Acknowledgements This work was, in part, supported by the Transportation Informatics Tier I University Transportation Center and the National Science Foundation. We would like to thank Dr. Younshik Chung at Yeungnam University in Korea for his involvement in this project.

Contents

1 Introduction and Background

Public transit planning is a complex design problem that involves a wide range of research topics and methodologies, such as strategic planning (e.g., network planning, customer behavior analysis, and demand forecasting), tactical planning (e.g., schedule adjustment in response to longitudinal and individual trip pattern changes), and operational planning (e.g., adjustment in response to demand fluctuations) (Pelletier et al., 2011). One key input into these problems is the Origin-Destination (OD) demand; the quality of this OD demand information is thus critical for informing policies.

The assessment of the routing preferences of transit system users is another component important for planning and assessing the efficiency of system operations. Such model-based assessments, with outputs in the form of (dis)utility coefficients for time, mode, fare, etc., have been carried out for decades (e.g., Wardman (2004)). Traditionally, the data required to inform such modeling efforts were collected through surveys, but the recently operationalized Automated Fare Collection systems can provide such data sets at almost no added cost (Sun et al., 2015; Sun and Xu, 2012; Zhou and Xu, 2012; Zhu et al., 2014).

The availability of the OD demand and traveler trajectory data is crucial for the effective and efficient operation and management of a transit system, yet at the same time, most difficult to secure. Much effort has been invested into helping analysts obtain OD demand data, both directly and indirectly. Various technologies and data collection methods serve this purpose, including manual onboard recording, surveys, road-side monitoring, automated passenger counting (APC) and automated vehicle location (AVL) tracking (e.g., that exploiting global positioning system (GPS) technology), as well as mobile phone tracking, and social media data analysis. Each of these, however, suffers from its own deficiencies: typically, the obtained data are expensive, hard-to-access, and/or noisy.

Automated Fare Collection (AFC) systems, often called smart transit card systems, come as a potent new data source that can be used to obtain accessible OD demand and user preference data. It has become more popular and have found use in public transportation systems worldwide in recent years (Pelletier et al., 2011). While the primary function of AFC systems is fare collection and user class validation for different fare rates, these systems also store travel information of all users, recorded as transactions. These data provide detailed travel information of transit users that can potentially be informative for transit operators and planners.

The information most commonly extracted from AFC data is station-to-station Origin-Destination (OD) travel demand. Travelers check-in while entering a transit system, and in certain applications, check-out while leaving the system. Thus, AFC systems record high-quality inbound and outbound station data that can be used to inform transit planning decisions. To this end, AFC records are processed to estimate OD demand with a greater accuracy and higher detail (Barry et al., 2002b; Chan, 2007; Cui, 2006a; Gordillo, 2006a; Hazelton, 2008; Lianfu et al., 2007; Munizaga and Palma, 2012; Pelletier et al., 2011; Trepanier et al., 2007; Zhao et al., 2007). An extracted OD demand matrix can then be used to allow operators and planners to better respond to the system's needs at an individual traveler level.

Previous studies developed algorithms to distill the OD information from AFC data. Their research objectives and methodologies varied, based on the data availability. Given AFC data with entry-only information, the authors have typically focused on inferring the destination stations by using rule-based approaches (Barry et al., 2002a; Nassir et al., 2011; Trépanier et al., 2007; Zhao, 2004). Barry et al. (2002a) came up with a rule-based model to synthesize AFC data to infer alighting stations. Zhao (2004) inferred alighting stations using modified rule-based model for Chicago AFC data; the destination estimation process here relied also on other data sources, in particular AVL, APC and GIS data. Trépanier et al. (2007) pointed out that individual trip

destinations could also be estimated by looking at similar trips made by the same card holder, found in the trip history database. Once a destination location is obtained (estimated), one can generate a passenger trip OD matrix (Cui, 2006b; Gordillo, 2006b).

Besides OD estimation, one can employ user-level information captured as AFC data to assess system-level transit network operation, more specifically, to distill users' travel patterns (Chakirov and Erath, 2011; Ma et al., 2013; Sun et al., 2012), perform route choice estimation analysis (Kusakabe et al., 2010; McMullan and Majumdar, 2012; Sun and Xu, 2012), trip purpose inference (Lee and Hickman, 2014), travel time analysis, and conduct an overall transit system reliability assessment (Sun et al., 2016; Sun and Xu, 2012).

The main issue with AFC data analyses is that such inference typically produces only *stop-level* OD estimates. The main deficiency of the *stop-level* OD estimation approaches is that they neglect the multi-modal nature of transit and lose the information of transit demand elasticity, which results in crude solutions for transit planning problems (Kumar et al., 2016). More broadly, AFC data analyses have been focused on understanding of statistical properties of data rather than that of the underlying fundamental behavior of users. It is, however, apparent that within-system inference captures just a part of a complete multi-modal route of a given traveler. Meanwhile, a route utilizing a transit system may start at the traveler's home and have him/her using bus, subway, or tram, etc., then completing the trip to a destination by foot. Indeed, check-out and check-in data at the stations of such a trip can be accessed as AFC records; however, the remainder of the complete OD-route, with its "true" OD locations and possible use of other modes of transportation as its parts, remains *hidden*. Also, stop-level ODs as well as mode choices for the same traveler can vary, depending on the traveler's time constraints and travel environment conditions.

This paper presents an approach for OD pair inference. Its logic and methodology relies on the observed travel trajectories, which reflect travelers' evaluation of travel times, travel prices, trip convenience, etc., depending on the state of the transit environment. The variability in the routing choices of a given observed traveler, in response to the changing transit environment parameters, can be gleaned from the AFC data. Using the traveler's trajectories under different travel environment settings, it is possible to identify their OD points from a set of reasonable OD candidates. These OD candidates are vital inputs to the problem of the inference of true OD pairs; they can be identified either by relying on side knowledge or via a rule-based selection process. This paper contributes to the transportation science both by pointing out ways to systematically select OD candidates and by inferring in a calculated way which of these candidates are the most likely ones, for all travelers.

The presented methodology also offers a way for quantifying general travel preferences (i.e., those characteristic of the entire population of travelers), which affect each traveler's route choice decisions. The literature that models and analyzes routing decisions has extensively used the mixed multinomial logit model for this purpose. In order to paramiterize this model, true OD pairs for system users are needed. In other words, the values of the inferred parameters can be rather sensitive to the availablity of only station-level OD or true OD information. Given the interdependence of the problems of (1) inferring travelers' true ODs and (2) inferring their route choice preferences from the same AFC data, we present an expectation maximization method to tackle both problems at once, in an iterative manner: intuitively, we fix one set of unknowns and refine our estimates of the other unknowns, and vice versa, until convergence is reached.

The contribution of this paper is thus two-fold. First, the paper presents a two step method to infer the true ODs of public transit users, as well as these users' preferences. Both of them are useful for planning and assessing the efficiency of a given public transit system. Second, the paper assesses the accuracy of the presented inference method in a real world case study, with a large AFC data set. The results reveal that most of the public transit user ODs can be correctly inferred. Moreover, the estimates of the user preferences are shown to turn out similar to what one would obtain by

using a traditional multinomial logit route choice model (the latter run under full information).

The remainder of this paper is structured as follows. Section 2 introduces the assumptions, notations and key ideas of our model. Section 3 provides computational results. The paper is concluded with Section 4, which summarizes the obtained insights and discusses some future research directions.

2 Modeling Framework

Several definitions and terms are needed to describe our model. *Travel history* of a traveler is a set of all the transactions that are registered when he/she adopts an AFC-equipped public transit system. A *transaction* is a record that contains the information about the traveler's ID (per the travel card they own), the boarding time and station, the alighting time and station, the fare paid, and perhaps, any other information the AFC system is set up to record. A sequence of transactions that a traveler generates while commuting from an origin to a destination is termed a trip. Note that a trip may or may not involve transfers between the bus and metro transportation modes. A *station-level OD* pair refers to the first boarding station and the last alighting station of a *trip*. Therefore, the trip OD is not really the "true" OD: the OD directly observable from AFC data misses the traveler's home address location and destination address location points. By using the term *true OD*, or simply, *OD*, this paper refers to the points beyond stations (of a station-level OD): these points typically are apartment buildings, workplaces, shopping places, gyms, etc. A *route* is a path from a true origin point to a true destination point; thus, each traveler has multiple routes to choose from when traveling *along* their true OD. A traveler's trip is only the observable part of his/her route, chosen on a particular occasion (day, or time of day). A vector of *route attributes* describes the convenience associated with traveling on a given route at a given time; the attribute values that this vector contains may include travel time, transfer time, crowdedness, price paid, and number of transfers, among others – they may depend on weather conditions, road closures and other factors. The numerical values of these attributes help a traveler determine the utility of each possible route on a particular occasion; naturally, a traveler is more likely to choose to travel along a route with a higher utility value than that with a lower utility value. *Trip history* of a traveler is a sequence of trips taken by him/her to travel along the true OD on different occasions. It is assumed that all the trips in the trip history of a traveler correspond to one and the same true OD (in practice, data post-processing and careful treatment of different destination points for each traveler is required to ensure that this assumption holds). Importantly, for a given true OD, the observed station-level ODs for the same traveler can be, and often are, different due to the fact that traveler tend to adjust their routes in response to the changing travel environment conditions.

A practical challenge that this paper addresses is that of inferring the true OD of each traveler, given the part of their trip history that constitutes the set of the trips that they took from the respective true origin to the respective true destination. It is thus firstly assumed that all such trips have been retrieved from the AFC data. Secondly, it is assumed that to infer a true OD means to select the correct one from a finite set of *eligible* (i.e., possible, or worth considering) ODs, among which the true OD is assumed to necessarily be present. Thirdly, it is assumed that every traveler makes decisions that generally aim to maximize their travel quality and convenience (represented as utility/disutility).

To express our modeling approach using mathematical notations, suppose there are N travelers in a public transit system, and their trip histories are on record and available. For each traveler $i = 1, 2, ..., N$, let his/her true OD be denoted by w_i. Note again that an individual may have multiple true ODs to make travels along, for different purposes: e.g., to travel regularly from home

to workplace, from workplace to a store, from the store to a gym, etc. To make the presentation of our model clear, it is worth emphasizing that hereafter, we treat the same individual taking trips for different purposes as different travelers. Hence, it is justified to assume that each traveler has only one true OD.

For any OD w, let a *set of route candidates* $R(w)$ contain all the *candidate* routes along this OD, i.e., the routes with high enough utility to consider taking them. Let $r_{i,k}$ denote the route that traveler $i = 1, 2, ..., N$, was observed to take along OD w_i on occasion $k = 1, 2, ..., K_i$. It is assumed that the choice of $r_{i,k} \in R(w_i)$ (i.e., among all the candidates) is consistent with the discrete choice model that assigns to it the probability $p_i^r(t)$, evaluated for the route $r \in R(w_i)$ at the time (t) when the traveler takes it. This probability is a function of the utility values of all the routes, $r \in R(w_i)$, considered as candidates at time t; the utility of any one such route is computed as the product of the fixed (but latent, i.e., originally unknown) preference vector $\boldsymbol{\beta}$ and this route's attributes $\boldsymbol{E}(t, r_{i,k})$ at time t. The vector $\boldsymbol{E}(t, r_{i,k})$ can contain such specific route attributes as travel time $T(t, r_{i,k})$, transfer time $W(t, r_{i,k})$, number of transfers $Z(r_{i,k})$, travel fare $F(t, r_{i,k})$, and crowdedness $C(t, r_{i,k})$, among others.

Under a multinomial logit route choice model, the notation for the probability that traveler i chooses route $r_{i,k} \in R(w_i)$ while traveling along OD w_i on occasion k at time t thus expands to $p_i^{r_{i,k}}(\boldsymbol{\beta}, \boldsymbol{E}(t, r_{i,k}), w_i)$.

The challenge that one has to tackle here is that $\boldsymbol{\beta}$, $\boldsymbol{E}(t, r_{i,k})$ and w_i, $i = 1, 2, ..., N$, are all unknown; we set out to build a model and devise an inference algorithm to estimate them simultaneously. As part of its output, our algorithm will return an inferred vector of real numbers $\hat{\boldsymbol{\beta}}$, and the inferred true ODs $\hat{w}_i$, selected from the sets of eligible true ODs $\boldsymbol{w_i}$, $i = 1, 2, ..., N$. Note that for more accurate inference, it is practical to restrict each set $\boldsymbol{w_i}$, $i = 1, 2, ..., N$, to include only some of the possible true ODs, so that this set would be small enough but would still contain the true OD in it. Thus, determining which eligible ODs should be included in each eligible set is an important task that requires care. $\boldsymbol{E}(t, r_{i,k})$ is the route attributes of a route $r_{i,k}$. In order to further clarify the object of our study, it should be pointed out that the routes $r_{i,k}$, $i = 1, 2, ..., N$, $k = 1, 2, ..., K_i$, are only partially observable: the observable part of route $r_{i,k}$ is denoted by $\widetilde{r}_{i,k}$, referred to as *trip*. All the trips are the inputs into our inference algorithm; for each traveler $i = 1, 2, ..., N$, the algorithm's output must clearly be one of the routes that have the recorded trip as their part and the inferred route should better have a high utility, among all the candidate routes. Therefore, $r_{i,k}$ will be inferred simultaneously with and in reference to the values $\boldsymbol{\beta}, w_i$ and $\widetilde{r}_{i,k}$, $i = 1, 2, ..., N$, $k = 1, 2, ..., K_i$. Specifically, for any eligible true OD w and recorded trip $\widetilde{r}_{i,k}$, the attributes of the part of the route between a true origin and a station-level origin, as well as the part of the route between a station-level destination and a true destination will be obtained. In other words, while the attributes $\boldsymbol{E}(t, \widetilde{r}_{i,k})$ are known and observable, the attributes $\boldsymbol{E}(t, r_{i,k})$ of the route $r_{i,k}$ are originally unknown and will be obtained during inference.

Given the observed travel history data, in order to estimate $\boldsymbol{\beta}$ and w_i, $i = 1, 2, ..., N$, we adopt a likelihood maximization approach. To this end, the product of all route choice probabilities will be maximized,

$$\max \prod_i \prod_k \prod_{w_i \in \boldsymbol{w}_i} p_i^{r_{i,k}}(\boldsymbol{\beta}, \boldsymbol{E}(t, r_{i,k}), w_i), \tag{1}$$

over $\boldsymbol{\beta}$ and w_i, $i = 1, 2, ..., N$.

The solution to the optimization problem (1) returns the best estimated true OD pairs and the best fitting numerical value of the utility weight vector $\boldsymbol{\beta}$, for each traveler $i = 1, 2, ..., N$, given their trip history $\widetilde{r}_{i,k}$. Section 3 describes the two stages one has to go through to solve (1), providing more details about the modeling and solution methodology.

3 Model and Formal Problem Statement

This section introduces the **OD Inference Problem (ODIP)**, which models route selection by travelers based on their ODs under variable travel environment conditions, and enables the inference of the ODs based on the observations of those routes. A solution to this problem returns the most likely OD for each traveler, along with a single utility weight vector that specifies the route utility calculation model, under the assumption that the traveler population is homogeneous in regards to utility valuation.

To describe ODIP, consider a traveler who travels between two often-visited location points (e.g., a home origin and a workplace destination) on weekdays using a public transit system: these two locations, together referred to as a true OD, are to be inferred, given the data of how the traveler has used the system. In response to the travel environment changes, the traveler may be found to have been changing their route accordingly to maximize their travel utility on each of the different travel occasions; it that's the case, then it opens the door for the true OD inference. The routes, the traveler would pick on different occasions, would be the ones with the higher utility values, i.e., the ones with lower travel times, fewer transfers, lower fare, lower crowdedness, etc. We assume that at the time of travel, the state of the travel environment is known to the traveler, and that he/she can make the route choice decisions accordingly. Given the above argument, and under the assumption of the traveler population homogeneity, the route choice behavior of all travelers can be modeled. In what follows, we specify our adopted model and present ODIP as the problem of inferring the travelers' true ODs, all at once for all trevelers, knowing each traveler's trip history and the travel environment state at the time that each trip was taken.

To build our model and formulate ODIP, the following sets, variables, parameters and functions are needed.

Table 1: Mathematical Notations used for the Modeling

β_{time}	Utility weight for travel time
β_{transfer}	Utility weight for transfer time
β_{num}	Utility weight for number of transfers
β_{fare}	Utility weight for travel fare cost
$\beta_{\text{crowdedness}}$	Utility weight for crowdedness
$\boldsymbol{\beta}$	Utility weight vector, with the elements β_{time}, β_{transfer}, β_{num}, β_{fare}, and $\beta_{\text{crowdedness}}$
$\hat{\boldsymbol{\beta}}$	Estimated utility weight vector
$T(t,r)$	Travel time for route r taken at time t
$W(t,r)$	Similarly, transfer time for route r at time t
$Z(r)$	Similarly, number of transfers for route r at time t
$F(t,r)$	Similarly, trip fare (cost) for route r at time t
$C(t,r)$	Similarly, crowdedness value for route r at time t
$\boldsymbol{E}(t,r)$	Route attribute vector for route r at time t, with elements $T(t,r)$, $W(t,r)$, $Z(r)$, $F(t,r)$, and $C(t,r)$
w	One element of a set of the eligible true OD pairs for traveler i ($w \in \boldsymbol{w_i}$)
w_i	True OD pair for traveler i ($w_i \in \boldsymbol{w_i}$)
$\hat{w}_i$	Estimated true OD pair for traveler i ($\hat{w}_i \in \boldsymbol{w_i}$)
$R(w)$	Set of route candidates for OD pair w
s_r	Path size overlap correction factor (measures the length of the overlap between route r and all other route candidates)

As we mentioned in the above section, ODIP can be modeled as a likelihood maximization problem, with the likelihood expressed as the product of the route choice probabilities $p_i^{r_{i,k}}(\boldsymbol{\beta}, \boldsymbol{E}(t, r_{i,k}), w)$, over all chosen routes $r_{i,k}$ (connecting all eligible ODs $\boldsymbol{w}$) by all travelers $i = 1, 2, ..., N$ on all occasions $k = 1, 2, ..., K_i$, evaluated given the states of the travel environment $\boldsymbol{E}(t, r_{i,k})$ at the respective times the trips were taken. Now, we are going to model the route choice probabilities using the multinomial logit model. The multinomial logit model and its variations are widely used to calculate the probability of choosing a specific route, from a known set of all route candidates. Note that this model assumes that travelers have enough information to make this choice, but the preference of one route over another remains probabilistic, emphasized when the utility values of the routes do not differ much.

Following the prior developments in multi-modal transit route choices modeling (Bovy and Hoogendoorn-Lanser, 2005; Hensher and Greene, 2003), and based on the random utility theory (Ben-Akiva and Lerman, 1985; McFadden et al., 1973), a linear utility function prescribes to compute the utility that traveler i assigns to each route r as follows:

$$\begin{aligned} V(t,r) = &-\beta_{\text{time}} * T(t,r) - \beta_{\text{transfer}} * W(t,r) - \beta_{\text{num}} * Z(r) \\ &- \beta_{\text{fare}} * F(t,r) - \beta_{\text{crowdedness}} * C(t,r), \end{aligned} \tag{2}$$

where $V(t,r)$ is the utility of a route r at time t. t is the time of making route choice decision. $T(t,r), W(t,r), Z(r), F(t,r)$ and $C(t,r)$ stand for travel time, transfer time, number of transfers, travel cost and crowdedness for route r at time t. These variables are figuring in the overall utility with the corresponding utility weights: $\beta_{\text{time}}, \beta_{\text{transfer}}, \beta_{\text{num}}, \beta_{\text{fare}}$, and $\beta_{\text{crowdedness}}$. These weights

are mentioned as utility weight vector $\boldsymbol{\beta}$. To simplify the model, we assume the homogeneity existed of all transit users, therefore, only one set of $\boldsymbol{\beta} = [\beta_{\text{time}}, \beta_{\text{transfer}}, \beta_{\text{num}}, \beta_{\text{fare}}, \beta_{\text{crowdedness}}]$ will be estimated for all travelers.

Then a traveler's route choice can be modeled by the following multinomial logit function:

$$p_i^r(\boldsymbol{\beta}, \boldsymbol{E}(t,r), w_i) = \frac{e^{V(t,r)}}{\sum_{r' \in R(w_i)} e^{V(t,r')}}, \tag{3}$$

where r is the chosen route, r' is a route candidate in set $R(w_i)$, and $R(w_i)$ is a set of all route candidates. All of these routes are connecting OD w_i and are evaluated at time t.

Path size logit model Ben-Akiva and Bierlaire (2003) is a modification of traditional multinomial logit model. It corrects the route choice probability with considering the situation that routes may have sharing segments. More specifically, sharing segments among routes may reduce the probability of such routes being chosen. Path size logit model states that the probability of a traveler choose route r against all other route candidates is given as follows:

$$p_i^r(\boldsymbol{\beta}, \boldsymbol{E}(t,r), w_i) = \frac{s_r e^{V(t,r)}}{\sum_{r' \in R(w_i)} s_{r'} e^{V(t,r')}}, \tag{4}$$

where s_r is the path size overlap correction factor for route r. $s_{r'}$ is the path size overlap correction factor for route r'. Generally, for any route r, $s_r = \sum_{a \in \tau_r} \frac{L_a}{L_r} \frac{1}{N_a}, \forall r \in \boldsymbol{R}(w)$. L_a is the length of an overlapping link a of route $r \in \boldsymbol{R}(w)$ on other route candidates. All the overlapping links of route r are denoted by set τ_r. L_r is the length of route r for the OD pair w ; N_a is the number of routes in the route set $\boldsymbol{R}(w)$ that through link a.

Therefore, for a given OD w_i, if route choice $r_{i,k}$ is known for traveler i on occasion k, with route attributes $E(t, r_{i,k})$, the probability for choose $r_{i,k}$ is denoted as $p_i^{r_{i,k}}(\boldsymbol{\beta}, \boldsymbol{E}(t, r_{i,k}), w_i)$. If the all the route choices are known, then, it is reasonable to use maximum likelihood estimation to estimate the utility weight vector $\boldsymbol{\beta}$ and true ODs w_i. That means, the product of all probabilities $\prod_i \prod_k \prod_t \prod_w p_i^{r_{i,k}}(\boldsymbol{\beta}, \boldsymbol{E}(t, r_{i,k}), w)$ will be maximized to obtain the best estimator.

However, there are still two problems need to be solved before applying the model. The first problem is that for any route r along OD w at time t, only the part of route that is in the transit system is observable, denoted as $\widetilde{r}$. How to obtain the route r's route attributes $\boldsymbol{E}(t,r)$ with only partial information? Let us denote the part of route between true origin to the first boarding station (station-level origin) as r_{begin}, the part of route between the last alighting station (station-level destination) to the true destination as r_{end}. The part of route between first boarding station (station-level origin) to last alighting station (station-level destination) is $\widetilde{r}$. For a given true OD w, a possible route r consists of $r_{begin}, \widetilde{r}$ and r_{end}. Usually r_{begin} and r_{end} are finished by walking and they are short in distance. So it is reasonable to calculate route of r_{begin} and r_{end} by choosing the shortest path. Furthermore, route r's route attribute can be calculated by summing up the attributes of r_{begin}, $\widetilde{r}$ and r_{end} ($\boldsymbol{E}(t,r) = \boldsymbol{E}(t, r_{begin}) + \boldsymbol{E}(t, \widetilde{r}) + \boldsymbol{E}(t, r_{end})$). Then $\boldsymbol{E}(t,r)$ will be used as the input - known route choice history - for ODIP.

The second problem is: what will be the eligible true ODs $\boldsymbol{w}_i$ for traveler i? There are many ways to do it, such as distance-based method, survey-based method, etc. We provide a data-driven method and it will be discussed in details in section 4.

Then ODIP can be summarized as a maximization problem. The inferred utility weight vector $\hat{\boldsymbol{\beta}}$ and true OD $\hat{w}_i$ for traveler i is as follows: :

$$\{\boldsymbol{\beta}, \{\hat{w}_i\}_{i=1,2,\ldots,N}\} = \arg\max \prod_i \prod_k \prod_t \prod_w p_i^{r_{i,k}}(\boldsymbol{\beta}, \boldsymbol{E}(t, r_{i,k}), w). \tag{5}$$

where $i = 1, 2, ..., N$ are travelers, w is an eligible true OD, $r_{i,k}$ is a chosen route along this true OD w by i on occasion k, t is the time when route choice is made, $\boldsymbol{\beta}$ is the weight utility vector, $\boldsymbol{E}(t, r_{i,k})$ is the route attributes along a route $r_{i,k}$, and $p_i^{r_{i,k}}(\boldsymbol{\beta}, \boldsymbol{E}(t, r_{i,k}), w)$ is the route choice probability of route $r_{i,k}$ along OD pair w under occasion k at time t for traveler i, and the utility weight vector is $\boldsymbol{\beta}$. Except for $\boldsymbol{\beta}$ and w, all other parameters are given. We will introduce our instance and data, and discuss the details of methodology in the following section.

4 Instance Specification and Methodology

4.1 Data Preparation

4.1.1 Data Preprocessing

AFC data records are not recorded in the format of trips but in transactions, therefore it cannot be directly used as input for ODIP model. We need to clean, preprocess and reorganize the data. Typically, one transaction record only has information such as boarding station/time, alighting station/time, price paid, etc. However, in a multi-modal travel environment, a complete station-level OD trip usually involved multiple transit modes, thus, multiple transaction records. To extract trip history of a traveler, such multiple transactions need to be combined for calculating station-level origin to destination route attributes such as travel time values, number of transfers, total priced paid, etc. In addition, to measure the crowdedness of a vehicle, number of individuals taking the same vehicle at the same time (crowdedness) is extracted by scanning the whole AFC dataset.

After the AFC data has been preprocessed, it is randomly divided into training and test dataset. The size of training and test dataset are 70% and 30% of the original data. Following eligible ODs are inferred based on training data. We'll use test data to validate the performance of our ODIP model.

4.1.2 Eligible OD Analysis

Specifying an instance of the ODIP, based on the historical data of AFC system operation, is a challenge in itself. In particular, eligible ODs have to be selected in such a way that ODIP could be properly solved. Eligible ODs are proper candidates, from which the true ODs can be inferred using our model. Not all of the origins and destinations in the transit system could be eligible ODs for a traveler. Eligible OD analysis allows us to learn true OD candidates from historical data and infer the most probable ones given traveler's same-OD trip history.

Eligible ODs are not readily available from data. They are some inferred true ODs that traveler likely to visit, that means, the probability of visiting such places is higher than others, given a traveler's trip history. This probability of visiting can be calculated in two steps: route choice frequency and bayesian inference.

Let us define observable route ($\widetilde{r}$) as the part of a complete route (r) that happened within the transit system. Observable route choice rate can be obtained from historical data - for example, surveys or simply prior knowledge. The rate of a route $\widetilde{r}_i$ being chosen given the true origin-destination pair is O_xD_y is showed as follows:

$$P(\widetilde{r}|O_xD_y) = \frac{Count(\widetilde{r}|O_xD_y)}{\sum_{\widetilde{r}' \in R(O_xD_y)} Count(\widetilde{r}'|O_xD_y)}. \tag{6}$$

More specifically, $O_x, x = 1, 2, ..., X$ are origins and $D_y, y = 1, 2, ..., Y$ are destinations. $Count(\widetilde{r}|O_xD_y)$ is the number of count of observable route $\widetilde{r}$ appears in the history data, given the OD pair of $\widetilde{r}$ is

O_xD_y. In order to avoid zero during calculation, laplace smoothing is used. It makes all $Count()$ start from at least 0.01. Now, the rate of $\widetilde{r}_i$ being taken given its OD is O_xD_y will be calculated using equation 7:

$$P(\widetilde{r}|O_xD_y) = \frac{Count(\widetilde{r}|O_xD_y) + 0.01}{\sum_{\widetilde{r}' \in R(O_xD_y)}(Count(\widetilde{r}'|O_xD_y) + 0.01)}. \tag{7}$$

Once we find out all observable route choice rate for all given ODs, the second step is to find out which ODs are the eligible ODs for a traveler i based on his/her trip history. Suppose his/her trip history contains $\widetilde{r}_{i,k}$ where $k = 1, 2, ..., K_i$ indicates the occasion of the trip. The probability of O_xD_y being this traveler's true OD can be calculated using Naive Bayes method as follows (equation 8):

$$\begin{aligned} P(O_xD_y|\widetilde{r}_{i,1}, \widetilde{r}_{i,2}, ..., \widetilde{r}_{i,k}) &= \frac{P(\widetilde{r}_{i,1}, \widetilde{r}_{i,2}, ..., \widetilde{r}_{i,k}|O_xD_y) * P(O_xD_y)}{\sum_{x,y} P(\widetilde{r}_{i,1}, \widetilde{r}_{i,2}, ..., \widetilde{r}_{i,k}|O_xD_y) * P(O_xD_y)} \\ &= \frac{P(\widetilde{r}_{i,1}|O_xD_y)P(\widetilde{r}_{i,2}|O_xD_y)...P(\widetilde{r}_{i,k}|O_xD_y)}{\sum_{x,y} P(\widetilde{r}_{i,1}|O_xD_y)P(\widetilde{r}_{i,2}|O_xD_y)...P(\widetilde{r}_{i,k}|O_xD_y)}, \end{aligned} \tag{8}$$

where $P(\widetilde{r}|O_xD_y)$ is calculated in equation 7. For each traveler i, any O_xD_y with corresponding $P(O_xD_y|\widetilde{r}_{i,1}, \widetilde{r}_{i,2}, ..., \widetilde{r}_{i,k}) > v$ will be the eligible ODs (denoted as w) for further inference. where v is a threshold value for picking out the eligible ODs. If this traveler only has one eligible OD left based on his/her trip history, this only OD pair will be his/her inferred true OD. Thus, further inference is not needed. For the travelers with more than one eligible OD pair after eligible OD analysis, we will use following Expectation Maximization method, to infer their most possible true OD among all eligible ODs.

4.2 Expectation Maximization

After eligible OD analysis, a set of eligible true OD $\boldsymbol{w}$ will be extracted for inferring traveler i's true OD based on his/her travel records. In this step, we developed an expectation maximization method to solve ODIP. Further, traveler i's true OD and his/her utility weight vector will be inferred.

Let $r_{i,k}$ denote the chosen route for a traveler i at time t with the route attributes are $\boldsymbol{E}(t, r_{i,k})$. Let w_i denote the true OD pair of this trip. We are planning to use Maximum Likelihood Estimation (MLE) algorithm to find out what is the best estimator for $\boldsymbol{\beta}$. If w_i is given for each traveler i,The product of likelihoods across all route observations are expressed as $\prod_i \prod_k \prod_t p_i^{r_{i,k}}(\boldsymbol{\beta}, \boldsymbol{E}(t, r_{i,k}))$, or $\prod_i \prod_k \prod_t p_i(r_{i,k}; \boldsymbol{\beta}, \boldsymbol{E}(t, r_{i,k}))$.

The Maximum Likelihood Estimator $\hat{\boldsymbol{\beta}}_{MLE}$ is:

$$\hat{\boldsymbol{\beta}}_{MLE} = \arg\max \prod_i \prod_k \prod_t p_i(r_{i,k}; \boldsymbol{\beta}, \boldsymbol{E}(t, r_{i,k})). \tag{9}$$

Taking the logarithm in above equation, we have:

$$L(\hat{\boldsymbol{\beta}}) = \arg\max \sum_i \sum_k \sum_t \log(p_i(r_{i,k}; \boldsymbol{\beta}, \boldsymbol{E}(t, r_{i,k}))). \tag{10}$$

However, the true OD pair w_i for traveler i is unknown and can be treated as a hidden/latent variable. The above $p_i(r_{i,k}; \boldsymbol{\beta}, \boldsymbol{E}(t, r_{i,k}))$ is marginal density function found by summing over all the latent variables w (eligible true OD). A new likelihood function can now be defined by working with the joint distribution of the route $r_{i,k}$ and the unobserved true OD pair w. Also, in this equation, Q_w^i is some distribution of w.

$$
\begin{aligned}
L(\hat{\boldsymbol{\beta}}) &= \arg\max \sum_i \sum_k \sum_t \log(\sum_w p_i(r_{i,k}, w; \boldsymbol{\beta}, \boldsymbol{E}(t, r_{i,k}))) \\
&= \arg\max \sum_i \sum_k \sum_t \log(\sum_w Q_w^i \frac{P_t(r_{i,k}, w; \boldsymbol{\beta}, \boldsymbol{E}(t, r_{i,k}))}{Q_w^i}).
\end{aligned} \tag{11}
$$

By using the Jensen's inequality for log-concave functions, the log likelihood becomes:

$$
L(\hat{\boldsymbol{\beta}}) >= \arg\max \sum_i \sum_k \sum_t \sum_w Q_w^i \log(\frac{p_i(r_{i,k}, w; \boldsymbol{\beta}, \boldsymbol{E}(t, r_{i,k}))}{Q_w^i}). \tag{12}
$$

To make a bound tight for a particular value of $\boldsymbol{\beta}$, Q_w^i can be constructed as follows for each traveler:

$$
Q_w^i = p_i(w|\boldsymbol{r_{i,k}}; \boldsymbol{\beta}, \boldsymbol{E}(t, r_{i,k})). \tag{13}
$$

Let's denote the trip history of a traveler t as $\boldsymbol{r_{i,k}}$. Since we assume each traveler will not change their true OD, Q_w^i now is:

$$
\begin{aligned}
Q_w^i &= p_i(w|r_{i,1}, r_{i,2}, ...r_{i,K_i}; \boldsymbol{\beta}, \boldsymbol{E}(t, r_{i,k})) \\
&= \frac{p_i(r_{i,1}, r_{i,2}, ...r_{i,K_i}|w; \boldsymbol{\beta}, \boldsymbol{E}(t, r_{i,k}))p_i(w)}{\sum_w p_i(r_{i,1}, r_{i,2}, ...r_{i,K_i}|w; \boldsymbol{\beta}, \boldsymbol{E}(t, r_{i,k}))p_i(w)} \\
&= \frac{p_i(r_{i,1}|w; \boldsymbol{\beta}, \boldsymbol{E}(t, r_{i,1}))p_i(r_{i,2}|w; \boldsymbol{\beta}, \boldsymbol{E}(t, r_{i,2}))...p_i(r_{i,K_i}|w; \boldsymbol{\beta}, \boldsymbol{E}(t, r_{i,K_i}))p_i(w)}{\sum_w p_i(r_{i,1}|w; \boldsymbol{\beta}, \boldsymbol{E}(t, r_{i,1}))p_i(r_{i,2}|w; \boldsymbol{\beta}, \boldsymbol{E}(t, r_{i,2}))...p_i(r_{i,K_i}|w; \boldsymbol{\beta}, \boldsymbol{E}(t, r_{i,K_i}))p_i(w)}.
\end{aligned} \tag{14}
$$

This means that Q_w^i is the probability of traveler i choosing w as his/her true OD given he/she takes the routes $r_{i,1}, r_{i,2}, ...r_{i,K_i}$ and his/her preference is $\boldsymbol{\beta}$. Therefore, the E step and M step for expectation maximization algorithm are clear:

The E step is plug in $\hat{\boldsymbol{\beta}}$ from M step and calculate Q_w^i:
for each observed route $r_{i,k}$ and each true OD pair w:

$$
\begin{aligned}
Q_w^t &= p_i(w|r_{i,1}, r_{i,2}, ...r_{i,K_i}; \boldsymbol{\beta}, \boldsymbol{E}(t, r_{i,k})) \\
&= \frac{p_i(r_{i,1}|w; \boldsymbol{\beta}, \boldsymbol{E}(t, r_{i,1}))p_i(r_{i,2}|w; \boldsymbol{\beta}, \boldsymbol{E}(t, r_{i,2}))...p_i(r_{i,K_i}|w; \boldsymbol{\beta}, \boldsymbol{E}(t, r_{i,K_i}))p_i(w)}{\sum_w p_i(r_{i,1}|w; \boldsymbol{\beta}, \boldsymbol{E}(t, r_{i,1}))p_i(r_{i,2}|w; \boldsymbol{\beta}, \boldsymbol{E}(t, r_{i,2}))...p_i(r_{i,K_i}|w; \boldsymbol{\beta}, \boldsymbol{E}(t, r_{i,K_i}))p_i(w)}.
\end{aligned} \tag{15}
$$

In above function, $p_i(r_{i,k}|w; \boldsymbol{\beta}, \boldsymbol{E}(t, r_{i,k}))$ is the probability of a traveler i's choice of route under occasion k: choose $r_{i,k}$ as his/her route given his/her true OD is w, his/her preference is $\boldsymbol{\beta}$, with the route attributes $\boldsymbol{E}(t, r_{i,k})$. This will be calculated from above-mentioned path size logit model. $p_i(w), w \in \boldsymbol{w}$ are prior probabilities that are assumed equal or calculated based on other information, such as population around origins and destinations.

The M step is to plug in Q_w^i from E step and find $\hat{\boldsymbol{\beta}}$:

$$
\begin{aligned}
L(\hat{\boldsymbol{\beta}}) &= \arg\max \sum_i \sum_k \sum_t \sum_w Q_w^i \log(\frac{p_i(r_{i,k}, w; \boldsymbol{\beta}, \boldsymbol{E}(t, r_{i,k}))}{Q_w^i}) \\
&= \arg\max \sum_i \sum_k \sum_t \sum_w Q_w^i \log(\frac{p_i(r_{i,k}|w; \boldsymbol{\beta}, \boldsymbol{E}(t, r_{i,k}))p_i(w)}{Q_w^i}),
\end{aligned} \tag{16}
$$

where $p_i(r_{i,k}|w;\boldsymbol{\beta},\boldsymbol{E}(t,r_{i,k}))$ is the route choice probability calculated based on path size logit model. Q_w^i is the output from above E step.

We will repeat these two steps until convergence to estimate the traveler preference $\boldsymbol{\beta}$. At the same time, the true OD of each observation could also be extracted, once we get the best estimator $\hat{\boldsymbol{\beta}}$:

$$\hat{w}_i = \arg\max \prod_k \prod_t p_i(w|r_{i,k};\boldsymbol{\beta},\boldsymbol{E}(t,r_{i,k})). \tag{17}$$

5 Case Study and Results

For our study, we use AFC data that contains all the metro and bus routes within Seoul Metropolitan area. This data consist of 12 weeks transactions providing sufficient cases of same OD trips. During those 12 weeks, more than 1 billion transaction records were generated. Each record contains information such as card ID number, boarding/alighting stations and times, etc. We develop computational algorithms to handle scalability given the large data set and discuss the results of the proposed methodology.

5.1 Scaling AFC data

Our original plan was to infer the true ODs. However, there is no ground truth of true OD information in AFC data, thus, it is impossible to verify the accuracy of inference. Therefore, we assume the first boarding station and last alighting station (station-level OD) are origin and destination for a multi-modal trip. If we have a multi-modal trip starting and ending both at bus station, but involve metro trip in between, we can scale back this multi-modal trip and use it as instances for ODIP model. By scaling back one level and assuming we only know a part of the AFC trip (the metro part), we will be able to apply and validate our ODIP model. Now the true origin and destination (w_t) of traveler t are the first boarding and last alighting stations. The known information would be all the travel environment between this true OD. The ODIP problem now becomes a station-level OD inference problem with only knowing the trip history of metro users(See Figure 1).

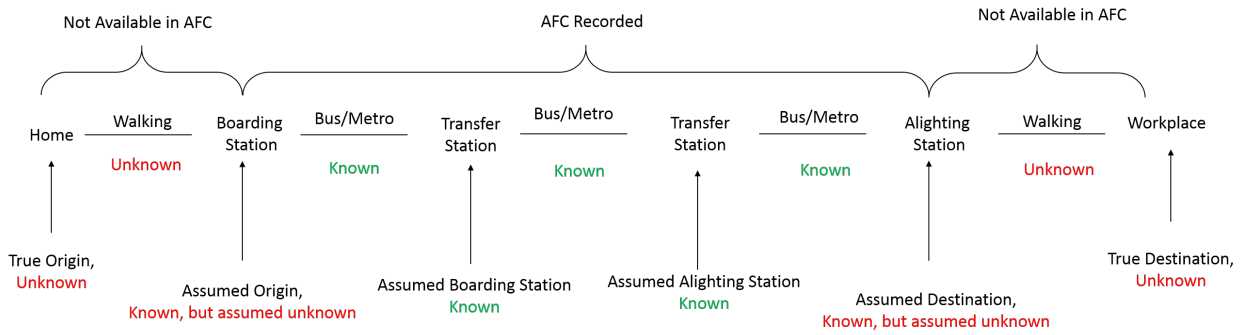

Figure 1: Typical Trip and Assumed Trip using AFC

5.2 OD estimation

By scanning each traveler's trip history and applying eligible OD analysis, we find out that around 93.5% of the travelers' true ODs can be limited to only one candidate. This only candidate

usually has a much higher probability of being true OD than the second-ranked candidate. That means, for those travelers, their ODs can be inferred directly without performing any further analysis(Expectation Maximization) .

However, there are still 6.5% of total metro users have more than one OD candidate after eligible OD analysis. We have to apply EM method to find out the true ODs of these 6.5% total metro users. In our study, we find out EM method will converge after certain iterations. Result shows that true ODs can be accurately identified by performing EM method for around 65% of these metro users, when comparing to the ground truth. By combining eligible OD analysis and EM method, our ODIP model can successfully infer traveler's true ODs for 93% of the total users.

	Inferred by Eligible OD Analysis	Inferred by EM method	Total
Number of Travelers	2851	197	3048
Percentage	93.54%	6.46%	100%
Accuracy	∼95%	∼65%	∼93%

Table 2: Results from Eligible OD Analysis and EM Inference Methods

5.3 Traveler preference estimation

Since the preference vector can also be estimated by using traditional discrete choice model, it can serve as a validation for the preference vector inferred from EM method. Our result shows that results from traditional discrete choice model and EM method are close to each other.

We tested two route choice models: in the first one, we choose to combine the multi-modal travel environment as one parameter to estimate. In the other model, we distinguish bus and metro segments in the same trip, and treat travel environment (travel time, trasfer time, etc.) in these two modals as separate parameters. By fitting training data, we could conclude that the latter model performs better, since it has a higher Macfadden's R square index. The summary table and fitted parameters are as follows.

Now one can assess the impact of different factors on a traveler's probability of choosing a particular route. In fact, all the factors turn out to have statistically significant non-zero estimated coefficients at a 1% significance level. Most of the factors have a negative effect on the probability of choosing a route, which indeed confirms that these factors (i.e., higher travel time, higher cost, etc.) increase disutility instead of improving utility. The number of transfers comes out as the most important factor for travelers to consider when choosing routes; we also find that they consider whether a route has common segments with other alternatives. Transfer time is the next biggest factor by importance, while travel cost and metro travel time are of smaller concern.

Level of service of bus influence the probability of choosing a route positively. This means that, if an increase of metro and/or bus levels of service is achieved, the probability of a traveler choosing an improved route will increase (for most travelers). However, the level of service of metro affect traveler's choice by the opposite way.

Possible reasons of such positive effects of bus travel time are that bus stations are more accessible compared to metro stations, which will lead to a preference to trips that having a higher proportion of bus segments. The attractiveness of crowdedness can be attributed to the high demand of such routes: instead of capturing the negative disutility of crowdedness, crowdedness is also representing how popular a route is.

Also, the bus travel time has a positive effect on the probability of choosing a route. Positive signs for bus travel time may be attributed to the high accessibility of bus stations: walking times

to bus stations are usually shorter than walking times to metro stations.

	Model 1	Macf R Sq: 0.008		Model 2	Macf R Sq: 0.019		EM	
Variable	Selected?	Coefficient	Sig	Selected?	Coefficient	Sig	Selected?	Coefficient
Intercept	Y	-5.20E-15		Y	-3.49E-15			
Travel Time Total	Y	4.20E-04	*					
Travel Time Bus				Y	7.49E-04	***	Y	9.86E-04
Travel Time Metro				Y	-4.01E-04	*	Y	-5.17E-04
Transfer Time	Y	-1.06E-03	***	Y	-1.01E-03	***	Y	-6.13E-04
Number fo Transfers	Y	-6.72E-01	***	Y	-7.01E-01	***	Y	-2.22E-01
Travel Cost(Price)	Y	3.18E-04	.	Y	3.12E-04	.	Y	4.35E-04
Crowdedness Total	Y	7.44E-05	*					
Crowdedness Bus				Y	5.26E-04	***	Y	2.98E-04
Crowdedness Metro				Y	-3.11E-05		Y	-4.74E-05
Path Size Correction Total	Y	1.43E+00	***					
Path Size Correction Bus				Y	4.60E-01	***	Y	3.65E-01
Path Size Correction Metro				Y	2.56E-01		Y	-3.10E-01
Sig: *** 0.001 ** 0.01 * 0.05 . 0.1 1								

Table 3: Comparison of Model Coefficient between fitted Discrete Choice Model and EM method

6 Conclusion

In this paper, we develop and apply an inference framework for distilling multi-modal routing preferences for transit system users and their true ODs, based on a real-world AFC data set, through probabilistic learning built upon Expectation Maximization approach. The inferential power of the proposed methods and tools stems from the observations of how the travelers revise their choices (recorded by transactions) on perturbations of travel environment conditions.

In order to solve our ODIP model, we use EM method to gradually updated the true OD andutility weightfor travelers to the optimal. This model had been applied and validated by using AFC data from Seoul, South Korea. Due to lack of true OD information as validation data, we scaled back our data and only infer the first boarding and last alighting bus stations (station-level OD) for metro users. By combining eligible OD analysis and EM method, we could solve ODIP problem and infer traveler's OD with an accuracy of 93%. At the same time, estimated travel preferences from EM method is close to the results from traditional discrete route choice model.

The methodology adopted in this paper can be extended to infer the true OD of AFC card users in the future. However, it is till a challenge to pick out eligible OD candidates without prior knowledge. The data-driven method proposed in this paper is highly restricted in many applications.

References

Barry, J., Newhouser, R., Rahbee, A., and Sayeda, S. (2002a). Origin and destination estimation in new york city with automated fare system data. *Transportation Research Record: Journal of the Transportation Research Board*, (1817):183–187.

Barry, J. J., Newhouser, R., Rahbee, A., and Sayeda, S. (2002b). Origin and destination estimation in new york city with automated fare system data. *Transportation Research Record*, 1817:183–187.

Ben-Akiva, M. and Bierlaire, M. (2003). Discrete choice models with applications to departure time and route choice. In *Handbook of transportation science*, pages 7–37. Springer.

Ben-Akiva, M. E. and Lerman, S. R. (1985). *Discrete choice analysis: theory and application to predict travel demand*. The MIT press.

Bovy, P. H. and Hoogendoorn-Lanser, S. (2005). Modelling route choice behaviour in multi-modal transport networks. *Transportation*, 32(4):341–368.

Chakirov, A. and Erath, A. (2011). Use of public transport smart card fare payment data for travel behaviour analysis in singapore. *[Arbeitsberichte/IVT]*, 729.

Chan, J. (2007). Rail transit od matrix estimation and journey time reliability metrics using automated fare data. Master's thesis, MIT.

Cui, A. (2006a). Bus passenger origin-destination matrix estimation using automated data collection systems. Master's thesis, MIT.

Cui, A. (2006b). Bus passenger origin-destination matrix estimation using automated data collection systems. Master's thesis, Massachusetts Institute of Technology.

Gordillo, F. (2006a). The value of automated fare collection data for transit planning: an example of rail transit od matrix estimation. Master's thesis, MIT.

Gordillo, F. (2006b). The value of automated fare collection data for transit planning: an example of rail transit od matrix estimation. Master's thesis, Massachusetts Institute of Technology.

Hazelton, M. L. (2008). Statistical inference for time varying origin–destination matrices. *Transportation Research Part B: Methodological*, 42(6):542–552.

Hensher, D. A. and Greene, W. H. (2003). The mixed logit model: the state of practice. *Transportation,*, 30:133–176.

Kumar, A. A., Kang, J. E., Kwon, C., and Nikolaev, A. (2016). Inferring origin-destination pairs and utility-based travel preferences of shared mobility system users in a multi-modal environment. *Transportation Research Part B: Methodological*, 91:270–291.

Kusakabe, T., Iryo, T., and Asakura, Y. (2010). Estimation method for railway passengers train choice behavior with smart card transaction data. *Transportation*, 37(5):731–749.

Lee, S. G. and Hickman, M. (2014). Trip purpose inference using automated fare collection data. *Public Transport*, 6(1-2):1–20.

Lianfu, Z., Shuzhi, Z., Yonggang, Z., and Ziyin, Z. (2007). Study on the method of constructing bus stops od matrix based on ic card data. *Wireless Communications, Networking and Mobile Computing WiCom*, pages 3147–3150.

Ma, X., Wu, Y.-J., Wang, Y., Chen, F., and Liu, J. (2013). Mining smart card data for transit riders travel patterns. *Transportation Research Part C: Emerging Technologies*, 36:1–12.

McFadden, D. et al. (1973). Conditional logit analysis of qualitative choice behavior.

McMullan, A. and Majumdar, A. (2012). Assessing the impact of travel path choice on london's rail network using an automatic fare collection system. *Transportation Research Record: Journal of the Transportation Research Board*, (2274):154–163.

Munizaga, M. A. and Palma, C. (2012). Estimation of a disaggregate multimodal public transport origin–destination matrix from passive smartcard data from santiago, chile. *Transportation Research Part C: Emerging Technologies*, 24:9–18.

Nassir, N., Khani, A., Lee, S., Noh, H., and Hickman, M. (2011). Transit stop-level origin-destination estimation through use of transit schedule and automated data collection system. *Transportation Research Record: Journal of the Transportation Research Board*, (2263):140–150.

Pelletier, M.-P., Trépanier, M., and Morency, C. (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, 19(4):557–568.

Sun, L., Lee, D.-H., Erath, A., and Huang, X. (2012). Using smart card data to extract passenger's spatio-temporal density and train's trajectory of mrt system. In *Proceedings of the ACM SIGKDD international workshop on urban computing*, pages 142–148. ACM.

Sun, L., Lu, Y., Jin, J. G., Lee, D.-H., and Axhausen, K. W. (2015). An integrated bayesian approach for passenger flow assignment in metro networks. *Transportation Research Part C: Emerging Technologies*, 52:116–131.

Sun, Y., Shi, J., and Schonfeld, P. M. (2016). Identifying passenger flow characteristics and evaluating travel time reliability by visualizing afc data: a case study of shanghai metro. *Public Transport*, 8(3):341–363.

Sun, Y. and Xu, R. (2012). Rail transit travel time reliability and estimation of passenger route choice behavior: Analysis using automatic fare collection data. *Transportation Research Record: Journal of the Transportation Research Board*, (2275):58–67.

Trepanier, M., Tranchant, N., and Chapleau, R. (2007). Individual trip destination estimation in a transit smart card automated fare collection system. *Journal of Intelligent Transportation Systems: Technology, Planning and Operations*, 11:1–14.

Trépanier, M., Tranchant, N., and Chapleau, R. (2007). Individual trip destination estimation in a transit smart card automated fare collection system. *Journal of Intelligent Transportation Systems*, 11(1):1–14.

Wardman, M. (2004). Public transport values of time. *Transport policy*, 11(4):363–377.

Zhao, J. (2004). The planning and analysis implications of automated data collection systems: rail transit od matrix inference and path choice modeling examples. Master's thesis, Massachusetts Institute of Technology.

Zhao, J., Rahbee, A., and Wilson, N. H. (2007). Estimating a rail passenger trip origin-destination matrix using automatic data collection systems. *Computer-Aided Civil and Infrastructure Engineering*, 22(5):376–387.

Zhou, F. and Xu, R.-h. (2012). Model of passenger flow assignment for urban rail transit based on entry and exit time constraints. *Transportation Research Record: Journal of the Transportation Research Board*, (2284):57–61.

Zhu, W., Hu, H., and Huang, Z. (2014). Calibrating rail transit assignment models with genetic algorithm and automated fare collection data. *Computer-Aided Civil and Infrastructure Engineering*, 29(7):518–530.