N}\left(\sum_{a\in r} c_a + m, \alpha_u^2 \sum_{a\in u_r} c_a^2 + \alpha_v^2 \sum_{a\in v_r} c_a^2 + \sigma_y^2\right), \tag{3}$$

where the links on path r are divided into two groups $r = u_r \cup v_r$, in which u_r represents the set of in-vehicle links and v_r is the set of transfer links.

IV. Solution Algorithm

The final log-likelihood of Eq. (2) contains $\log\left[\sum f(x)\right]$ terms, which is difficult to simplify and optimize. A previous article [2] applied a Markov chain Monte Carlo (MCMC) algorithm to sample the unknown parameters based on the likelihood function Eq. (2) in a Bayesian setting. The MCMC method provides good estimates on the unknown parameters. However, in doing so the MCMC method has to extensively compute the log-likelihood for all proposed candidates. In the variable-at-a-time sampling framework, the number of log-likelihood function calls in each iteration is twice the number of parameters. Because of the lack of simplification for the log-likelihood function, a single iteration to update all parameters may take about 15 sec. In practice we need to run MCMC for thousands of iterations to reach the stationary state and thus the process becomes very computationally intensive.

In order to simplify the estimation of unknown parameters, a natural alternative is to adopt the EM algorithm by introducing a set of latent variables describing route choice outcomes for each travel time observation $\boldsymbol{z} = \{\boldsymbol{z}_i : i = 1, \ldots, W^n\}$. In this set, each element $\boldsymbol{z}_i = \left\{z_i^j : j = 1, \ldots, T_i^n, z_i^j \in \{1, \ldots, R_i^n\}\right\}$ represents the choice outcome of travel time t_i^j. In this sense, $z_i^j = k$ means that travel time t_i^j is the outcome of choosing alternative path r_i^k among all potential choices. The EM algorithms is useful in the situation where the maximum likelihood is easier to compute when data were fully observed, which means latent variables are known. By introducing latent variables z_i^j, the complete likelihood can be expressed as

$$\mathcal{L}(\boldsymbol{\Theta}|\boldsymbol{T},\boldsymbol{z}) = \prod_{i=1}^{W^n} \prod_{j=1}^{T_i^n} p\left(t_i^j, z_i^j | \boldsymbol{\Theta}\right), \tag{4}$$

where the joint probability $p\left(t_i^j, z_i^j = k|\boldsymbol{\Theta}\right)$ can be given as

$$\begin{aligned} p\left(t_i^j, z_i^j = k|\boldsymbol{\Theta}\right) &= p\left(t_i^j | z_i^j = k, \boldsymbol{\Theta}\right) p\left(z_i^j = k|\boldsymbol{\Theta}\right) \\ &= \pi_i^k(\boldsymbol{\Theta})\, h\left(t_i^j | r_i^k, \boldsymbol{\Theta}\right). \end{aligned} \tag{5}$$

Eq. (5) comes from the fact that $p\left(t_i^j | z_i^j = k, \boldsymbol{\Theta}\right) = h\left(t_i^j | r_i^k, \boldsymbol{\Theta}\right)$ and the route choice probability is characterized by the MNL model $p\left(z_i^j = k|\boldsymbol{\Theta}\right) = \pi_i^k(\boldsymbol{\Theta})$. Therefore, the generalized joint probability can be formulated as pure production

$$p\left(t_i^j, z_i^j | \boldsymbol{\Theta}\right) = \prod_{k=1}^{R_i^n} \left[\pi_i^k(\boldsymbol{\Theta})\, h\left(t_i^j | r_i^k, \boldsymbol{\Theta}\right)\right]^{\mathbb{I}\left(z_i^j = k\right)}, \tag{6}$$

where $\mathbb{I}(e)$ is an indicator function equals to 1 if e is true and 0 otherwise.

Therefore, by integrating Eq. (6) into Eq. (4), the complete data likelihood becomes

$$\mathcal{L}(\boldsymbol{\Theta}|\boldsymbol{T},\boldsymbol{z}) = \prod_{i=1}^{W^n} \prod_{j=1}^{T_i^n} \prod_{k=1}^{R_i^n} \left[\pi_i^k(\boldsymbol{\Theta})\, h\left(t_i^j | r_i^k, \boldsymbol{\Theta}\right)\right]^{\mathbb{I}\left(z_i^j = k\right)}. \tag{7}$$

Applying the EM algorithm starts with random guesses of all known parameters $\boldsymbol{\Theta}^{(0)}$. Next, the EM algorithm tries to iteratively infer the latent variables given current estimates of parameters (E-step), and then update new parameter estimates by maximizing the expectation of complete data log-likelihood, which is often denoted as an auxiliary function $Q\left(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t-1)}\right)$ (M-step). The expectation is taken with repect to the previous parameter estimates $\boldsymbol{\Theta}^{(t-1)}$ and the observed data $\boldsymbol{T}$.

$$\begin{aligned} &Q\left(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t-1)}\right) = \mathbb{E}_{z|\boldsymbol{T},\boldsymbol{\Theta}^{(t-1)}}\left[\log \mathcal{L}(\boldsymbol{\Theta}|\boldsymbol{T},\boldsymbol{z})\right] \\ &= \sum_{i=1}^{W^n} \sum_{j=1}^{T_i^n} \sum_{k=1}^{R_i^n} \mathbb{E}_{z|\boldsymbol{T},\boldsymbol{\Theta}^{(t-1)}}\left[\mathbb{I}\left(z_i^j = k\right)\right] \log\left[\pi_i^k(\boldsymbol{\Theta})\, h\left(t_i^j | r_i^k, \boldsymbol{\Theta}\right)\right] \\ &= \sum_{i=1}^{W^n} \sum_{j=1}^{T_i^n} \sum_{k=1}^{R_i^n} p\left(z_i^j = k | t_i^j, \boldsymbol{\Theta}^{(t-1)}\right) \log\left[\pi_i^k(\boldsymbol{\Theta})\, h\left(t_i^j | r_i^k, \boldsymbol{\Theta}\right)\right] \\ &= \sum_{i=1}^{W^n} \sum_{j=1}^{T_i^n} \sum_{k=1}^{R_i^n} \nu_{i,j}^k \left[\log \pi_i^k(\boldsymbol{\Theta}) + \log h\left(t_i^j | r_i^k, \boldsymbol{\Theta}\right)\right], \end{aligned} \tag{8}$$

where $\nu_{i,j}^k \triangleq p\left(z_i^j = k | t_i^j, \boldsymbol{\Theta}^{(t-1)}\right)$ is the responsibility that path r_i^k takes for travel time t_i^j, which is computed in the E-step. The EM steps for our problem are summarized as below.

1. E-step

Given travel time $\boldsymbol{T}$ and $\boldsymbol{\Theta}^{(t-1)}$, compute $Q\left(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t-1)}\right)$. In computing Q, the responsibility is calculated using Bayes' theorem:

$$\nu_{i,j}^k = \frac{\pi_i^k\left(\boldsymbol{\Theta}^{(t-1)}\right) h\left(t_i^j | r_i^k, \boldsymbol{\Theta}^{(t-1)}\right)}{\sum_{k\prime=1}^{R_i^n} \pi_w^{k'}\left(\boldsymbol{\Theta}^{(t-1)}\right) h\left(t_i^j | R_i^{k'}, \boldsymbol{\Theta}^{(t-1)}\right)}. \tag{9}$$

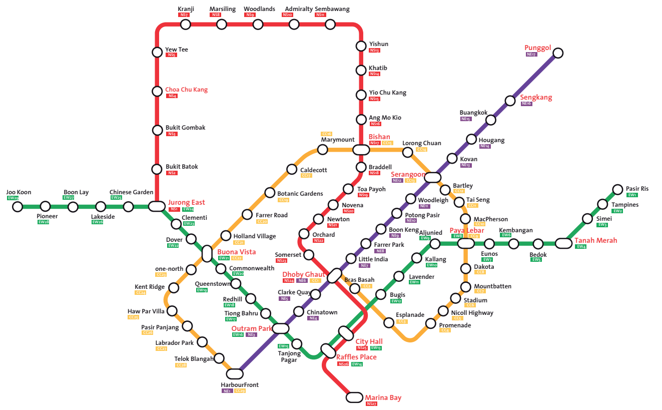

Fig. 1. Adapted MRT network of Singapore used in this study (source: http://exploresg.com/mrt/)

2. M-step

Update parameter estimates by maximizing $Q\left(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t-1)}\right)$ with respect to $\boldsymbol{\Theta}$. Essentially, the M-step in normal EM procedure also involves the update of weight parameter π_i^k. However, the number of parameters will be largely increased if we take π_i^k as free parameters. To simplify the model, we assume as mentioned before that π_i^k is solely characterized by the MNL model in Eq. (1), which is controlled by $\boldsymbol{\theta}$. Therefore, in the revised M-step we update π_i^k using Eq. (1) and concentrate on the estimation of other parameters $\boldsymbol{\Theta}$. In doing so, we set

$$\boldsymbol{\Theta}^{(t)} = \arg\max_{\boldsymbol{\Theta}} Q\left(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t-1)}\right), \tag{10}$$

and $t \leftarrow t+1$.

The EM algorithm should be run for multiple iterations unless a stopping criterion, such as convergence, has been met. One of the appeals of the EM algorithm is that the maximization of Q is often simpler than working on the incomplete likelihood in Eq. (2). We next integrate $h\left(t|r,\boldsymbol{\Theta}\right)$ and $\pi_i^k\left(\boldsymbol{\Theta}\right)$ into the auxiliary function.

Still, we do not have analytical solutions for the maximization problem in Eq. (10). However, given Q is twice continuous differentiable, the problem can be considered an unconstrained non-linear optimization problem. For the numerical experiment in the next section, we develop code in Matlab, in which BFGS algorithm is implemented, to solve the optimization problem in the M-step.

V. Case Study

A. Smart card data and test network

In this section we apply the proposed modeling framework on Singapore's Mass Rapid Transit (MRT) network. The same network and travel time observations as in [2] are used in this paper. Fig. 1 depicts the reconstructed metro network. The network contains 99 nodes, 95 in-vehicle links and 12 transfer links. Travel time observations are extracted from smart card transactions on 19th, March, 2012.

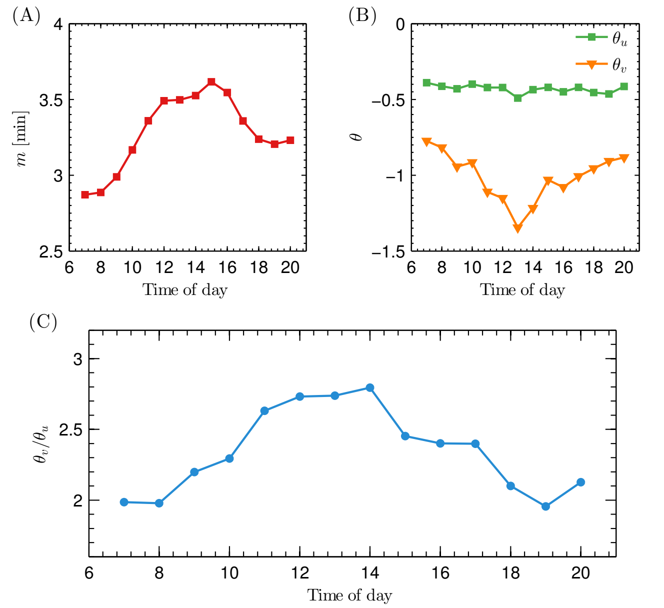

Fig. 2. Temporal variation of (A) extra cost m, (B) route choice parameters $\boldsymbol{\theta}$ and (C) relative ratio θ_v/θ_u.

B. Route choice behavior

We apply a simple MNL model to characterize passenger route choice behavior. We assume that the representative utility is solely determined by two attributes: (1) mean in-vehicle time $\sum_{a\in u_r} c_a$ and (2) mean transfer time $\sum_{a\in v_r} c_a$:

$$V_i^k\left(\boldsymbol{\Theta}\right) = \theta_u \sum_{a\in u_r} c_a + \theta_v \sum_{a\in v_r} c_a; \;\; r = r_i^k. \tag{11}$$

Here, θ_u and θ_v are scalar parameters modeling passengers' perception on in-vehicle time and transfer time, respectively. The size of final parameter set $\boldsymbol{\Theta}$ is 112, with $|\boldsymbol{c}| = 107$, $|\boldsymbol{\alpha}| = 2$, $|\boldsymbol{\theta}| = 2$ and $|m| = 1$.

C. Temporal variations

We divided all travel time observations into hourly groups given their starting time. For example, a trip initiated (with tapping-in time) between 7:30 a.m. and 8:30 a.m. is assigned to the group of 8 a.m. By doing so, we can infer the temporal variation of network attributes and passenger choice behavior by applying the estimation framework on each group of data. In total, we created 14 sets of travel time observations, ranging from 7 a.m. to 8 p.m. The estimation results are provided in the following subsection.

D. Results and analyses

The computational experiments are conducted on a PC with an Intel Core i7 3.40GHz CPU and 16 GB RAM. To avoid bias in estimation, we discarded those OD pairs with less than 25 travel time transactions. For computation efficiency, we keep only 100 travel time observations for those OD pairs with large numbers of observations ($n > 100$). In doing so, we generated a subset with 100 elements by sampling without replacement ($\max\left\{T_i^n\right\} = 100$). The set of alternative routes are generated using the same methods as in [2], by enumerating all potential routes using Breadth First Search (BFS) algorithm while only preserving those rational ones using pre-defined selection criteria. With regard to the extra cost y, we assume its variance is known in advance with $\sigma_y^2 = 1.5$. In the M-step, we restrict the maximum number of BFGS iterations to 10. We applied the proposed EM algorithm on each group of data. The EM algorithm shows good convergence properties in the numerical experiments. We restricted maximum number of EM iterations as 50. The estimation takes about 75 to 120 min for each data group, depending on the size of $\boldsymbol{T}$. Since $\boldsymbol{\alpha}$ appears in the auxiliary function as squared terms, in each step we update α as its absolute value.

Fig. 2(A) shows the temporal variation of extra cost m. Essentially, the result is consistent with service schedules—service frequency in morning and evening peaks [7-9 a.m and 6-8 p.m.] are higher than off-peaks. And thus passenger waiting time during off-peak hours is expected to be higher than during peak hours. Fig. 2(B) shows the variation of coefficients θ_u and θ_v, respectively. As can be seen, passengers always value transfer time more than in-vehicle time. In terms of temporal variation, θ_u does not varies substantially, while we observe a clear peak of θ_v at 1 p.m. To further quantify the relative value of transfer time against in-vehicle time, we depict the variation of θ_v/θ_u in Fig. 2(C). We found that θ_v/θ_u exhibits a similar pattern as m, being higher during off-peak hours than morning/evening peaks. The penalty of transfer time on utility is about twice of that of in-vehicle time during morning/evening peaks. However, during off-peak hours [11 a.m. to 2 p.m.] the relative ratio almost reaches 3, suggesting that passengers care more on transfer time during off-peak than peak hours. Overall speaking, the relative variation indicates that passengers show more concern about total travel time during peak time, whereas they care more about comfort (less transfer time) during off-peaks. On the other hand, the temporal variation could also result from other service factors, such as level of crowdedness and availability of seats, which are not modeled in the utility function.

The temporal variation of route choice coefficients $\boldsymbol{\theta}$ and other parameters also impacts route choice probability π_i^k and passenger flow assignment patterns. By using the estimated parameters we computed path probability for each temporal group of data using Eq. (1). We selected four OD pairs in which there is no dominated route. In Fig. 3 we show temporal variation of π_i^k for four selected OD pairs from 7 a.m. to 8 p.m. For simplicity, here we show only those routes with $\pi_i^k > 0.01$. In each case, a path and its temporal probability are drawn in the same color. For these selected OD pairs, we observed clear temporal heterogeneity of π_i^k due to the temporal change of model parameters, suggesting the temporal effects do play an important role in characterizing passenger route choice behavior and the final flow assignment patterns.

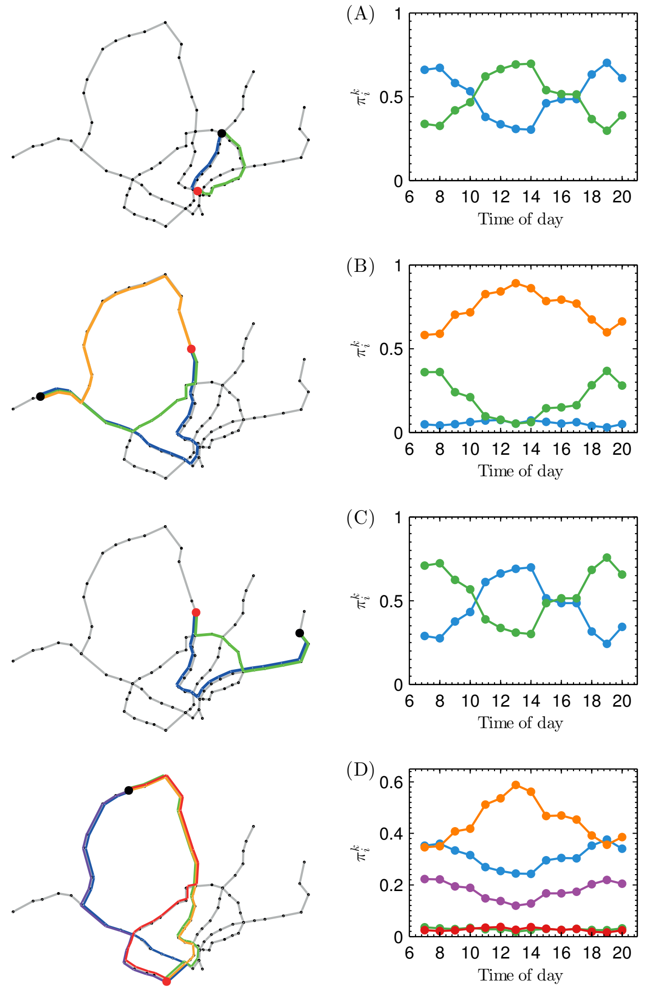

Fig. 3. Temporal variation of π_i^k for four OD pairs (only those routes with $\pi_i^k > 0.01$ are shown in this figure): (A) STN Bras Basah—STN Serangoon, (B) STN Boon Lay—STN Yio Chu Kang, (C) STN Ang Mo Kio—STN Tampines, and (D) STN Woodlands—STN Harbour Front. Each route is shown in the same color on the map and the corresponding plot.

VI. Summary and Conclusion

In this paper, we presented a framework to study passenger flow assignment problem in temporal scales using smart card transactions. In doing so, we divided all passenger travel time observations into time-stamped slices. In order to efficiently infer the temporal variation of network attributes and passenger choice behavior, we modified the MCMC based estimation model in Sun et al. [2] by introducing latent variables on route choice outcomes. The adapted model can be efficiently estimated using EM algorithm.

Previous studies on route choice in metro systems mainly rely on stated preference data. However, conducting such a study requires large quantities of choice observations collected from field surveys [1]. With the wide deployment of AFC systems in public transit, we are now provided with a wealth of smart card transactions with detailed spatial-temporal information. In this sense, inferring passenger route choices by applying a statistical learning model based on full travel time observations could be more advantageous than using limited survey data. This paper focuses on building a general learning framework that can be applied on other metro networks with passenger travel time information available (e.g., extracted from smart card transactions). Considering the lack of real choice data, the empirical validation of the proposed framework still requires more effort in obtaining real route choice outcomes, such as conducting large-scale surveys or tracking passenger activities using video recognition techniques.

Nevertheless, the results of this study have demonstrated good capacity of the proposed framework, in conducting passenger flow assignment in temporal scales. Applying the assignment model could further help us estimate train load profiles and level of crowdedness, and infer transfer demand at interchanging stations. These results could be used to identify critical links, platforms and transfer facilities, helping operators and agencies to better design and evaluate metro systems.

Supplementary Information

The Matlab source code and test data sets can be downloaded at https://github.com/lijunsun/MetroAssignment.

Acknowledgment

This research was partially supported by the Singapore National Research Foundation and ETH Zurich as a project at the Singapore-ETH Centre for Global Environmental Sustainability under the programme of the Future Cities Laboratory (NRF FI No. 370071002), Shanghai Philosophy and Social Sciences Foundation (Project No. 2014EGL008), and National Natural Science Foundation of China (Project No. 71401101).

References

[1] S. Raveau, Z. Guo, J. C. Muoñz, 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.

[2] 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.

[3] J. G. Jin, K. M. Teo, and L. Sun, "Disruption response planning for an urban mass rapid transit network," in *Transportation Research Board (TRB) 92nd Annual Meeting*. Washington DC, USA: Transportation Research Board, 2013, Conference Proceedings.

[4] 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, pp. 557–568, 2011.

[5] L. Sun, D.-H. Lee, A. Erath, and X. Huang, "Using smart card data to extract passenger's spatio-temporal density and train's trajectory of MRT system," in *ACM SIGKDD International Workshop on Urban Computing*. Beijing, China: ACM, 2012, pp. 142–148.

[6] L. Sun, J. G. Jin, D.-H. Lee, K. W. Axhausen, and A. Erath, "Demand-driven timetable design for metro services," *Transportation Research Part C: Emerging Technologies*, vol. 46, pp. 284–299, 2014.

[7] J. G. Jin, L. C. Tang, L. Sun, and D.-H. Lee, "Enhancing metro network resilience via localized integration with bus services," *Transportation Research Part E: Logistics and Transportation Review*, vol. 63, pp. 17–30, 2014.

[8] T. Kusakabe, T. Iryo, and Y. Asakura, "Estimation method for railway passengers' train choice behavior with smart card transaction data," *Transportation*, vol. 37, no. 5, pp. 731–749, 2010.

[9] F. Zhou and R. Xu, "Passenger flow assignment model for urban rail transit based on entry and exit time constraints," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2284, pp. 57–61, 2012.

[10] W. Zhu, H. Hu, and Z. Huang, "Calibrating rail transit assignment models with genetic algorithm and automated fare collection data," *Computer-Aided Civil and Infrastructure Engineering*, vol. 29, no. 7, pp. 518–530, 2014.

[11] X. Zhan and S. V. Ukkusuri, "Probabilistic urban link travel time estimation model using large-scale taxi trip data," in *Transportation Research Board (TRB) 94th Annual Meeting*, Washington DC, USA, 2015.