Transportation Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Ex Post Path Choice Estimation for Urban Rail Systems Using Smart Card Data: An Aggregated Time-Space Hypernetwork Approach

Baichuan Mo, Zhenliang Ma, Haris N. Koutsopoulos, Jinhua Zhao

https://pubsonline.informs.org/journal/trsc

TRANSPORTATION SCIENCE
Articles in Advance, pp. 1–23
ISSN 0041-1655 (print), ISSN 1526-5447 (online)

Ex Post Path Choice Estimation for Urban Rail Systems Using Smart Card Data: An Aggregated Time-Space Hypernetwork Approach

Baichuan Mo,[a] **Zhenliang Ma,**[b,*] **Haris N. Koutsopoulos,**[c] **Jinhua Zhao**[d]

[a] Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139; [b] Department of Civil and Architectural Engineering, KTH Royal Institute of Technology, Stockholm 11428, Sweden; [c] Department of Civil and Environmental Engineering, Northeastern University, Boston, Massachusetts 02115; [d] Department of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139
*Corresponding author

Contact: baichuan@mit.edu, https://orcid.org/0000-0003-0323-0066 (BM); zhema@kth.se, https://orcid.org/0000-0002-2141-0389 (ZM); h.koutsopoulos@northeastern.edu (HNK); jinhua@mit.edu, https://orcid.org/0000-0002-1929-7583 (JZ)

Received: December 28, 2021
Revised: June 21, 2022; August 14, 2022
Accepted: August 16, 2022
Published Online in Articles in Advance:
October 18, 2022

https://doi.org/10.1287/trsc.2022.1177

Abstract. This paper proposes an ex post path choice estimation framework for urban rail systems using an aggregated time-space hypernetwork approach. We aim to infer the actual passenger flow distribution in an urban rail system for any historical day using the observed automated fare collection (AFC) data. By incorporating a schedule-based dynamic transit network loading (SDTNL) model, the framework captures the crowding correlation among stations and the interaction between the path choice and passenger left behind, which is important for the path choice estimation in a "near-capacity" operated urban rail system. The path choice estimation is formulated as an optimization problem, which aims to minimize the difference between the model-derived and observed information with path choice parameters as decision variables. The original problem is intractable because of nonlinear (logit model) and nonanalytical (SDTNL) constraints. A solution procedure is proposed to decompose the original problem into three tractable subproblems, which can be solved efficiently. Solving the decomposed problem is equivalent to finding a fixed point. We prove that the solution to the original problem is the same as the decomposed problem (i.e., the fixed point) when passenger path choices follow the predefined behavior model. If this condition does not hold, the solution of the original problem is proved to be an "almost fixed point" for the decomposed problem. The model is validated using both synthetic and real-world AFC data from a major urban railway system. The analysis with synthetic data validates the model's effectiveness in estimating path choice parameters and left behind probabilities, which outperforms state-of-art simulation-based optimization methods and probabilistic models in both accuracy and efficiency. The analysis using actual data shows that the estimated path shares are more reasonable than the baseline uniform path shares and survey-derived path shares. The model estimation is robust to different initial parameter values and AFC data from various dates.

Funding: The project is partially funded by the TRENoP Strategic Research funding at KTH.
Supplemental Material: The online appendices are available at https://doi.org/10.1287/trsc.2022.1177.

Keywords: path choice estimation • smart card data • time-space hypernetwork • simulation-based optimization

1. Introduction

With increasing urbanization, the urban rail systems are playing an important role in urban transportation. Understanding passenger flow distribution in urban rail systems is crucial for designing operating strategies and better accommodating passengers. Simulation and transit loading models are powerful instruments to infer and predict passenger flows in the network and hence monitor and evaluate system performance. Two important inputs are required for these models: the origin-destination (OD) demand and passengers' path choices. With the availability of data from automated fare collection (AFC) systems, station-to-station OD flows in urban rail networks are readily available, especially for close systems with both tap-in and tap-out fare validations (Koutsopoulos et al. 2019). However, path choices are not observed directly. Therefore, estimating path choices is an essential requirement for understanding passenger flow distributions and monitoring system performance.

On-site surveys are typically used to estimate path choices. However, surveys are time-consuming and labor intensive. In addition, given the changes in operating characteristics and performance of an urban rail system, survey results may quickly become outdated. To overcome these disadvantages, researchers have proposed path choice estimation methods using AFC data. In closed urban rail systems, AFC data include locations and times of both tap-in and tap-out transactions. AFC data-based methods for path choice estimation can be categorized into two groups: path-identification methods (Kusakabe, Iryo, and Asakura 2010, Zhou and Xu 2012, Kumar, Khani, and He 2018, Zhu, Koutsopoulos, and Wilson 2021) and parameter-inference methods (Sun and Xu 2012, Sun et al. 2015, Zhao et al. 2017, Xu et al. 2018). The former studies aim to identify the exact path chosen by a user. The path attributes (e.g., walking time and in-vehicle time) are used to evaluate how likely a certain path is for a passenger. The later studies formulated probabilistic models to describe the random process of passengers' path choice behavior. Bayesian inference is usually used to estimate the corresponding choice parameters or path choice fractions. Despite using different methods, the key idea for those AFC data-based approaches are similar. They all attempt to match the model-derived journey times with the observed journey times from AFC data. Because the model-derived journey times are determined by the choice parameters, observed journey times provide indirect measurements to calibrate path choices. However, this type of method may fail if left behind (also called denied boarding, which means passengers are not able to board the first train on their arrival on the platform due to limited train capacity) is not taken into account.

Left behind causes passengers' waiting time on the platform to increase, thus increasing their total travel times. It may happen that the journey time for a longer route without left behind is close to that of a shorter route with left behind, which makes the two routes indistinguishable using the pure journey time-based methods (Zhu 2017, Zhu, Koutsopoulos, and Wilson 2021). Several studies have taken left behind into consideration explicitly or implicitly. For example, Sun et al. (2015) considered the delay caused by the left behind as part of travel time variability. This method is unable to distinguish the choice of routes with very similar journey time distributions. Sun and Xu (2012) and Zhao et al. (2017) assumed that the left behind probabilities for different stations are independent and explicitly estimated the left behind probability before inferring the path choice fraction. These partially addressed the left behind problem.

However, the independence assumption neglects important interactions among stations. In the real world, left behind is caused by the interaction between supply and demand. A station with high entry demand may cause the adjacent stations to be congested because the remaining capacity for the next station will become limited. Therefore, the left behind probabilities for different stations are dependent. Moreover, it is not reasonable to consider path choice and left behind separately. These two components are interconnected and affect passenger journey times collectively. Thus, the path choice estimation model needs to consider the correlation of left behinds among platforms, as well as the interactions between path choice and left behind. One of the solutions is to incorporate a schedule-based dynamic transit network loading (SDTNL) model (Mo et al. 2020) for the model-derived journey time estimation.[1] The SDTNL simulates train movements and passenger boarding, alighting, and left behind behaviors explicitly with OD demands and path choices as inputs. The left behind correlation at different stations and its interactions with path choices can be naturally endogenized and included. However, as is known in the literature, the SDTNL is a complicated simulation process with no analytical formulation (Song et al. 2017). There is no direct way to write the mathematical formulations of the model-derived journey time as a function of the SDTNL. Hence, none of the aforementioned studies have used the SDTNL model for the path choice estimation problem (ignore important interactions as mentioned previously).

This paper proposes a novel path choice estimation framework to incorporate the SDTNL model. It captures left behind correlations among platforms and interactions between path choice and left behind. The key idea is to convert the "model-derived journey time" to a new concept named "path exit rates," in which an aggregated time-space (TS) hypernetwork is introduced along with the newly defined aggregated network flows (i.e., OD entry flows, OD entry-exit flows, and path flows). The aggregated TS hypernetwork facilitates fast computation by taking all passengers' information into consideration, thus leading to more efficient path choice estimations for large-scale urban rail systems. The objective is to minimize the difference between model-derived and observed "OD entry-exit flows" with path choice parameters in the choice model as decision variables (details in Section 3).

The original problem is intractable due to the nonanalytical SDTNL model and the nonlinear logit choice model constraints. We decompose the original problem into three tractable subproblems: rough path shares estimation, choice parameters estimation, and path exit rates estimation, and solve them iteratively to approximate the original solutions. We prove that the solution of the decomposed problem is equivalent to that of the original problem under specific conditions. The model is validated using data from the Hong Kong Mass Transit Railway (MTR) system. The results affirm the effectiveness and robustness of the proposed method in path choice estimation.

The contributions of the paper are as follows:

- Proposing a novel path choice estimation framework that incorporates the SDTNL model. It captures left behind correlations among platforms and dynamic interactions between passenger's path choices and left behind. Similar to the commonly used probabilistic approach in existing studies, the proposed model makes use of the information of observed trip journey times and flows in travel records. However, it is purely data driven without using prior information or making independent assumptions, whereas typical probabilistic approaches assume independence of individual travels and trip components and need prior information on trip component time distributions (e.g., left behind distribution).

- The proposed path choice estimation framework is optimization-based and uses the information of all passengers' travel records (i.e., AFC data of all passengers, captured by time-dependent OD entry-exit flows as illustrated in Section 3.1). Typical probabilistic approaches only use samples (i.e., selected AFC data records) from certain OD pairs in the network due to the computational challenges (Sun et al. 2015). Results in the case study show that the probabilistic models, although only use 20% of the passengers in the system, have a longer model running time than the proposed model.
- Proposing an aggregate TS hypernetwork using the "path exit rates" to capture the model-derived trip travel time information. In the aggregated TS hypernetwork, the vehicle travel times and congestion information cannot be directly modeled as the link cost and left behind probability, respectively, as in previous studies. Path exit rates capture the vehicle travel times and congestion loads in an aggregate way. The aggregate TS hypernetwork simplifies the temporal representation of network flows, which facilitates the modeling of large-scale networks while preserving the flow dynamics across time.
- Proposing a decomposition method and solution algorithm with three tractable subproblems to solve the original problem efficiently. We show that solving the decomposed problem is equivalent to finding a fixed point. We prove that the solution to the original problem is the same as the decomposed problem (i.e., the fixed point) when the passenger's path choices follow the predefined behavior model. When this condition does not hold, the solution of the original problem is proved to be an "almost fixed point" for the decomposed problem.

It is worth noting that the "ex post" path choice estimation is different from the typical transit assignment and traffic assignment problems in the literature. The former infers the actual (realized) flow distribution using the observed AFC data, whereas the latter forecasts "hypothetical" flow patterns by assuming user equilibrium or system optimal criteria. Thus, these problems have different settings and challenges.

The remainder of this paper is organized as follows: Section 2 reviews related studies in the literature. Section 3 describes the modeling framework, including network representation, problem definition, and solution procedures. The model's effectiveness and robustness are validated using both synthetic and actual data in Section 4. The main findings and future research directions are summarized in Section 5.

2. Literature Review

2.1. Path Choice Estimation for Urban Rail Systems

Considerable literature exists on rail transit path choice estimation. Stated preference (SP) and revealed preference (RP) surveys are often used for the estimation of path choices. For example, Lam and Xie (2002) applied a path-size logit model to estimate route choices in Singapore's urban rail system using the mixed SP and RP data. Nazem, Trépanier, and Morency (2011) adopted a discrete choice model to estimate passengers' route choice behavior for different demographic groups using the household travel survey in Canada. Eluru, Chakour, and El-Geneidy (2012) developed a mixed logit model to study transit route choices in Montreal, Canada, using data from a Google Map–based RP survey. A methodological review on survey-based route choice estimation can be found in Prato (2009).

The emergence of smart card data has shifted the research toward data-driven path choice estimation using historical fare transaction records rather than using surveys. As mentioned in Section 1, these studies can be categorized into two categories: path-identification methods and parameter-inference methods. In terms of path identification, Kusakabe, Iryo, and Asakura (2010) proposed an algorithm to identify the exact train that a passenger boarded using smart card data, which then gave the path choice. Based on a case study in Japan, the model was implicitly validated using the train load weight data and GPS trajectories of probe passengers. Zhou and Xu (2012) proposed a path identification method using the "maximum likelihood boarding plan" method. It assumes that each individual will choose the path with the highest degree of match between the model-derived and observed journey times. Actual passenger data from the Beijing subway system were used for a case study. Kumar, Khani, and He (2018) proposed a trip chaining method to infer the most likely trajectory of transit passengers using AFC and General Transit Feed Specification (GTFS) data. The method was applied using smart card data from the Twin Cities subway system and implicitly verified using automatic passenger count data. Zhu, Koutsopoulos, and Wilson (2021) formulated the likelihood of passengers choosing an itinerary (i.e., time-dependent path) based on their tap-in and tap-out times, and estimated passengers' path choices using the maximum likelihood estimation method. Path-identification methods have several limitations. For example, they require and are sensitive to predetermined model parameters (e.g., walking speed, crowding) that are difficult to calibrate in advance. Also, they assume independence of journey time components for an individual. It tends to be computationally expensive to construct the likelihood of individuals' observations in large-scale networks with high travel demand. The estimation results could be biased for OD pairs with limited travel observations.

The parameter-inference methods estimate the network-level path choice that is more suitable for system performance evaluation than the path-identification methods. They model path shares as a function of path attributes using behavioral models (e.g., discrete choice models) and estimate the corresponding model parameters using observed AFC data. Sun and Xu (2012) proposed a probabilistic model for path choice estimation using AFC data. They first estimated the platform elapsed times for transfer and through stations and used a Gaussian mixture model to estimate path choice fractions based on the journey time distribution. The model was validated with a simple synthetic data set and applied to the Beijing urban rail system. Sun et al. (2015) proposed an integrated Bayesian approach to estimate the network-level path choices. Path choices are captured by a multinomial logit model with parameters to be estimated. The model is implemented with data from the Singapore MRT system. Zhao et al. (2017) proposed a probabilistic model to estimate path choice fractions using AFC data. They first estimated the number of trains waited by passengers, which is equivalent to the left behind rate. Then the path choice fractions were modeled and estimated based on a Gaussian mixture model. Xu et al. (2018) proposed a Bayesian inference approach to estimate the path choice parameters of a logit model using AFC data. Metropolis-Hasting sampling was used to calibrate the model parameters.

2.2. Time-Space Hypernetworks

The time-space (TS) hypernetwork representation is commonly used for representation of scheduled transit systems. In the spatial representation, the nodes represent the entrance, boarding platform, alighting platform, and exit, whereas the arcs represent entering, waiting, boarding, service (on trains), transferring, and exiting. The TS hypernetwork expands the spatial representation and captures the temporal flow interactions (e.g., left behind from previous trips) (Schmöcker, Bell, and Kurauchi 2008, Stasko, Levine, and Reddy 2016). Depending on the application, different variants of the TS hypernetwork are proposed in the transit assignment literature, for example, a node-path network (Han et al. 2015, Kroon, Maróti, and Nielsen 2015), route-section representation (a section arc is created to classify passengers belonging to the same OD pair) (Szeto and Jiang 2014), service line-node network (Szeto et al. 2013), line link-node network (Hamdouch et al. 2011), and diachronic graph (Nuzzolo, Crisalli, and Rosati 2012). In essence, the network representation is a tradeoff between realistic flow assignment (strict capacity constraints, first come first serve, left behind) and computational complexity to effectively solve the problem.

The typical fine-grained TS hypernetwork representation is modeled at a granular level. For example, the hypernetworks have explicit links to represent left behind or walking behavior. This granular modeling requires the time interval to be at least as small as a headway. However, in our study, we propose an aggregated TS hypernetwork. The time interval can be 10 minutes or 15 minutes (much greater than a headway). In our aggregate TS hypernetwork, there is no explicit left behind links or in-vehicle links. The travel time and congestion (i.e., left behind) between two TS nodes are jointly captured by the "path exit rate." Details can be found in Section 3.1.

2.3. Research Gaps

As mentioned in Section 1, the left behind is important in estimating network-level path choices. However, few studies have addressed this problem satisfactorily. The pure journey time-based methods (Sun et al. 2015, Xu et al. 2018) considered the waiting time caused by left behind as part of the total journey time, which cannot distinguish long paths without left behind and short paths with left behind because they have very similar total journey times. Sun and Xu (2012) and Zhao et al. (2017) assumed left behinds are independent across stations and considered the left behind and path choice problems separately, which neglects the interaction between supply and demand in the network. Thus, a comprehensive path choice estimation framework that can capture the left behind correlations among platforms, as well as the interactions between path choice and left behind, is needed to advance the current state-of-the-art.

To capture these interactions, the SDTNL model is essential (Hamdouch and Lawphongpanich 2008, Hamdouch et al. 2011, Mo et al. 2020). In the SDTNL model, a vehicle's movement is assumed to follow a fixed schedule or real-world automated vehicle location (AVL) data. The model defines an event as a vehicle's arrival or departure at a specific station (or platform). Each event can be indexed by the corresponding platform and time. Passengers boarding and alighting at each vehicle's arrival and departure events are processed one at a time and in the topological and chronological order, that is, an event whose platform with no predecessor and with the smallest time index is processed first. In this way, when loading passengers into a vehicle, the available capacity of a vehicle (which is determined by previously boarding passengers) is known, and the left behind can be modeled if the number of waiting passengers exceed the capacity. After processing all events, the model-derived passengers' journey times are obtained. These journey times take left behind into consideration in an endogenous way.

The frequency-based static transit network loading (FSTNL) model, although it has good mathematical formulations, cannot estimate the model-derived journey times well due to the difficulty in estimating waiting times caused by the left behind. In the FSTNL model (De Cea and Fernández 1993, Wu, Florian, and Marcotte 1994, Spiess and Florian 1989, Nielsen 2000, Schmöcker et al. 2011), the waiting time is either assumed to be reversely proportional to the (effective) service frequency (Wu, Florian, and Marcotte 1994, Nielsen 2000, Schmöcker et al. 2011), or modeled as a congestion function (e.g., BRP) of previously boarded flows and new arrival flows with exogenously calibrated parameters (De Cea and Fernández 1993). The former method does not consider the left behind (or considers the left behind exogenously). The latter method only outputs a generalized waiting cost (rather than the waiting time) as the vehicle capacity is not explicitly modeled. Therefore, the FSTNL model is not suitable for path choice estimation in a near-capacity transit system where left behind has to be considered. Therefore, in this paper, we use the SDTNL model for the travel time (in-vehicle time and waiting time with left behind) estimation and propose a decomposition method to solve the nonanalytical SDTNL problem.

3. Methodology

3.1. Network Representation

To capture the path choice and left behind interactions, the model-derived journey times are estimated from an SDTNL model. A network loading process assumes that the passengers' path choices are known and treated as an input (Song et al. 2017). A typical way to represent a transit network with schedule information is using a TS hypernetwork (Nguyen, Pallottino, and Malucelli 2001, Hamdouch and Lawphongpanich 2008, Hamdouch et al. 2011), where each station in the urban rail system is expanded into a series of nodes, representing the station at different time intervals. The length of the time interval τ is usually set as the minimum headway. For example, assuming train departing the terminal every two minutes, a station i in the urban rail system will be expanded to nodes $(i_1, i_2, \ldots, i_N)$, where i_1 represents station i at time 7:00-7:02 a.m.; i_2 at time interval 7:02-7:04 a.m., and so on. This fine-grained method may not be practical for real-world applications because the TS network can be extremely large. Consider an urban rail system with 100 stations and a headway of two minutes (e.g., the peak hour in the Hong Kong MTR system). For a two-hour network loading, each station will be expanded to 60 TS nodes. The total number of OD pairs in this TS network is approximately 36 million, causing great computational challenges. However, the path choice calibration problem actually does not require such a fine-grained representation. Instead, too granular information may be sensitive to observation errors in the system. Therefore, an aggregate network representation is preferred for the problem.

Let us consider a study time period T divided into N elementary time intervals of length τ (e.g., $\tau = 15$ minutes). Each time interval may include several headways, indicating a more aggregate representation. Considering a station i, we expand i into a sequence of TS nodes, denoted as $(i_1, \ldots, i_m, \ldots, i_N)$, where i_m represents station i at time interval m. The aggregate TS representation thus consists of N layers of the network with each layer representing a time interval. Let $\mathcal{S}$ be the set of all physical stations in the network, $\mathcal{N}$ be the set of all TS nodes, where $\mathcal{N} = \{i_m : \forall i \in \mathcal{S}, m = 1, \ldots, N\}$ and $|\mathcal{N}| = |\mathcal{S}| \times N$.

In the aggregate TS network, some detailed aspects (e.g., left behind) cannot be explicitly modeled. However, the tradeoff is that we can obtain a small-scale TS network, which is scalable to large networks. Moreover, the value of τ should be consistent with the granularity of information. A small τ, although allowing for modeling detailed data, may increase estimation errors caused by external and uncaptured factors such as walking speed. Consider, for example, an extreme scenario where $\tau = 1$ second. It allows us to model detailed passenger flows per second. However, passenger flows per second may be easily distorted if the walking time measurement error is greater than one second (which is common in reality). The sensitivity to external factors may introduce errors in the path choice estimation. On the contrary, a large value of τ may hide useful temporal variations, which makes the model insensitive to path choices. Therefore, considering the tradeoff of computational tractability and information granularity, $\tau = 15$ minutes is used in this study. Online Appendix C tests the impact of different values of τ.

Consider two stations i and j in an urban rail system with different routes connecting them. The route set is denoted as $\mathcal{R}_{i,j}$. Our purpose is to calculate the choice fractions of these routes for different time intervals. Using the previous network representation, the key variables are defined here:

- **OD entry flow**, $q^{i_m,j}$: Number of people with origin i and destination j entering station i during time interval m. It can be obtained from the AFC data directly. The vector of all $q^{i_m,j}$ is denoted as $\boldsymbol{q_e} = (q^{i_m,j})_{i_m \in \mathcal{N}, j \in \mathcal{S}}$.
- **OD entry-exit flow**, q^{i_m,j_n}: Number of passengers who enter station i in time interval m and exit at station j in time interval n $(m \le n)$. By definition,

$$\sum_{\{n: m \le n \le N\}} q^{i_m,j_n} = q^{i_m,j}, \quad \forall i_m \in \mathcal{N}, j \in \mathcal{S}. \tag{1}$$

The expression q^{i_m,j_n} is an output of the network loading model. For a closed urban rail system with both tap-in and tap-out records, q^{i_m,j_n} is also directly observed from the AFC data. OD entry-exit flows capture both passenger flow and journey time information at an aggregate level. Therefore, it can be used to calibrate the path choice.

- **Path choice fraction (or path share)**, $p_r^{i_m,j}$: The probability that path r is chosen in time interval m, where $r \in \mathcal{R}_{i,j}$. By definition, $0 \le p_r^{i_m,j} \le 1$ and $\sum_{r\in\mathcal{R}_{i,j}} p_r^{i_m,j} = 1$. The vector of all $p_r^{i_m,j}$ is denoted as $\boldsymbol{p} = (p_r^{i_m,j})_{i_m\in\mathcal{N},j\in\mathcal{S},r\in\mathcal{R}_{i,j}}$.
- **Path flow**, $q_r^{i_m,j_n}$: Number of passengers who enter station i in time interval m and exit at station j in time interval n using path r. $q_r^{i_m,j_n}$ is an output of the network loading process.
- **Path exit rate**, $\mu_r^{i_m,j_n}$: Number of passengers who enter station i in time interval m and exit station j in time interval n using path r divided by number of passengers who enter station i in time interval m and exit station j using path r, that is,

$$\mu_r^{i_m,j_n} = \frac{q_r^{i_m,j_n}}{\sum_{\{n':m\le n'\le N\}} q_r^{i_m,j_{n'}}}, \quad \forall i_m \in \mathcal{N}, j_n \in \mathcal{N}, r \in \mathcal{R}_{i,j}. \tag{2}$$

This variable captures the information on how many passengers exit the system at different time intervals, which, for a given path r, depends only on the train schedule and left behind. Because the schedule is known, the path exit rate can be seen as an indicator of left behind. The vector of all $\mu_r^{i_m,j_n}$ is denoted by $\boldsymbol{\mu} = (\mu_r^{i_m,j_n})_{i_m\in\mathcal{N},j_n\in\mathcal{N},r\in\mathcal{R}_{i,j}}$.

Given the previous notation, the following relationships hold.

- OD entry-exit flow equals the sum of all the path flows of the corresponding OD.

$$q^{i_m,j_n} = \sum_{r\in\mathcal{R}_{i,j}} q_r^{i_m,j_n}, \quad \forall i_m \in \mathcal{N}, j_n \in \mathcal{N} \tag{3}$$

- The path flow can be expressed as a product of the OD entry flow, the path share, and the path exit rate.

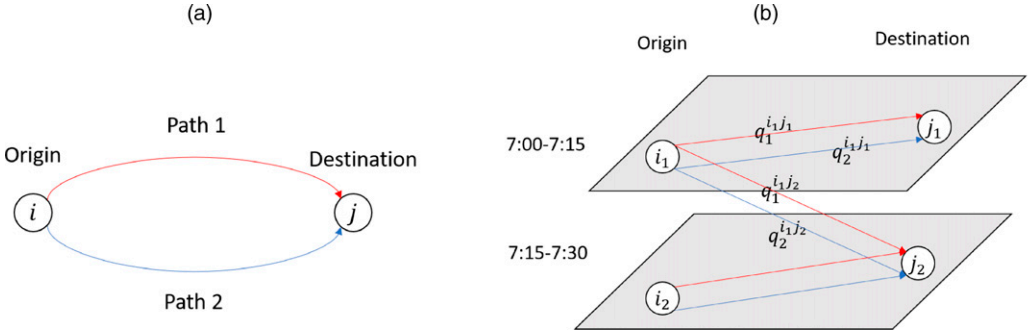
$$q_r^{i_m,j_n} = q^{i_m,j} \cdot p_r^{i_m,j} \cdot \mu_r^{i_m,j_n}, \quad \forall i_m \in \mathcal{N}, j_n \in \mathcal{N}, r \in \mathcal{R}_{i,j} \tag{4}$$

This relationship is the major procedure for the network loading. It assigns the OD demand (OD entry flows) to path flows.

We use a simple example to illustrate the network representation. Consider two stations, i and j, where i is the origin and j is the destination (Figure 1(a)). Assume there exists two different paths connecting this OD pair, that is, $\mathcal{R}_{i,j} = \{1,2\}$. The upper arrows represent path 1, and the lower arrows represent path 2. The time period of interest is from 7:00 a.m. to 7:30 a.m. and the time interval is τ = 15 minutes. The TS network representation of this example is shown in Figure 1(b). For example, i_1 in the figure represents the station i at time 7:00 a.m. to 7:15 a.m. Let us assume that the only OD entry flow is $q^{i_1,j} = 10$. The path shares are $p_1^{i_1,j} = 0.3$ and $p_2^{i_1,j} = 0.7$. Then there are 10 passengers arriving at station i during 7:00 a.m. to 7:15 a.m. Three of them use path 1 and seven use path 2. They all head to destination j, but currently, we do not know when they will arrive at the destination. Given the available information, a transit loading model can determine passengers' exit times. For illustration, let us assume that the exit times (which can be used to calculate the path exit rate) are known. For the three passengers who use path 1, two of three tap out at station j during 7:00 a.m. to 7:15 a.m. (i.e., $\mu_1^{i_1,j_1} = 2/3$) and one of three tap out at station j during 7:15 a.m. to 7:30 a.m. (i.e., $\mu_1^{i_1,j_2} = 1/3$). Then we have $q_1^{i_1,j_1} = q^{i_1,j} \cdot p_1^{i_1,j} \cdot \mu_1^{i_1,j_1} = 2$ and $q_1^{i_1,j_2} = q^{i_1,j} \cdot p_1^{i_1,j} \cdot \mu_1^{i_1,j_2} = 1$. These equations correspond to Equation (4), which assigns the OD entry flow ($q^{i_1,j}$) to the path flows ($q_1^{i_1,j_1}$ and $q_1^{i_1,j_2}$). Similarly, for the seven passengers who use path 2, assume four of seven passengers tap out at station j during 7:00 a.m. to 7:15 a.m. (i.e., $\mu_2^{i_1,j_1} = 4/7$), and three of seven passengers tap out at station j during 7:15 a.m. to 7:30 a.m. (i.e., $\mu_2^{i_1,j_2} = 3/7$). Then we have $q_2^{i_1,j_1} = 4$ and $q_2^{i_1,j_2} = 3$.

From the relationship between OD entry-exit flows and path flows (Equation (3)), we have $q^{i_1,j_1} = q_1^{i_1,j_1} + q_2^{i_1,j_1} = 6$, and $q^{i_1,j_2} = q_1^{i_1,j_2} + q_2^{i_1,j_2} = 4$. Also, the sum of the OD entry-exit flows over all exit time intervals is the OD entry flow, that is, $q^{i_1,j} = q^{i_1,j_1} + q^{i_1,j_2} = 10$.

Figure 1. (Color online) Network Representation Example



Notes. (a) Physical network. (b) Time-space hypernetwork.

3.2. Problem Formulation

3.2.1. Model Assumptions. We assume that path shares can be formulated as a C-logit model (Cascetta et al. 1996), which is an extension of the multinomial logit (MNL) model to correct for the correlation among paths due to overlapping (Prato 2009):

$$p_r^{i_m,j} = \frac{\exp(\beta_X \cdot X_{r,m} + \beta_{CF} \cdot CF_r)}{\sum_{r' \in \mathcal{R}_{i,j}} \exp(\beta_X \cdot X_{r',m} + \beta_{CF} \cdot CF_{r'})} := \frac{\exp(\beta Y_{r,m})}{\sum_{r' \in \mathcal{R}_{i,j}} \exp(\beta Y_{r',m})}, \tag{5}$$

where $X_{r,m}$ are the attributes of path r in time interval m (e.g., in-vehicle time, number of transfers, and transfer walking time). CF_r is the commonality factor of path r that measures the degree of similarity of path r with the other paths of the same OD; β_X and β_{CF} are the corresponding coefficients to be estimated; and β and $Y_{r,m}$ represent the combination of the two terms in the utility function (i.e., $\beta = [\beta_X, \beta_{CF}], Y_{r,m} = [X_{r,m}, CF_r]$). CF_r is defined as

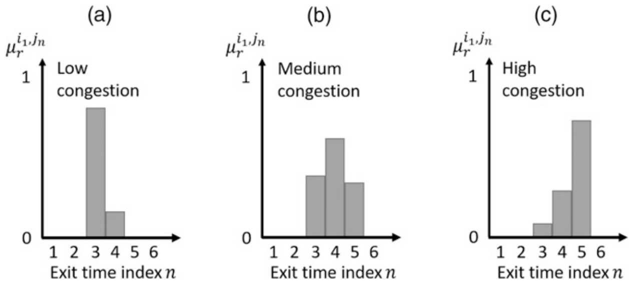
$$CF_r = \ln \sum_{r' \in \mathcal{R}_{i,j}} \left(\frac{L_{r,r'}}{L_r L_{r'}} \right)^{\gamma}, \tag{6}$$

where $L_{r,r'}$ is the number of common stations of path r and r'. L_r and $L_{r'}$ are the number of stations for path r and r', respectively; γ is a fixed positive parameter that determines the degree of overlapping impact on path utilities. In this study, $\gamma = 5$ is used following the setting in Li (2014). It is possible to use other path choice models as long as they are convex. A detailed discussion on the model extensions can be found in Online Appendix D.

It is worth noting that $Y_{r,m}$ (i.e., the path attribute vector) is known and fixed. It is assumed to represent the historical path conditions based on which passengers make their habitual choices. Different from typical transit/traffic assignment problems where path choices are estimated by assuming user equilibrium (for planning purposes), the AFC data-based estimation aims to find the actual realized path choices based on real-world observations (i.e., OD entry-exit flows). Because passengers make decisions before knowing the actual travel or waiting times, $Y_{r,m}$ should reflect passengers' historical perceptions of path attributes and should not change within the model estimation process. Therefore, although $\mu_r^{i_m,j_n}$ captures the actual path left behind and travel time information, it should *not* be included in the path choice formulation as passengers make decisions before knowing $\mu_r^{i_m,j_n}$.

In the formulation of the problem, we assume that passengers waiting on a platform board trains based on a first-in-first-board (FIFB) principle. Every train has a capacity. When a train reaches its capacity, the remaining passengers on the platform will be left behind for the next train with available capacity to board.

Figure 2. Pattern of $\mu_r^{i_m,j_n}$



Notes. (a) Low congestion. (b) Medium congestion. (c) High congestion.

These constraints are formulated as the relationship among all $\mu_r^{i_m,j_n}$, because $\mu_r^{i_m,j_n}$ represents when and how many passengers exit the system, which is a reflection of the network loading mechanism (NLM). For example, we should have $\mu_r^{i_m,j_n} = 0$ for all $(n-m)\cdot\tau$ smaller than travel time of path r, which indicates no passenger exits the system before the earliest possible time. We should also have $\boldsymbol{\mu}$ following the patterns in Figure 2 given different congestion levels. However, formulating all these constraints analytically, based on the aggregate network representation, is difficult. We thus temporally denote the constraints for $\mu_r^{i_m,j_n}$ as

$$\mu_r^{i_m,j_n} \text{ satisfies the NLM}, \quad \forall i_m \in \mathcal{N}, j_n \in \mathcal{N}, r \in \mathcal{R}_{i,j}. \tag{7}$$

The constraints will be addressed explicitly in the following sections.

3.2.2. Formulation. Because we assume that path choices can be captured by a C-logit model, the β in the C-logit model are decision variables. As mentioned before, the OD entry-exit flow (q^{i_m,j_n}) can be obtained from the transit network loading process, for which the ground truth value can also be observed from AFC data. Hence, minimizing the difference between the estimated and observed OD entry-exit flows can be the optimization objective. The reasons for using this difference as the objective function rather than individual journey times in literature are as follows. (1) The model is framed based on the aggregate TS hypernetwork. Individual-based journey times are not available under this framework. (2) Estimating individual-based journey times is difficult given many latent factors (e.g., various walking speeds, in-station activities). The model may become sensitive to parameters when matching individual-level information (Ma et al. 2019). Aggregate information (e.g., q^{i_m,j_n}) has the potential to offset some latent errors, thus providing more reliable calibrations. (3) Considering the computational cost, using aggregate information along with the aggregate TS hypernetwork facilitates the application of the model in large-scale urban rail systems, whereas individual-based models are usually applied to a small sample of AFC data (Sun et al. 2015).

If prior information about path choices was available (e.g., estimated from a prior survey), the difference between estimated β and prior β's can be incorporated into the objective function. Based on the previous discussions, the original problem is formulated as follows:

$$\min_{\beta,\boldsymbol{\mu},\boldsymbol{p}} \quad w_1 \sum_{i_m \in \mathcal{N},\, j_n \in \mathcal{N}} (q^{i_m,j_n} - \tilde{q}^{i_m,j_n})^2 + w_2 \|\beta - \tilde{\beta}\|^2 \tag{8a}$$

$$\text{s.t.} \quad q^{i_m,j_n} = \sum_{r \in \mathcal{R}_{i,j}} q_r^{i_m,j_n} \qquad \forall i_m \in \mathcal{N}, j_n \in \mathcal{N}, \tag{8b}$$

$$q_r^{i_m,j_n} = q^{i_m,j} \cdot p_r^{i_m,j} \cdot \mu_r^{i_m,j_n} \qquad \forall i_m \in \mathcal{N}, j \in \mathcal{S}, r \in \mathcal{R}_{i,j}, \tag{8c}$$

$$p_r^{i_m,j} = \frac{\exp(\beta Y_{r,m})}{\sum_{r' \in \mathcal{R}_{i,j}} \exp(\beta Y_{r',m})} \qquad \forall i_m \in \mathcal{N}, j \in \mathcal{S}, r \in \mathcal{R}_{i,j}, \tag{8d}$$

$$\mu_r^{i_m,j_n} \text{ satisfies the NLM} \qquad \forall i_m \in \mathcal{N}, j \in \mathcal{S}, r \in \mathcal{R}_{i,j}, \tag{8e}$$

$$\sum_{r \in \mathcal{R}_{i,j}} p_r^{i_m,j} = 1 \qquad \forall i_m \in \mathcal{N}, j \in \mathcal{S}, \tag{8f}$$

$$0 \le p_r^{i_m,j} \le 1 \qquad \forall i_m \in \mathcal{N}, j \in \mathcal{S}, r \in \mathcal{R}_{i,j}, \tag{8g}$$

$$q_r^{i_m,j_n} \ge 0 \qquad \forall i_m \in \mathcal{N}, j \in \mathcal{S}, r \in \mathcal{R}_{i,j}, \tag{8h}$$

where $\tilde{q}^{i_m,j_n}$ is the observed OD entry-exit flow; $\tilde{\beta}$ is prior estimates of β. w_1 and w_2 are the corresponding weights. In the case study section, we assume no prior knowledge is available (i.e., $w_2 = 0$). Constraints (8b) and (8c) are the relationships described in Section 3.1. Constraints (8d) and (8e) represent the assumptions we made in Section 3.2.1. Constraints (8f), (8g), and (8h) are given by definition.

The original problem is hard to solve due to the following constraints: First, Constraints (8c) and (8d) are nonlinear equality constraints because both $\mu_r^{i_m,j_n}$ and $p_r^{i_m,j}$ are decision variables. This makes the original problem a nonconvex optimization problem. Second, Constraint (8e) is nonanalytical because NLM constraints cannot be formulated analytically in terms of $\mu_r^{i_m,j_n}$. Therefore, the original problem is intractable. In the following sections, we propose a decomposition approach to deal with these constraints and approximately solve the original problem.

3.3. Problem Decomposition

Although Constraints (8e) cannot be formulated analytically, the corresponding $\mu_r^{i_m,j_n}$ values can be obtained from the output of a network loading process. Therefore, we decompose the original problem into two subproblems as follows.

- Subproblem 1:

$$\begin{aligned} \min_{\beta,\boldsymbol{p}} \quad & w_1 \sum_{i_m \in \mathcal{N},\, j_n \in \mathcal{N}} (q^{i_m,j_n} - \tilde{q}^{i_m,j_n})^2 + w_2 \|\beta - \tilde{\beta}\|^2 \\ \text{s.t.} \quad & \text{Constraints (8b)–(8d)}, \\ & \text{Constraints (8f)–(8h)}. \end{aligned} \tag{9}$$

- Subproblem 2:

$$\boldsymbol{\mu} = \text{Network Loading } (\beta, \boldsymbol{q_e}). \tag{10}$$

This decomposition shares the same idea as the Expectation–Maximization algorithm or alternating optimization. The idea is that when fixing some variables, the original problem will be easier to solve.

Subproblem 2 is a network loading model, which takes the route choice parameter β and OD entry demand $\boldsymbol{q_e}$ as inputs and outputs the path exit rate $\boldsymbol{\mu}$. In this study, we use an event-based network loading model proposed by Mo et al. (2020). The model is well calibrated and validated with real-world data for the case study. Two events are considered: train arrival events, in which passengers who need to transfer or exit the station are offloaded; and train departure events, in which passengers are loaded into the train based on the FIFB principle. All events are processed sequentially according to their occurrence time. This network loading model shares the same NLM and model assumptions as described before. Therefore, the estimated $\boldsymbol{\mu}$ from the model satisfies the NLM constraints.

Subproblem 1 is a variation of the original problem (Equation (8)) with the nonanalytical Constraint (8e) removed because $\boldsymbol{\mu}$ is treated as constants in Subproblem 1. Moreover, Constraint (8c) is now linear to $p_r^{i_m,j}$. However, the problem is still intractable because of the nonlinear Constraint (8d), which we refer to as the *logit constraint*. In the following sections, we will show how we linearize Subproblem 1 and solve it as a quadratic programming problem.

3.4. Linearization for Subproblem 1

Addressing the logit constraints is difficult. Davis, Gallego, and Topaloglu (2013) and Atasoy et al. (2015) showed that when the logit structure is in an objective function, and utilities are constants, but choice sets are unknown, this assortment planning–type problem can be reformulated as linear programming. However, for our problem, the logit structure is a constraint and β in the utility function is unknown. To the best of our knowledge, there is no equivalent transformation of this logit constraint to a tractable form. In this study, we propose two procedures to approximately linearize the logit constraints in Subproblem 1.

3.4.1. Approximate Linear Constraints (ALC).

The logit constraint reflects the relationship between β and $p_r^{i_m,j}$. Because dealing with the nonlinear constraints directly is difficult, we first replace the decision variables β with $p_r^{i_m,j}$

and remove Constraint (8d). Then Subproblem 1 becomes a simple quadratic programming problem because all constraints have become linear. However, as the degrees of freedom for $p_r^{i_m,j}$ are much larger than β, directly replacing decision variables will cause problems of overfitting. Introducing additional constraints on $p_r^{i_m,j}$ can reduce the feasible solution space and create a tighter problem.

In this study, we propose a Monte-Carlo sampling method to construct a series of linear constraints for $p_r^{i_m,j}$. The basic idea is that, in urban rail networks, there exists some i_m, j, r and $i'_{m'}, j', r'$ such that $p_r^{i_m,j} = p_{r'}^{i'_{m'},j'}$. This property is based on the argument that if the paths for two different OD pairs share the same transfer patterns (i.e., transfer stations and the number of transfers), the corresponding path choice fractions may be the same under logit constraints.

A simple example is shown to illustrate the nature of this constraint. Consider the OD pairs 1-5 and 2-5 in Figure 3. There are two paths for each OD pair. Path 1 has a transfer at station 4 and path 2 at station 3. The path choice fractions are denoted as $p_1^{1,5}, p_2^{1,5}, p_1^{2,5}, p_2^{2,5}$, respectively. For simplicity of notation, we ignore the time index. We further assume that there are four path attributes affecting passengers' path choices: in-vehicle time, the number of transfers and transfer walking time, and the commonality factor. Then we can show that, under logit constraints, $p_1^{1,5} = p_1^{2,5}$ and $p_2^{1,5} = p_2^{2,5}$. The proof is shown later.

Denote the utility (in the C-logit model) for path r of OD pair i and j as $V_r^{i,j}$. Because path 1 of OD 1-5 and path 1 of OD 2-5 share the same transfer patterns, the number of transfers and transfer walking time for them are the same. The commonality factors for these two paths are also the same (see Equation (6)). Therefore, the difference in utilities for path 1 of the two different OD pairs only contains in-vehicle time. Let the in-vehicle time for path r of OD pair (i, j) be $tt_r^{i,j}$, the in-vehicle time for link (i, j) be $tt^{i,j}$, and the coefficients of in-vehicle time be β_{tt}. Then, we have

$$V_1^{1,5} - V_1^{2,5} = \beta_{tt} \cdot (tt_1^{1,5} - tt_1^{2,5}) = \beta_{tt} \cdot tt^{1,2}. \tag{11}$$

Similarly, for path 2 of OD 1-5 and OD 2-5, we have

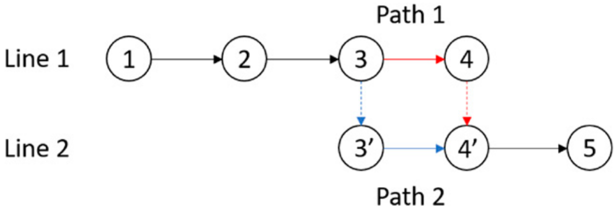
$$V_2^{1,5} - V_2^{2,5} = \beta_{tt} \cdot (tt_2^{1,5} - tt_2^{2,5}) = \beta_{tt} \cdot tt^{1,2}. \tag{12}$$

According to the logit constraint,

$$\begin{aligned} p_1^{1,5} &= \frac{1}{1 + \exp(V_2^{1,5} - V_1^{1,5})} \\ &= \frac{1}{1 + \exp((V_2^{2,5} + \beta_{tt} \cdot tt^{1,2}) - (V_1^{2,5} + \beta_{tt} \cdot tt^{1,2}))} \\ &= \frac{1}{1 + \exp(V_2^{2,5} - V_1^{2,5})} = p_1^{2,5}. \end{aligned} \tag{13}$$

Similarly, for path 2, $p_2^{1,5} = p_2^{2,5}$.

Figure 3. (Color online) Network Example for Approximated Linear Constraints



This example network represents a subcomponent of many real-world urban rail networks. Therefore, this property also holds in for many OD pairs in reality.

Besides equality constraints, there are also inequality constraints that can be introduced to further limit the feasible space. For example, because all cost coefficients (e.g., in-vehicle time, transfer times) should be negative, if there is a path that has smaller costs than other paths for the same OD pair, it should always have a higher share regardless of the value of β. Therefore, we can construct linear constraints of the form $p_r^{i_m,j} \geq p_{r'}^{i_m,j}$ to capture this information.

To automatically extract all these linear constraints in the system, we propose a Monte-Carlo sampling method. We first define a reasonable range for all β (i.e., $\beta \in [L_\beta, U_\beta]$) based on prior knowledge (e.g., survey results from previous years), where L_β (U_β) is the vector of lower (upper) bounds for β. It is worth noting that the selection of L_β and U_β has a limited impact on the construction of ALC. The equality constraints are independent of the value of β (as shown in Equations (11)–(13)); L_β and U_β only affect the construction of inequality constraints, and from our numerical tests, the impact is very small. Generally, we only need to set the cost coefficients to be negative (i.e., $L_\beta = -\infty$ and $U_\beta = 0$).

The detailed ALC construction steps are shown in Algorithm 1. The maximum number of sampling points is S. The choice of S is a tradeoff between computational efficiency and constraint accuracy. Larger S can help avoid including erroneous constraints. We sample β uniformly from $[L_\beta, U_\beta]$ because (1) it reflects no prior distribution knowledge of β is used, and (2) this facilitates getting a wide range of β covering different scenarios so that the constructed linear constraints are valid.

Algorithm 1 (Monte Carlo–Based ALC Construction)

```
Initialize $s = 0$
**while** $s < S$ **do**
    $s = s + 1$
    Sample $\beta$ from the uniform distribution $U(L_\beta, U_\beta)$, denoted as $\beta^{(s)}$.
    Calculate the path choice fractions for all paths based on $\beta^{(s)}$, denote them as $p_r^{i_m,j(s)}$
Initialize $\Omega_{\text{Eq}} = \emptyset$, $\Omega_{\text{Ineq}} = \emptyset$
```

```
for all $i_m, j, r$ in path sets do
    for all $i_m', j', r'$ in path sets do
        if $p_r^{i_m,j^{(s)}} = p_{r'}^{i_m',j'^{(s)}}$ for all $s = 1, \dots, S$ then
            $\Omega_{\text{Eq}} = \Omega_{\text{Eq}} \cup \{(i_m, j, r, i_m', j', r')\}$.
for all $i_m, j$ in OD pairs sets do
    for all $r \in \mathcal{R}_{i,j}$ do
        for all $r' \in \mathcal{R}_{i,j}$ do
            if $p_r^{i_m,j^{(s)}} \geq p_{r'}^{i_m,j^{(s)}}$ for all $s = 1, \dots, S$ then
                $\Omega_{\text{Ineq}} = \Omega_{\text{Ineq}} \cup \{(i_m, j, r, r')\}$.
return $\Omega_{\text{Eq}}$, $\Omega_{\text{Ineq}}$
```

These linear constraints partially capture the effect of the logit constraints and retain the model tractability. We denote all the constructed linear constraints for $p_r^{i_m,j}$ as

$$p_r^{i_m,j} = p_{r'}^{i_m',j'} \quad \forall (i_m, j, r, i_m', j', r') \in \Omega_{\text{Eq}}, \tag{14}$$

$$p_r^{i_m,j} \geq p_{r'}^{i_m,j} \quad \forall (i_m, j, r, r') \in \Omega_{\text{Ineq}}. \tag{15}$$

Then Subproblem 1 can be reformulated as

$$\begin{aligned} \min_{\boldsymbol{p}} \quad & w_1 \sum_{i_m \in \mathcal{N}, j_n \in \mathcal{N}} (q^{i_m,j_n} - \tilde{q}^{i_m,j_n})^2 \\ & + w_2 \sum_{i_m \in \mathcal{N}, j \in \mathcal{S}} \sum_{r \in \mathcal{R}_{i,j}} (p_r^{i_m,j} - \tilde{p}_r^{i_m,j})^2 \\ \text{s.t.} \quad & \text{Constraints (8b)–(8c),} \\ & \text{Constraints (14)–(15),} \\ & \text{Constraints (8f)–(8h).} \end{aligned} \tag{16}$$

The decision variables in the new formulation are $\boldsymbol{p}$ instead of β; $\tilde{p}_r^{i_m,j}$ are path shares derived from $\tilde{\beta}$ to capture the prior knowledge. Equation (16) is a quadratic program because all constraints are linear and can be solved efficiently. Based on the numerical results in the case study, after adding the ALC, the total degrees of freedom decrease by 40%, which demonstrates a narrower feasible space. However, we still need to go one step further to make all estimated path shares satisfy the actual logit constraints.

3.4.2. Logit Correction.

The estimated $\boldsymbol{p}$ from Equation (16) has two issues. The first is possible overfitting due to the high degrees of freedom. The second is that some path shares cannot be identified due to few or no observed OD entry-exit flows; for example, if there are no observed passengers for an OD pair (i, j) in time interval m and historical information is not available. The variable $p_r^{i_m,j}$ can take any values and does not affect the objective function. Hence, its value cannot be estimated. Both of these problems can be attributed to the same source: The estimated $p_r^{i_m,j}$ violates the original logit constraints (they only satisfy the ALC of logit).

To address this problem, we use the estimated $p_r^{i_m,j}$ from Equation (16) (called *rough* path shares hereafter) to obtain β and then use the β to generate new path shares. This procedure is referred to as *logit correction*. Path shares after the logit correction will naturally satisfy the logit constraints by definition. However, not all rough path shares are equally reliable. Because more observed passengers provide more information for the path shares estimation, the reliability of the estimated $p_r^{i_m,j}$ ($\forall r \in \mathcal{R}_{i,j}$) can be measured by the corresponding OD entry flow $q^{i_m,j}$. Therefore, we formulate the logit correction problem as follows, which can be seen as a weighted fractional logit model (Papke and Wooldridge 1996).

$$\max_{\beta} \sum_{i_m \in \mathcal{N}, j \in \mathcal{S}} q^{i_m,j} \sum_{r \in \mathcal{R}_{i,j}} p_r^{i_m,j} \cdot \log \frac{\exp(\beta Y_{r,m})}{\sum_{r' \in \mathcal{R}_{i,j}} \exp(\beta Y_{r',m})} \tag{17}$$

In Equation (17), $p_r^{i_m,j}$ and $q^{i_m,j}$ are known. The objective function is concave because $\log \frac{\exp(\beta Y_{r,m})}{\sum_{r' \in \mathcal{R}_{i,j}} \exp(\beta Y_{r',m})}$ is concave in terms of β. Hence, it is a convex optimization problem (maximizing a concave function) without constraints and can be solved efficiently. The term $q^{i_m,j}$ is the weight for corresponding path shares ($p_r^{i_m,j}$, $\forall r \in \mathcal{R}_{i,j}$). Equation (17) has the advantage that if there are no passengers observed for a specific OD pair ($q^{i_m,j} = 0$), the corresponding term will have zero weight and will not contribute to the objective function. Using the estimated β, new values of $\boldsymbol{p}$ that satisfy the logit constraints exactly can be calculated.

Besides the weighted fractional logit model, another way to look at Equation (17) is to treat it as a maximum likelihood estimation. For a given i_m, j, there are $q^{i_m,j} \cdot p_r^{i_m,j}$ passengers choosing path r. The probability of choosing path r is $\frac{\exp(\beta Y_{r,m})}{\sum_{r' \in \mathcal{R}_{i,j}} \exp(\beta Y_{r',m})}$. Therefore, the likelihood function can be formulated as

$$\mathcal{L} = \prod_{i_m \in \mathcal{N}, j \in \mathcal{S}} \prod_{r \in \mathcal{R}_{i,j}} \left(\frac{\exp(\beta Y_{r,m})}{\sum_{r' \in \mathcal{R}_{i,j}} \exp(\beta Y_{r',m})} \right)^{q^{i_m,j} \cdot p_r^{i_m,j}}. \tag{18}$$

Taking the logarithm of $\mathcal{L}$ and maximizing the log-likelihood lead to Equation (17). The empirical results on the effect of linearization of Subproblem 1 (i.e., ALC and logit correction) can be found in Online Appendix E.

3.5. Solution Algorithm

Thus far, we have formulated three subproblems to approximate the solution of the original problem. These subproblems can be summarized by Equations (19)–(21). In Subproblem 1a, given $\mu_r^{i_m,j_n}$, we estimate the rough path shares by solving a quadratic programming problem. In Subproblem 1b, given the rough path shares, we estimate the corresponding β through a weighted fractional logit model formulation. In Subproblem 2, given β, we load passengers to the network and return the $\mu_r^{i_m,j_n}$ values that satisfy the NLM constraints.

- Subproblem 1a:

$$[\mathrm{SP1A}(\boldsymbol{\mu})] \quad \min_{\boldsymbol{p}} \quad w_1 \sum_{i_m\in\mathcal{N}, j_n\in\mathcal{N}} (q^{i_m,j_n} - \tilde{q}^{i_m,j_n})^2 + w_2 \sum_{i_m\in\mathcal{N}, j\in\mathcal{S}} \sum_{r\in\mathcal{R}_{i,j}} (p_r^{i_m,j} - \tilde{p}_r^{i_m,j})^2 \tag{19a}$$

$$\text{s.t.} \quad \text{Constraints (8b)–(8c)}, \tag{19b}$$

$$\text{Constraints (14)–(15)}, \tag{19c}$$

$$\text{Constraints (8f)–(8h)}. \tag{19d}$$

- Subproblem 1b:

$$[\mathrm{SP1B}(\boldsymbol{p})]$$

$$\max_{\beta} \sum_{i_m\in\mathcal{N}, j\in\mathcal{S}} q^{i_m,j} \sum_{r\in\mathcal{R}_{i,j}} p_r^{i_m,j} \cdot \log \frac{\exp(\beta Y_{r,m})}{\sum_{r'\in\mathcal{R}_{i,j}} \exp(\beta Y_{r',m})}. \tag{20}$$

- Subproblem 2:

$$[\mathrm{SP2}(\beta)] \quad \boldsymbol{\mu} = \text{Network Loading}\,(\beta, \boldsymbol{q}_e, \theta). \tag{21}$$

We solve these three subproblems iteratively and approximate the solution for the original problem. This process is summarized in Figure 4. The terms $\boldsymbol{\mu}$ and β are updated in each iteration, which suggests that the interactions between path choice and left behind are captured. It is worth noting that this process is equivalent to finding a fixed point of the following problem:

$$\beta = \mathrm{SP1B} \circ \mathrm{SP1A} \circ \mathrm{SP2}(\beta), \tag{22}$$

where SP2 is the solution function of Subproblem 2, that is, $\boldsymbol{\mu} = \mathrm{SP2}(\beta)$; SP1A is the solution function of Subproblem 1a, that is, $\boldsymbol{p} = \mathrm{SP1A}(\boldsymbol{\mu})$; SP1B is the solution function of subproblem 1b, that is, $\beta = \mathrm{SP1B}(\boldsymbol{p})$; and "$\circ$" is the sign of function composition, that is, $f \circ g(x) = f(g(x))$. The existence and uniqueness of the solution in Equation (22) and its relationship to the original problem in Equation (8) are important questions.

Lemma 1. *If $\boldsymbol{p}$ satisfies the logit constraints in terms of β^*, that is, $p_r^{i_m,j} = \frac{\exp(\beta^* Y_{r,m})}{\sum_{r'\in\mathcal{R}_{i,j}} \exp(\beta^* Y_{r',m})}$ for all $i_m \in \mathcal{N}, j \in \mathcal{S}$, $r \in \mathcal{R}_{i,j}$, then β^* is the solution of Subproblem 1b with respect to rough path share $\boldsymbol{p}$ (i.e., $\beta^* = \mathrm{SP1B}(\boldsymbol{p})$)*

Proof. Define $h_r^{i_m,j}(\beta) := \frac{\exp(\beta Y_{r,m})}{\sum_{r'\in\mathcal{R}_{i,j}} \exp(\beta Y_{r',m})}$. Then Equation (20) can be rewritten as

$$\max_{\beta} \sum_{i_m\in\mathcal{N}, j\in\mathcal{S}} q^{i_m,j} \sum_{r\in\mathcal{R}_{i,j}} p_r^{i_m,j} \cdot \log h_r^{i_m,j}(\beta), \tag{23}$$

which has the form of a cross-entropy function. The maximum is reached when $p_r^{i_m,j} = h_r^{i_m,j}(\beta)$, $i_m \in \mathcal{N}$, $j \in \mathcal{S}, r \in \mathcal{R}_{i,j}$. From the known condition of the lemma, we know that $p_r^{i_m,j}$ satisfies the logit constraints in terms of β^* $\left(\text{i.e., } p_r^{i_m,j} = \frac{\exp(\beta^* Y_{r,m})}{\sum_{r'\in\mathcal{R}_{i,j}} \exp(\beta^* Y_{r',m})}\right)$, feeding β^* into $h_r^{i_m,j}(\beta)$ gives the desired condition ($p_r^{i_m,j} = h_r^{i_m,j}(\beta^*)$). Thus, β^* is the optimal solution of Subproblem 1b (i.e., $\beta^* = \mathrm{SP1B}(\boldsymbol{p})$).

Because $\log \frac{\exp(\beta Y_{r,m})}{\sum_{r'\in\mathcal{R}_{i,j}} \exp(\beta Y_{r',m})}$ is strictly concave in terms of β, the solution of Subproblem 1b is unique. □

Proposition 1 discusses the question of existence of a fixed point in Equation (22).

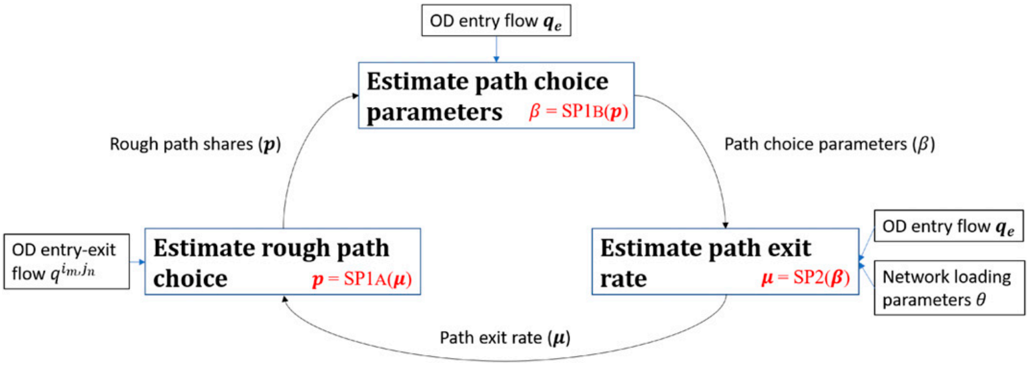
Proposition 1. *The optimal solution β^* for the original problem (Equation (8)) is a fixed point for Equation (22) if the following condition holds: the optimal objective function for the original problem Equation (8) is zero.*

Proof. Denote $\boldsymbol{\mu}^* := \mathrm{SP2}(\beta^*)$. Define $\boldsymbol{p}^*$ such that its element $p_r^{i_m,j*} = \frac{\exp(\beta^* Y_{r,m})}{\sum_{r'\in\mathcal{R}_{i,j}} \exp(\beta^* Y_{r',m})}$ for all $i_m \in \mathcal{N}, j \in \mathcal{S}$, $r \in \mathcal{R}_{i,j}$. We claim that $\boldsymbol{p}^* = \mathrm{SP1A}(\boldsymbol{\mu}^*)$. The proof is shown later.

By definition, $\boldsymbol{p}^*$ satisfies the logit constraints with respect to β^*. The term $\boldsymbol{\mu}^*$ satisfies the NLM constraints with respect to β^* because it is the output from the network loading process (i.e., SP2($\cdot$)). Therefore, $(\beta^*, \boldsymbol{\mu}^*, \boldsymbol{p}^*)$ is the optimal solution for the original problem.

If $\boldsymbol{\mu}^*$ is used in Subproblem 1a, the optimal objective function of Subproblem 1a should be less than or equal to that of the original problem because $\boldsymbol{p}$ has a larger feasible space in Subproblem 1a than in the

Figure 4. (Color online) Summary of Solution Procedure


OD entry flow $\boldsymbol{q_e}$
Estimate path choice parameters $\beta = \mathrm{SP1B}(\boldsymbol{p})$
Rough path shares ($\boldsymbol{p}$)
Path choice parameters (β)
OD entry-exit flow q^{i_m,j_n}
Estimate rough path choice $\boldsymbol{p} = \mathrm{SP1A}(\boldsymbol{\mu})$
Estimate path exit rate $\boldsymbol{\mu} = \mathrm{SP2}(\boldsymbol{\beta})$
OD entry flow $\boldsymbol{q_e}$
Network loading parameters θ
Path exit rate ($\boldsymbol{\mu}$)

original problem (because we use ALC to approximate the logit constraints). However, given the condition that the optimal objective function of the original problem is zero (cannot be decreased), the optimal objective function of Subproblem 1a is zero as well. Because the objective function for these two problems are the same, it follows that $\boldsymbol{p}^*$ is the optimal solution for Subproblem 1a (i.e., $\boldsymbol{p}^* = \text{SP1A}(\boldsymbol{\mu}^*)$).

By definition, $\boldsymbol{p}^*$ satisfies the logit constraints in terms of β^*. According to Lemma 1, $\beta^* = \text{SP1B}(\boldsymbol{p}^*)$. This leads to $\beta^* = \text{SP1B} \circ \text{SP1A} \circ \text{SP2}(\beta^*)$. □

Remark 1. Proposition 1 proves the existence of a fixed point for Equation (22), which is exactly the solution of the original problem. However, it only holds under the condition that the path choice behavior is perfectly captured by the behavior model (so that the optimal objective function is zero). This may not be true in reality. Nevertheless, even if the decomposed method may give solutions not exactly the same as the original problem, the proof under the specific condition illustrates the reasonableness of the approach.

It is worth noting that the nonconvexity of the original problem does not affect the conclusion of Proposition 1 because the condition of zero optimal objective function implies the global optimal. The proof of Proposition 1 does not require the convexity of the original problem. As long as β^* is a global optimal point of the original problem, it is a fixed point of Equation (22).

When the optimal objective function of the original problem (Equation (8)) is greater than zero, its optimal solution may not be the fixed point of Equation (22). Under this condition, Proposition 2 discusses the upper bounds of the difference in terms of the estimated β between the original problem and the decomposed method. For simplicity, we assume $w_2 = 0$ for the discussion, which corresponds to the setting in our case study.

Proposition 2. *Let $(\beta^*, \boldsymbol{\mu}^*, \boldsymbol{p}^*)$ be the optimal solution for the original problem. Define $\beta^{\text{D}} = \text{SP1B} \circ \text{SP1A} \circ \text{SP2}(\beta^*)$, then $\|\beta^{\text{D}} - \beta^*\|$ is bounded from above. The upper bound is illustrated in the proof.*

Proof. Because $(\beta^*, \boldsymbol{\mu}^*, \boldsymbol{p}^*)$ are the optimal solution to the original problem, by definition, $\boldsymbol{\mu}^* = \text{SP2}(\beta^*)$ and $p_r^{i_m,j*} = \frac{\exp(\beta^* \Upsilon_{r,m})}{\sum_{r' \in \mathcal{R}_{i,j}} \exp(\beta^* \Upsilon_{r',m})}$ for all $i_m \in \mathcal{N}, j \in \mathcal{S}, r \in \mathcal{R}_{i,j}$. According to Lemma 1, $\beta^* = \text{SP1B}(\boldsymbol{p}^*)$.

Let $\boldsymbol{p}^D = \text{SP1A}(\boldsymbol{\mu}^*)$. We claim that there exists an upper bound $U(\Omega_{\text{Eq}}, \Omega_{\text{Ineq}}) \ge 0$ such that $\|\boldsymbol{p}^D - \boldsymbol{p}^*\| \le U(\Omega_{\text{Eq}}, \Omega_{\text{Ineq}}) \le \mathbf{1}_{|\boldsymbol{p}|}$. The expression $U(\Omega_{\text{Eq}}, \Omega_{\text{Ineq}})$ will decrease if we construct more effective linear cuts in Ω_{Eq} and Ω_{Ineq}. The vector $\mathbf{1}_{|\boldsymbol{p}|}$ is a vector of all elements 1 with the same dimension as $\boldsymbol{p}$. The proof is as follows.

Define

$$Z(\boldsymbol{p}; \boldsymbol{\mu}, \boldsymbol{q}^e, \tilde{\boldsymbol{q}}) = \sum_{i_m \in \mathcal{N}, j_n \in \mathcal{N}} \left(\sum_{r \in \mathcal{R}_{i,j}} q^{i_m,j} \cdot p_r^{i_m,j} \cdot \mu_r^{i_m,j_n} - \tilde{q}^{i_m,j_n} \right)^2, \tag{24}$$

$$\mathcal{P}_{1\text{A}}(\Omega_{\text{Eq}}, \Omega_{\text{Ineq}}) = \{0 \le \boldsymbol{p} \le 1 : \text{Eqs. 8f, 14, 15}\}, \tag{25}$$

$$\mathcal{P}_{OP} = \{0 \le \boldsymbol{p} \le 1 : \text{Eq.8f}, \exists \beta \in \mathbb{R}^{|\beta|} \text{ s.t. Eq. 8d is satisfied}\}, \tag{26}$$

where $\tilde{\boldsymbol{q}} = (\tilde{q}^{i_m,j_n})_{i_m \in \mathcal{N}, j_n \in \mathcal{N}}$. By definition, $\mathcal{P}_{1\text{A}} \subseteq \mathcal{P}_{OP}$. Then $\max_{\boldsymbol{p} \in \mathcal{P}_{1\text{a}}} Z(\boldsymbol{p}; \boldsymbol{\mu}^*, \boldsymbol{q}^e, \tilde{\boldsymbol{q}})$ is Subproblem 1a and $\max_{\boldsymbol{p} \in \mathcal{P}_{OP}} Z(\boldsymbol{p}; \boldsymbol{\mu}^*, \boldsymbol{q}^e, \tilde{\boldsymbol{q}})$ is the original problem (when $w_2 = 0$). Hence, Subproblem 1a and the original problem have the same objective function. However, Subproblem 1a has a larger feasible region because we use ALC to approximate the logit constraints. Therefore, the upper bound of $\|\boldsymbol{p}^D - \boldsymbol{p}^*\|$ can be expressed as the maximum difference between points under optimal conditions in two feasible regions:

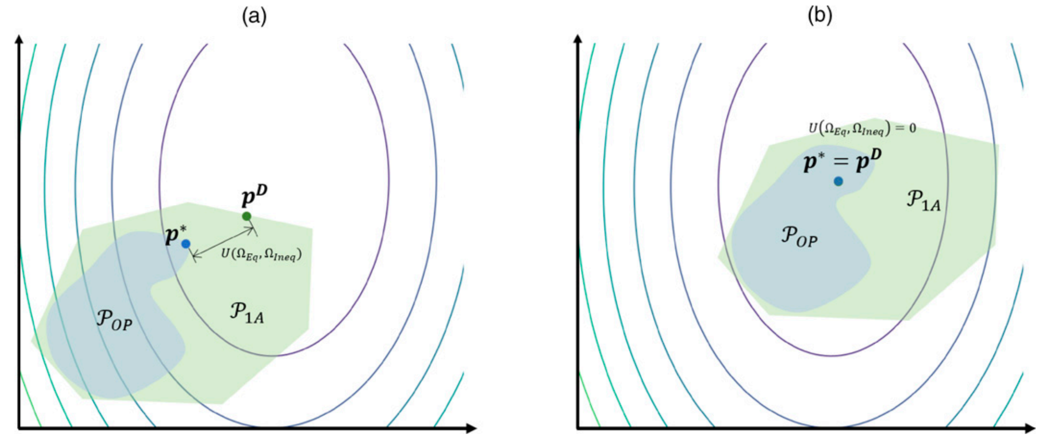
$$\begin{aligned} U(\Omega_{\text{Eq}}, \Omega_{\text{Ineq}}) = \max_{\boldsymbol{\mu}^*, \boldsymbol{q}^e, \tilde{\boldsymbol{q}}} \{ \|\boldsymbol{p}_1 - \boldsymbol{p}_2\| : & \\ \boldsymbol{p}_1 \in \arg\max_{\boldsymbol{p} \in \mathcal{P}_{1\text{A}}} Z(\boldsymbol{p}; \boldsymbol{\mu}^*, \boldsymbol{q}^e, \tilde{\boldsymbol{q}}), & \\ \boldsymbol{p}_2 \in \arg\max_{\boldsymbol{p} \in \mathcal{P}_{OP}} Z(\boldsymbol{p}; \boldsymbol{\mu}^*, \boldsymbol{q}^e, \tilde{\boldsymbol{q}}) \}. & \end{aligned} \tag{27}$$

The term $U(\Omega_{\text{Eq}}, \Omega_{\text{Ineq}})$ is bounded because both $\mathcal{P}_{1\text{A}}$ and $\mathcal{P}_{OP}$ are bounded. A trivial upper bound for $U(\Omega_{\text{Eq}}, \Omega_{\text{Ineq}})$ is $\mathbf{1}_{|\boldsymbol{p}|}$. An illustration for $U(\Omega_{\text{Eq}}, \Omega_{\text{Ineq}})$ is shown in Figure 5. Figure 5(a) shows that when the optimal objective function value of the original formulation is greater than zero, $\boldsymbol{p}^D$ and $\boldsymbol{p}^*$ may be different. Figure 5(b) shows that when the optimal objective function value of the original formulation equals zero, $\boldsymbol{p}^D = \boldsymbol{p}^*$ (reduced to Proposition 1). If we construct more ALCs, we could have a tighter feasible region for Subproblem 1a and thus a smaller $U(\Omega_{\text{Eq}}, \Omega_{\text{Ineq}})$.

Given the upper bound of $\|\boldsymbol{p}^D - \boldsymbol{p}^*\|$, we now prove the upper bound for $\|\beta^{\text{D}} - \beta^*\|$. The proof follows the maximum theorem (Ok 2011) and is illustrated later.

The maximum theorem states that if the elements of an optimization problem are sufficiently continuous, then some of that continuity is preserved in the solutions. The goal is to prove that $\text{SP1B}(\boldsymbol{p})$ (the maximizer function) is continuous in terms of $\boldsymbol{p}$. Specifically, for Subproblem 1b, because it has no constraints, we know that the correspondence (i.e., set-valued function) that maps $\boldsymbol{p}$ to the feasible region of Subproblem 1b (which is $\mathbb{R}^{|\beta|}$) is continuous. Also, we know that the objective function of Subproblem 1b is strictly concave, according to the maximum theorem, $\text{SP1B}(\boldsymbol{p})$, is continuous. We further observe that because $\boldsymbol{p}$ is in a compact set, $\text{SP1B}(\boldsymbol{p})$ is also uniformly continuous (by the Heine-Cantor theorem; Cadenas Aldana 2007).

Figure 5. (Color online) Upper Bound of $\|\boldsymbol{p}^D - \boldsymbol{p}^*\|$



Notes. (a) $Z(\boldsymbol{p}^*; \boldsymbol{\mu}^*, \boldsymbol{q}^e, \tilde{\boldsymbol{q}}) > 0$. (b) $Z(\boldsymbol{p}^*; \boldsymbol{\mu}^*, \boldsymbol{q}^e, \tilde{\boldsymbol{q}}) = 0$.

According to Vanderbei (1991), uniform continuity is almost Lipschitz continuity, leading to

$$\|\beta^D - \beta^*\| = \|\text{SP1B}(\boldsymbol{p}^D) - \text{SP1B}(\boldsymbol{p}^*)\| \le L \cdot \|\boldsymbol{p}^D - \boldsymbol{p}^*\| + \epsilon \le L \cdot U(\Omega_{\text{Eq}}, \Omega_{\text{Ineq}}) + \epsilon, \tag{28}$$

where L and ϵ are two parameters for the "almost Lipschitz continuity" (more details in Vanderbei 1991). □

Remark 2. Proposition 2 shows that when the optimal objective function of the original problem is not zero, the optimal solution of the original problem β^* may not be the fixed point of the decomposed method. However, because $\|\beta^D - \beta^*\|$ are bounded above, β^* is essentially an "almost fixed point" (Afif Ben Amar 2022). Define $\Gamma := L \cdot U(\Omega_{\text{Eq}}, \ \Omega_{\text{Ineq}}) + \epsilon$. Then β^* is a Γ-fixed point of SP1B ∘ SP1A∘ SP2.

Remark 3. Denote the fixed point (if it exists) of SP1B ∘ SP1A ∘ SP2 as $\tilde{\beta}^*$ such that $\tilde{\beta}^* = \text{SP1B} \circ \text{SP1A} \circ \text{SP2}(\tilde{\beta}^*)$. One may be interested in the difference between $\tilde{\beta}^*$ and β^* (i.e., $\|\tilde{\beta}^* - \beta^*\|$). In general, answering this question is hard. This is because, although β^* is a Γ-fixed point, it is not necessarily close to the actual fixed point $\tilde{\beta}^*$ (i.e., weak approximation). Finding a point that is close to $\tilde{\beta}^*$ (i.e., strong approximation) is not computationally feasible for general functions because it requires anticipating the limit of a sequence from a finite amount of data (Scarf 1967).

Proposition 1 discusses the existence of fixed points under the condition that the optimal objective function of the original problem is zero. For a more general situation, we show the existence of fixed points in Proposition 3.

Proposition 3. *There is a fixed point for* SP1B ∘ SP1A ∘ SP2$(\cdot)$ *if the following conditions hold:*

- *There are closed boundaries for β across the estimation process (i.e., $-\infty < L_\beta \le \beta \le U_\beta < +\infty$)*
- *The composed function* SP1B ∘ SP1A ∘ SP2$(\cdot)$ *is continuous*

Proof. Because $L_\beta \le \beta \le U_\beta$ is convex and compact, the proposition directly follows Brouwer's fixed-point theorem (Afif Ben Amar 2022). □

Remark 4. Proposition 3 provides the conditions for the existence of a fixed point using Brouwer's fixed-point theorem. However, the continuity of SP1B ∘ SP1A∘ SP2$(\cdot)$ is not clear. In the proof of Proposition 2, we show that SP1B$(\cdot)$ is continuous. However, as *SP*2$(\cdot)$ is a transit network loading process that simulates the integer number of passengers, it is inherently noncontinuous. Moreover, we can also show that the gradient of SP2$(\cdot)$ can be arbitrarily large due to network crowding and left behind (i.e., a slight change in path choices can lead to a large change in $\boldsymbol{\mu}$, see Online Appendix B for an example). Although the large gradient does not change the continuity mathematically, it does affect the numerical estimation and may result in precision issues.

In addition to the existence, we also discuss the uniqueness of the fixed point in Proposition 4.

Proposition 4. *The composed function* SP1B ∘ SP1A ∘ SP2$(\cdot)$ *has a unique fixed point if for any β and β',* $\|\text{SP1B} \circ \text{SP1A} \circ \text{SP2}(\beta) - \text{SP1B} \circ \text{SP1A} \circ \text{SP2}(\beta')\| \le \delta\|\beta - \beta'\|$, *where $\delta \in [0, 1)$ is a constant.*

Proof. The proof directly follows Banach's fixed-point theorem (Luan and Xia 2015). □

However, because SP1B ∘ SP1A ∘ SP2 has no analytical expression (due to the network loading process), it is hard to prove the contraction of SP1B ∘ SP1A ∘ *SP*2. According to the Banach fixed-point theorem, if the contraction holds, the unique fixed point can be obtained by the following procedure: start with an arbitrary $\beta^{(0)}$

and define a sequence $\{\beta^{(n)}\}$ by $\beta^{(n)} = \text{SP1B} \circ \text{SP1A} \circ \text{SP2}(\beta^{(\text{n}-1)})$ for $n \geq 1$. Then, the $\lim_{n\to\infty}\beta^{(n)}$ exists and the converged value is the fixed point.

Although we cannot prove the contraction of SP1B ∘ SP1A ∘ SP2 analytically, we apply the Banach fixed-point theorem to develop the solution approach (Algorithm 2) and validate the convergence later numerically. If $\lim_{n\to\infty}\beta_n$ exists, and the converged value is the solution of the original problem, then a *necessary* condition for the uniqueness is satisfied, which provides more evidence of the reasonableness of the decomposed method. The results of the numerical validation of the approach using synthetic data are discussed in Section 4.4. It is shown that $\lim_{n\to\infty}\beta_n$ converges, and the converged value is close to the solution of the original problem. (In Table 1, we show that the root mean square error for estimated path share is only 1.16%.)

Algorithm 2 (Solution Procedures for Path Choice Estimation)

```
Initialize $\beta^{(0)}$ and specify $K_t, K_b, \epsilon$.
$\boldsymbol{\mu}^{(0)}$ = Network Loading ($\beta^{(0)}, \boldsymbol{q}_e$) (Subproblem 2)
Set iteration counter $k = 0$.
do
    $k = k + 1$
    Solve Subproblem 1(a) with fixed $\boldsymbol{\mu}^{(k-1)}$ and return $\boldsymbol{p}^{(k)}$
    Solve Subproblem 1(b) with fixed $\boldsymbol{p}^{(k)}$ and return $\beta^{(k)}$
    Solve Subproblem 2 with $\beta^{(k)}$ as input and return $\boldsymbol{\mu}^{(k)}$
while $\|\beta^{(k)} - \beta^{(k-1)}\| > \epsilon$ and $k < K_t$
if $k < K_t$ then
    $\beta = \beta^{(k)}$
else
    $\beta = \sum_{k=K_b}^{K_t} \beta^{(k)} / (K_t - K_b + 1)$
return $\beta$
```

In Algorithm 2, $\beta^{(0)}$ is the initial value of β. The variable ϵ is a predetermined threshold for algorithm termination. To address the randomness in the network loading model, we also define a "burn-in" iteration K_b and a maximum iteration K_t. When β fluctuates because of randomness, we take the average of the last $K_t - K_b$ values of β as the final estimation.

4. Case Study

For the purpose of model illustration and validation, we apply the proposed modeling framework using data from Hong Kong's MTR network. The model is validated using both synthetic and real-world AFC data.

4.1. MTR Network

The map for the Hong Kong MTR system is shown in Figure 6. In this study, the airport express and light rail transit services are not considered because they are separated from the urban railway lines, and passengers who enter the urban railway lines from these services need to tap in again. The system consists of 10 lines and 114 stations, out of which 16 are transfer stations. In this network, most transfer stations connect only two lines. A special case is the Admiralty station on Hong Kong Island, where three lines pass through the same transfer station. The Admiralty station is in the CBD area of Hong Kong. During peak hours, it is very busy. Despite its near-capacity operation, the MTR system offers a good level of service with high on-time performance.

In the aggregated TS hypernetwork, the dimension of all TS paths (i.e., $|\boldsymbol{p}|$) is reasonable. This is because (1) in urban rail networks, the number of possible paths between an OD pair is limited (as opposed to road networks). For example, in the Hong Kong MTR network, there is a total of 91 stations, 8,190 OD pairs, and 13,545 paths. On average, there are 1.65 paths per OD pair. (2) In the aggregated TS hypernetwork, because we are using 15-minute intervals, the total number of space-time paths (i.e., combinations of (i_m, j, r)) only increases to 54,180 in the one-hour study period. Therefore, we do not encounter the typical column management problems as in typical road networks or fine-grained TS hypernetworks.

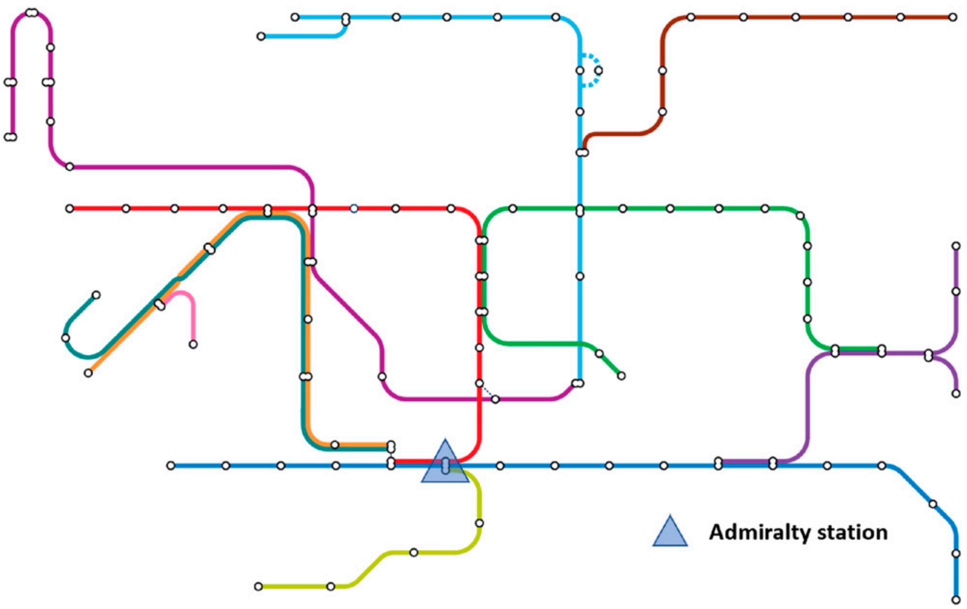
4.2. Validation Setting

We use AFC data from March 16, 2017 (Thursday), for model validation. The path sets for each OD pair are provided by MTR. Li (2014) conducted a revealed-preference (RP) route choice survey of more than 20,000 passengers in the MTR system and used the data to estimate a C-logit model. The estimation results are included in Online Appendix A. According to Li (2014), the

Table 1. β Estimation Results of Synthetic Data

		Estimated					
Variable	Synthetic ("true")	Proposed	BYO	CORS	Prob1	Prob2	$[L_\beta, U_\beta]$
In-vehicle time (β_1)	−0.147	−0.156	−0.205	−0.231	−0.085	−0.095	[−2, 0]
Number of transfers (β_2)	−0.573	−0.544	−1.218	−1.189	−0.840	−0.791	[−4, 0]
Relative walking time (β_3)	−1.271	−1.291	−2.499	−2.316	−1.015	−1.518	[−6, 0]
Commonality factor (β_4)	−3.679	−3.413	−6.184	−6.537	−2.881	−2.906	[−10, 0]
Objective function	—	10,328.8	42,390.6	37,066.1	58,123.0	57,043.3	—
RMSE (%)	—	1.16	7.34	6.74	7.67	7.44	—

Figure 6. (Color online) Hong Kong MTR Urban Rail System Map



following attributes were used in the path utility specification: (a) total in-vehicle time, (b) the number of transfers, (c) relative walking time (total walking time divided by total route distance), and (d) the commonality factor (Equation (6)). Other attributes such as average waiting time and average left behind rate were not significant based on the survey results and thus are not included. We follow the model specifications of Li (2014) so that there are reference parameters to generate synthetic data (details in the following) and to be compared with real-world estimation results (Section 4.5).

The evening peak (6:00 p.m. to 7:00 p.m.) is selected for validation. For simplicity, we assume the path shares are static during this hour. The weights in the objective function of Subproblem 1(a) are set as $w_1 = 1$ and $w_2 = 0$, which means no prior knowledge is available. The maximum number of iterations, K_t, is set to 15, and the "burn-in" iteration number, K_b, is set as 13. According to our numerical testing, a larger K_t does not change the estimation results of β. The term $\beta^{(0)}$ is set to zero for all parameters. The parameters of the network loading model are summarized here.

- Access, egress, and transfer walking times. They are platform-specific, obtained from field measurements.
- Train arrival and departure times: Approximated by the timetable. Future research can use AVL data to get actual train arrival and departure information.
- Capacity: 235 passengers per car based on MTR's calibration.
- Warm-up and cool-down time: 60 minutes warm-up and cool-down time for simulation.

Access/egress walking time is defined as the walking time between the fare machine and the train boarding platform. Warm-up (cool-down) time indicates the time before (after) simulation period starts (ends). It is needed because the simulation system usually starts from an empty state (no train and passengers).

Because the actual path choice information is usually unavailable, it is common to quantitatively validate the model with synthetic data. To generate the synthetic data, we first extract the OD entry flow from the real-world AFC records. Choice parameters β estimated in Li (2014) are treated as passengers' "true" behavior parameters (called synthetic β hereafter). We use the network loading model with the true OD entry flows (actual number of tap-in passengers according to the AFC data) and the synthetic β as input to simulate the travel of passengers in the system and record their tap-out time. The records of tap-out times and their true tap-in times are treated as the *synthetic AFC data*. The proposed path estimation approach is applied to the synthetic AFC data and validated based on its ability to recover the synthetic β values (i.e., the difference between the estimated and synthetic β values). The methodology is also applied to the real-world AFC data, providing further qualitative analysis and indirect comparison.

4.3. Benchmark Model

To evaluate the model performance, we compare the results from the proposed model to two types of methods: the simulation-based optimization (SBO) method (Mo et al. 2021) and probabilistic models (Sun and Xu 2012, Zhao et al. 2017).

4.3.1. SBO The formulation of the SBO benchmark approach is shown as follows:

$$\min_{\beta, p} \quad w_1 \sum_{i_m, j_n} (q^{i_m,j_n} - \tilde{q}^{i_m,j_n})^2 + w_2 \|\beta - \tilde{\beta}\|^2 \tag{29a}$$

$$\text{s.t.} \quad \text{Constraints (8b), (8d), (8f)–(8h),} \tag{29b}$$

$$q_r^{i_m,j_n} = \text{Network Loading } (\boldsymbol{p}, q^{i_m,j}) \quad \forall i_m \in \mathcal{N}, j_n \in \mathcal{N}, r \in \mathcal{R}_{i,j}, \tag{29c}$$

$$L_\beta \le \beta \le U_\beta, \tag{29d}$$

where L_β and U_β are predetermined lower and upper bounds of β. Equation (29) is equivalent to Equation (8). Constraints (8c) and (8e) are embedded into the network loading process (Equation (29c)). We set $w_1 = 1$ and $w_2 = 0$ as previously. Compared with our proposed model, the pure SBO methods need more function evaluations (i.e., simulation) to construct surrogate functions and calculate gradients. The terms L_β and U_β are usually required to narrow the feasible space and guide the algorithm to obtain reasonable results. The values of L_β and U_β are shown in Table 1. By introducing L_β and U_β, we actually provide the benchmark model with more information.

Many solution algorithms have been proposed to solve SBO problems. These algorithms generally belong to three major classes: direct search, gradient-based, and response surface methods (Osorio and Bierlaire 2013, Amaran et al. 2016). According to Osorio and Bierlaire (2013) and Cheng et al. (2019), response surface methods have good performance and are gaining popularity in the transportation literature. In this study, we adopt two response surface methods to solve the benchmark model: Bayesian optimization (BYO) (Snoek, Larochelle, and Adams 2012) and constrained optimization using response surfaces (CORS) (Regis and Shoemaker 2005).

- BYO aims to construct a probabilistic model of the objective function (response surface) and then exploit this model to determine where to evaluate the objective function for the next step. In each iteration, the probabilistic model is updated according to the posterior distribution of the objective function.
- CORS constructs a response surface model and updates the model based on all previously probed points. The criteria for selecting the next points to be evaluated are (a) finding points that have lower objective function value, and (b) improving the fitting of the response surface model by sampling feasible regions where little information exists.

SBO methods are usually unstable due to the randomness in the search process. For this reason, we perform 10 replications of each algorithm and report the mean and standard deviation of the objective function. The SBO methods are only used with the synthetic data so that we can compare the estimated path choice parameters to the synthetic ones.

4.3.2. Probabilistic Models.

The other type of benchmark is probabilistic models. We implement the Sun and Xu (2012) model (referred to as model "Prob1") and the Zhao et al. (2017) model (referred to as model "Prob2") methods because their models can directly output the left behind probabilities, which can be used to compare with our proposed approach. The ideas of the methods of Sun and Xu (2012) and Zhao et al. (2017) are similar: they both first estimate the left behind probabilities based on a subset of users that only have one possible route in the system. Then, based on the estimated left behind probabilities, they further estimate the route choices. The difference is that Sun and Xu (2012) assume that the number of trains a passenger needs to wait for follows a geometric distribution, whereas Zhao et al. (2017) assume a multinomial distribution. As Sun and Xu (2012) and Zhao et al. (2017) only output the path choice fractions, for comparison purposes, after getting the estimated left behind probabilities, we use the following maximum likelihood estimation to obtain path choice parameters for these two models:

$$\max_{\beta} \sum_{i_m\in\mathcal{N}, j\in\mathcal{S}} \sum_{u\in\mathcal{U}^{i_m,j}} \sum_{r\in\mathcal{R}_{i,j}} p_r^{i_m,j}(\beta)\cdot \mathbb{P}(TT_u^{\text{Obs}} \mid r, \text{Est. left behind}), \tag{30}$$

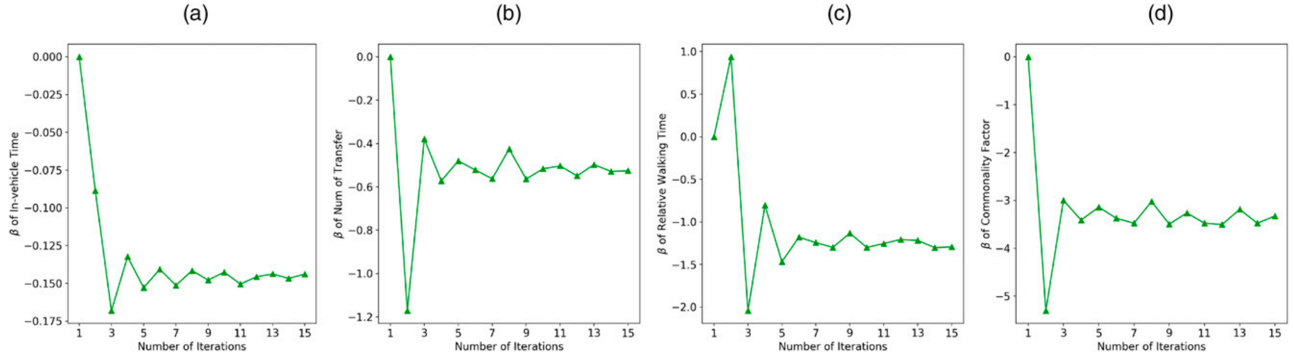
where $\mathcal{U}^{i_m,j}$ is the set of passengers with OD pair (i, j) and departure time index m. The expression $\mathbb{P}(TT_u^{\text{Obs}} \mid r, \text{Est. left behind})$ is the probability that passenger u's total journey time is TT_u^{Obs} given the estimated left behind information and that he/she has chosen path r. To also capture the distribution of access and egress walking times, we adopt a state-of-the-art formulation for $\mathbb{P}(TT_u^{\text{Obs}} \mid r, \text{est. left behind})$ based on Zhu et al. (2021)[2] The maximum likelihood estimation is implemented with the Python Scipy package and the BFGS optimizer (Nocedal and Wright 2006). Due to the computational burden, only 20% of the passengers (around 160,000) are used for the estimation. These passengers are randomly drawn from OD pairs with multiple route choices. This setting is similar to Sun et al. (2015), who randomly draw 190,000 passengers from different OD pairs in the Singapore metro network.

4.4. Synthetic Data Results

4.4.1. Convergence of $\boldsymbol{\beta}$.

The convergence results of β for the fixed-point algorithm (Algorithm 2) is depicted in Figure 7, which shows the value of β for each iteration. All β values appear to converge despite slight fluctuation in the tail. The results support the proposed solution approach (Algorithm 2) and validate the theoretical arguments on the solution's existence and uniqueness made in Section 3.5. There are two possible reasons for the fluctuation in the tail. First, due to the discontinuity of network loading (as discussed in Remark 4 and Online Appendix B), the fixed point of SP1B ∘ SP1A ∘ SP2(·) may not exist (thus the iteration may not converge). Second, although Subproblem 1B is strictly concave, due to precision issues, there may exist multiple sets of β values that result in close path shares,

Figure 7. (Color online) Convergence Behavior of Estimated β



Notes. (a) In-vehicle time. (b) Number of transfer. (c) Relative walking time. (d) Commonality factor.

especially when some path attributes are not statistically significant. However as shown in Algorithm 2, the final β will be the average over the last $K_t - K_b$ iterations, which reduces the influence of fluctuations.

4.4.2. Performance. Two indicators are reported during the iterations for the SBO and the proposed approaches. One is the objective function value of Subproblem 1a, which shows the square error of OD entry-exit flows. Another is the root-mean-square-error (RMSE) of estimated path shares:

$$\text{RMSE} = \sqrt{\sum_{i_m,j}\sum_{r\in\mathcal{R}_{i,j}} (p_r^{i_m,j} - \hat{p}_r^{i_m,j})^2 \Big/ \sum_{i,j} R_{i,j}}, \tag{31}$$

where $p_r^{i_m,j}$ are the estimated path shares and $\hat{p}_r^{i_m,j}$ are the synthetic path shares (unit is %). The term $\sum_{i,j} R_{i,j}$ is the total number of paths in the system.

Figure 8 shows the value of the objective function in Subproblem 1a as a function of the number of evaluations of the transit loading process. The error bars for the benchmark methods represent the standard deviation. The proposed method outperforms the benchmark models both in convergence rate and final solution quality. The RMSE comparison results are shown in Figure 9. The proposed method approaches the "true" path shares rapidly and has a lower estimation error than the benchmark models. The RMSE may not always decrease with the reduction of the objective function. This is because the relationship between path choices and OD entry-exit flows is highly nonlinear.

The comparison of estimated β and "true" (synthetic) β are shown in Table 1. The β values estimated using the proposed method are close to the "true" ones. The quality of the estimated solution is highlighted by the RMSE values. The RMSE of the β values estimated from the proposed method is significantly lower than the RMSE of those estimated from the benchmark methods. The probabilistic models show the worst estimation results. The possible reasons include (1) only using part of the user information and (2) not capturing the interaction between left behind and path choices.

4.4.3. Left Behind Estimation Comparison. Because the probabilistic models first estimate left behind then route choices, it cannot capture the interaction between path choices and left behind. Hence, we expect that the proposed model has a better left behind estimation.

For comparison, we calculate the left behind probability for each station in every 15-minute interval from 1800 to 1900 hours. Denote the probability of being left behind k times at station i and time index m as $LB_k^{i_m}$. $LB_k^{i_m}$ are obtained from the simulation model using the estimated path choices as input. Figure 10 shows the comparison of estimated and actual $LB_k^{i_m}$ for different models. The

Figure 8. (Color online) Subproblem (1a) Objective Function Results (Synthetic Data)

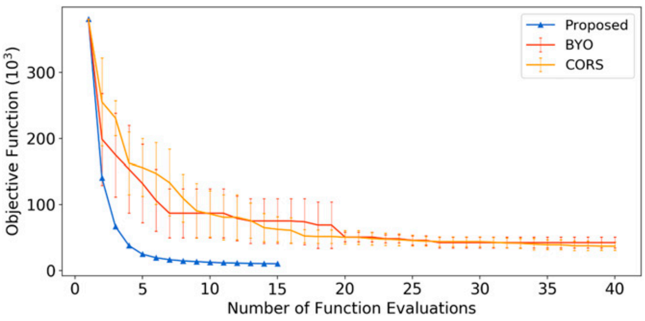


Figure 9. (Color online) RMSE of Path Fractions (Synthetic Data)

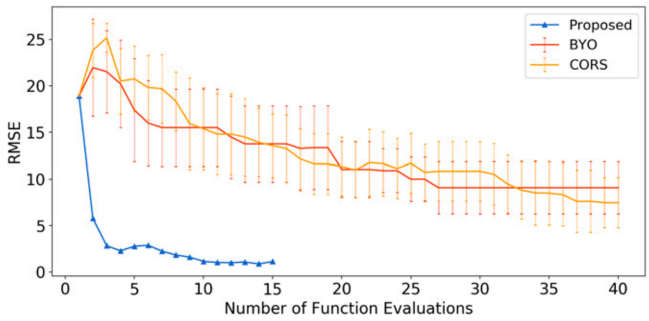
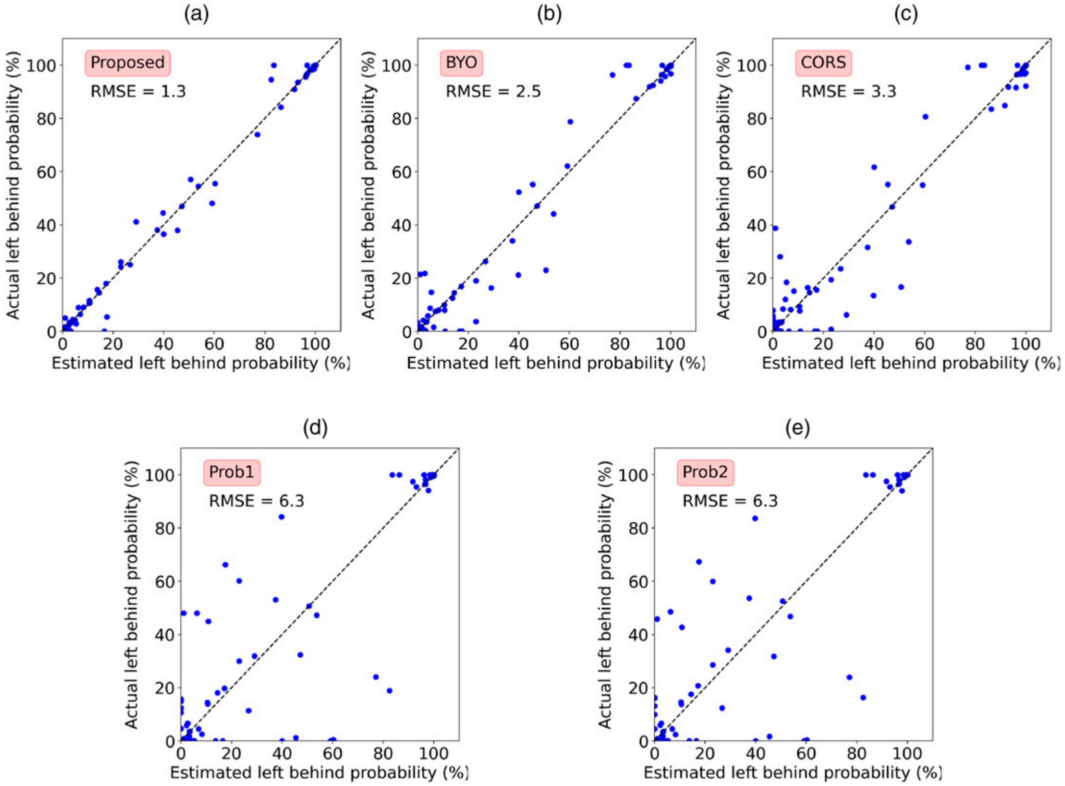


Figure 10. (Color online) Comparison of Left Behind Estimation



Notes. (a) Proposed. (b) BYO. (c) CORS. (d) Prob1. (e) Prob2.

RMSE is calculated using the similar way as Equation (31) by replacing $p_r^{i_m,j}$ with $LB_k^{i_m}$ (unit is percentage). The results show that the proposed model can well capture the left behind patterns in the system, with RMSE = 1.3%. Although the two probabilistic models have the worst left behind probability estimation, with RMSE = 6.3% for both models.

4.4.4. Computational Time. All models are run on a personal computer with a single-core CPU I9-9900K. The typical running time for a network loading (simulation) process is four minutes. For the probabilistic model (Sun and Xu 2012, Zhao et al. 2017), to efficiently estimate the route choices, we precalculate $\mathbb{P}(TT_u^{\text{Obs}} \mid r,$ Est. left behind) for all u and r before solving the maximum likelihood estimation (Equation (30)). It turns out that the probability precalculation takes a significant amount of time in the large-scale MTR network. This is because, for each path r and user u, there are many combinations of boarded trains and left behind situations that can result in the observed travel time, especially in the case of a path with multiple transfers (i.e., the passenger may be either left behind at the boarding station or the transfer stations). Because Sun and Xu (2012) assumes geometric distribution for the number of waiting vehicles, there are more nonzero left behind probabilities and it takes more computational time.

The comparison results in Table 2 show that the time bottleneck for the probabilistic models is the probability precalculation. After that, the maximum likelihood estimation can be efficiently solved. The total running times for the two probabilistic models are around six and four hours, respectively. There may exist better implementation techniques that could accelerate the probabilistic models, which is beyond the scope of this study. In terms of optimization-based methods, because the SBO methods require more function evaluations (i.e., simulation), they take a longer time than the proposed approach. Most of the computational time in optimization-based methods is spent in the simulation process, meaning that it can be

Table 2. Model Running Time Comparison

	CPU time (min)			
Models	LB estimation	Prob precalculation	Choice estimation	Total
Prob1	0.1	343.1	1.9	345.1
Prob2	0.3	210.9	1.7	212.9
CORS	NA	NA	169.2	169.2
BYO	NA	NA	178.6	178.6
Proposed	NA	NA	65.0	65.0

Note. NA, not applicable.

Table 3. β Estimation Results of Real-World Data

	In-vehicle time	Number of transfers	Relative walking time	Commonality factor
Estimated	−0.116	−0.920	−1.457	−1.775
Li (2014)	−0.147	−0.573	−1.271	−3.679

further accelerated if the simulation coding efficiency is improved (e.g., using C/C++ or Java).

4.5. Real-World Data Results

Because ground-truth path fractions are not available, we use the real-world AFC data to estimate the β values with the proposed method and compare them to ones obtained by Li (2014) (summarized in Table 3). The results show that the scale of all coefficients is similar. The tradeoff between the in-vehicle time and the number of transfers is reasonable, where one transfer is equivalent to 7.9 minutes of in-vehicle travel time compared with 3.9 minutes in Li (2014). The tradeoff between in-vehicle time and walking time is relatively small for long trips but significant for short trips. The results indicate that for a trip with four stations, 1 minute of transfer walking time is equivalent to 3.14 minutes of in-vehicle travel time (2.16 minutes in Li (2014)). For a trip with eight stations, 1 minute of walking time is equivalent to 1.57 minutes of in-vehicle travel time (1.08 minutes in Li (2014)). Hence, the marginal rates of substitution are reasonable and similar to the previous results (Li 2014). It is worth noting that because there are no ground truth path choices, the comparison of results using real-world data should be treated qualitatively rather than quantitatively. Moreover, it should be pointed out that since 2014, a number of changes have taken place in the MTR network. There was not only a growth in demand but also the opening of a new line that can result in path choice pattern changes.

Although we cannot directly compare path shares, other measurements can also reflect the quality of path shares. Field observation data for the Admiralty station Northbound platform during the testing period (1800–1900 hours) is available, which contains information on left behind rates (proportion of passengers who are not able to board the first train), the total number of arrival passengers (sum of new tap-in and transfer passengers), and the total number of boarding passengers. These measures can also be obtained from the network loading model using the path shares estimated from the proposed method as input. For comparison purposes, we also run the network loading model using two other path shares. The first is generated by a naive model that results in equal shares among all paths (referred to as "uniform" path shares). The second is based on the path choice model in Li (2014) to calculate path fractions.

The comparison results are shown in Table 4. Compared with the ground truth, the network loading model using the estimated path shares replicates closely the left behind rate, the number of arriving passengers, and the number of boarding passengers. The squared error of OD entry-exit flow (i.e., the objective function) is also the lowest. The performance in terms of left behind rate and number of arriving passengers for the estimated path shares are similar to that of Li (2014). However, the model with path shares estimated using the proposed method can outperform that of Li (2014) in the number of boarding passengers and OD entry-exit flows.

From an evaluation point of view, it is important to consider the performance of each path share generation method across all metrics in Table 4. The generated path shares are reasonable only if it performs well across all metrics. For example, the naive uniform path shares report a good left behind rate but significantly underestimates the number of boarding passengers and have a large squared error of OD entry-exit flows, which should be seen as reasonable path shares. We also observe that these metrics have different sensitivity to path shares. For example, the estimated left behind rates at Admiralty station from four methods are similar. This may be because the left behind rate at Admiralty station is largely determined by tap-in demands and not sensitive to the path shares. These metrics also reflect the performance of path shares at different levels. The first three metrics (related to Admiralty station) are station-level (local) indicators, whereas the fourth one (i.e., the squared error of OD entry-exit flows) is network level (global). As we aim at estimating path choices for the whole network, the fourth metric is more representative in terms of the quality of path shares as it captures the flows at the network level. The proposed model has the best performance in the fourth metric, indicating consistent estimation quality for all OD pairs.

4.6. Robustness

Model robustness is important for real-world applications. We test the performance of the model under

Table 4. Comparison of Various Measurement of Admiralty Station (1800 to 1900 hours)

	Left behind rate	Number of arriving passengers	Number of boarding passengers	Squared error of OD entry-exit flows
Ground-truth	0.747	24,945	23,926	—
Proposed model	0.724	24,589	23,570	1,044,692
Li (2014)	0.742	24,959	22,357	1,166,814
Uniform	0.779	25,683	18,767	1,323,594

different $\beta^{(0)}$ (initial β values) and data from different days. A robust model should output similar estimated β values regardless of initial values. The estimated β for different days should also be similar because passengers' choice behavior is stable in the short term.

4.6.1. Sensitivity to $\beta^{(0)}$. The sensitivity analysis to different values of $\beta^{(0)}$ are conducted using the synthetic data. Twelve different $\beta^{(0)}$ are drawn from a uniform distribution $U(L_\beta, U_\beta)$. Figure 11(a) shows the convergence of the objective function for different $\beta^{(0)}$ values. In early iterations, the initial objective function values vary a lot. However, after around 10 iterations, they all converge to the same value regardless of $\beta^{(0)}$. These results demonstrate the robustness of the approach with respect to initial β values. Figure 11(b) is the boxplot of the estimated β parameters for different $\beta^{(0)}$. The variables that $\beta_1, \ldots, \beta_4$ correspond to can be found in Table 1. The estimated β_1, β_2, and β_3 values are very stable regardless of $\beta^{(0)}$. The estimated β_4 value (commonality factor) shows some fluctuations but still within a small range (the 95% confidence interval is $[-3.2, -3.6]$). This also corresponds to the survey estimation results where β_4 has a relatively low t value (see Online Appendix A).

4.6.2. Sensitivity to Data from Different Days. To test the robustness of the model in terms of OD demands from different days, the model was applied using actual AFC data on different days in one month (from March 6 to 30, all weekdays). Figure 12 compares the estimated β values for different days of week (Monday to Friday). In general, all estimated values are consistent across days except for the coefficient of relative walking time on Friday. This may be because Friday nights are the start of the weekend, and passengers may have different travel patterns and behavior. Walking time, for example, is less important for entertainment trips that may take place during the evening peak on Friday.

Figure 13 shows the mean and standard deviation of estimate β for different weeks. We find that the parameters for in-vehicle time, number of transfers, and commonality factors are stable across different weeks. The values are similar and standard deviations are relatively small. However, the estimated coefficient for relative walking time is not stable with high variance. Because we expect that the choice behavior does not change too much within a month, the nonstability indicates that relative walking time may not be statistically significant in determining path choices.

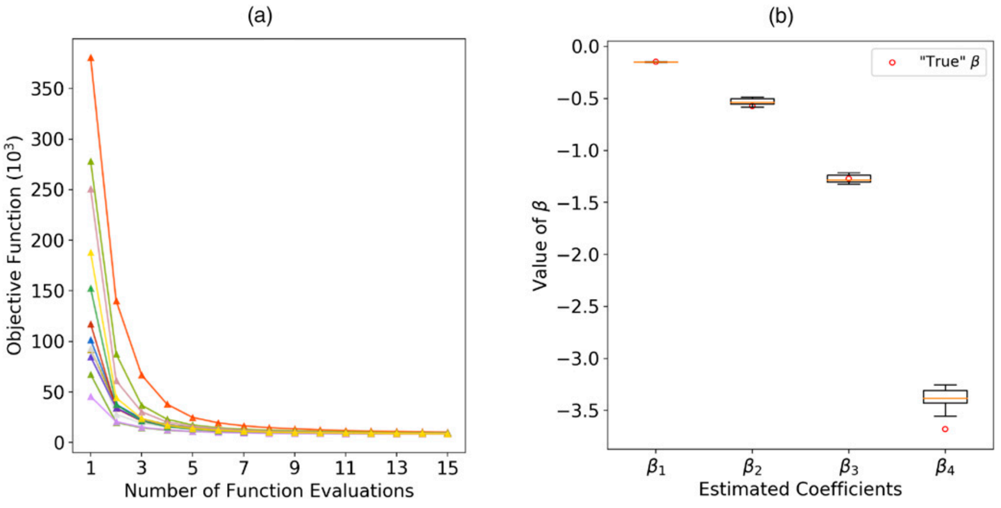
Overall, the proposed model is robust with respect to data from different weekdays in terms of estimating significant parameters (such as in-vehicle time, number of transfers, and commonality factors). For nonsignificant parameters (such as relative walking time), the estimation has more variances.

5. Conclusion

This paper presents an aggregated time-space hypernetwork approach to infer passenger route choice behavior in urban rail systems with both entry and exit AFC transactions. The approach models explicitly the interactions between path choices and left behind and the interactions among stations in terms of crowding. The path choice estimation is modeled as an optimization problem. The original intractable problem is decomposed into three tractable subproblems that can be solved efficiently. Case studies using synthetic and actual data validate the effectiveness and robustness of the approach.

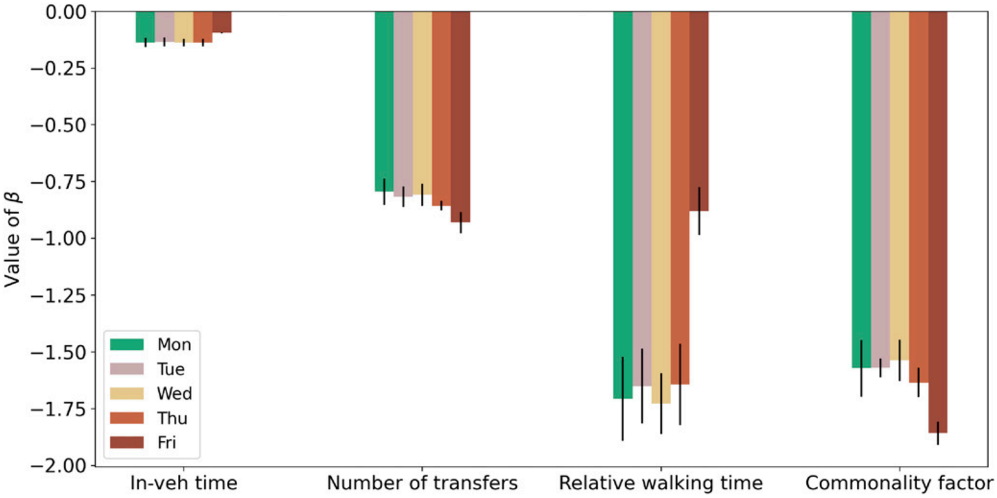
The proposed ex post path choice estimation framework is general. It is applicable to accommodate different choice model structures (nested logit models) by

Figure 11. (Color online) Model Sensitivity to Initial β Values



Notes. (a) Convergence of the objective function (different curves correspond to different $\beta^{(0)}$). (b) Boxplot of estimated coefficients for different $\beta^{(0)}$.

Figure 12. (Color online) Comparison of Estimated β Values for Different Days

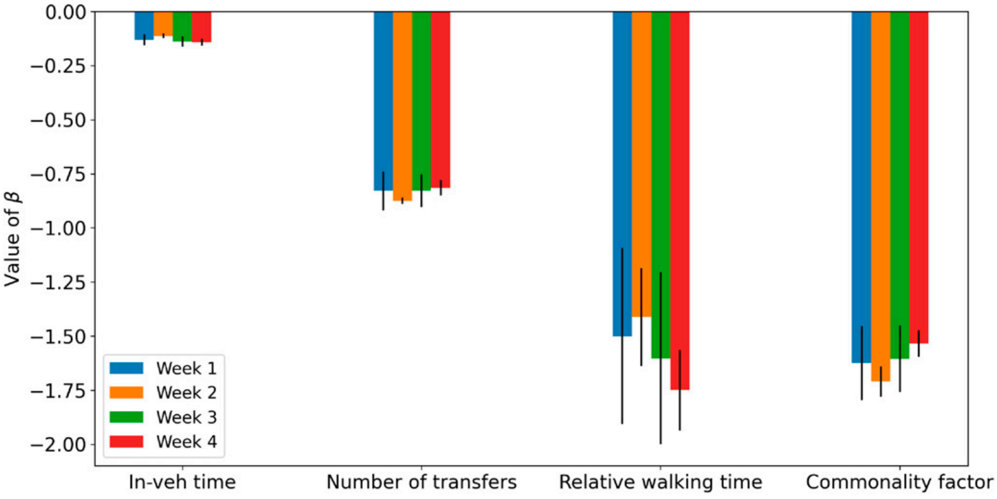


Note. Black lines indicate ±standard deviation.

changing Subproblem (1b). Also, it can be extended to allow different β values for different groups of passengers to better capture the heterogeneity in real-world route choice behavior. It could drive applications in performance monitoring (e.g., network crowding) and evaluation of control/management strategies in planning and operations (e.g., train timetabling and network change).

A limitation in this study is assuming that a passenger's path choice is based on the attractiveness of a path as measured by its attributes. However, some complex choice behaviors, such as that passengers may choose whatever line comes first when they wait at a station with multiple lines to their destinations (i.e., common line case), may not be captured by the standard discrete choice models. In the literature, an alternative setting is to estimate path choice fractions $\boldsymbol{p}$ directly, instead of β. This setting does not assume any behavior models for path choices. However, there are pros and cons for estimating $\boldsymbol{p}$ and β. When estimating $\boldsymbol{p}$ directly, we can capture complex path choice behavior beyond behavior models. However, the estimation errors can be large for some OD pairs with sparse or small samples. When estimating β, the domain knowledge (i.e., behavior assumption) is incorporated. All OD pairs and path shares can obtain reasonable results even if there is sample sparsity. However, the drawback is that complex choice behaviors beyond the behavior model (such as the common line case mentioned previously) cannot be captured. In this study, we estimate β considering the estimation accuracy of all OD pairs and application potentials.

Figure 13. (Color online) Comparison of Estimated β Values for Different Weeks



Note. Black lines indicate ±standard deviation.

Acknowledgments

The authors thank the Hong Kong Mass Transit Railway for support and providing data for this research and the reviewers for valuable comments on enhancing model proofs and arguments.

Endnotes

[1] We illustrate why SDTNL is needed instead of frequency-based static network loading models in Section 2.3.

[2] In Zhao et al. (2017), they used the probability of a travel plan to represent $\mathbb{P}(TT_u^{\text{Obs}} \mid r$, est. left behind) but ignoring the access and egress walking time distribution.

References

Afif Ben Amar DO (2022) *Topology and Approximate Fixed Points* (Springer, Berlin).

Amaran S, Sahinidis NV, Sharda B, Bury SJ (2016) Simulation optimization: A review of algorithms and applications. *Ann. Oper. Res.* 240(1):351–380.

Atasoy B, Ikeda T, Song X, Ben-Akiva ME (2015) The concept and impact analysis of a flexible mobility on demand system. *Transportation Res., Part C Emerging Tech.* 56:373–392.

Cadenas Aldana RA (2007) Continuity and theorem of heine-cantor. *Divulgaciones Matemáticas* 15(1):71–76.

Cascetta E, Nuzzolo A, Russo F, Vitetta A (1996) A modified logit route choice model overcoming path overlapping problems. specification and some calibration results for interurban networks. Transportation and traffic theory. *Proc. 13th Internat. Sympos. on Transportation and Traffic Theory*. https://trid.trb.org/view/481284.

Cheng Q, Wang S, Liu Z, Yuan Y (2019) Surrogate-based simulation optimization approach for day-to-day dynamics model calibration with real data. *Transportation Res. Part C Emerging Tech.* 105:422–438.

Davis J, Gallego G, Topaloglu H (2013) Assortment planning under the multinomial logit model with totally unimodular constraint structures. Department of IEOR 335–357, Cornell University, Ithaca, NY.

De Cea J, Fernández E (1993) Transit assignment for congested public transport systems: An equilibrium model. *Transportation Sci.* 27(2):133–147.

Eluru N, Chakour V, El-Geneidy AM (2012) Travel mode choice and transit route choice behavior in Montreal: Insights from McGill University members commute patterns. *Public Transportation (Berlin)* 4(2):129–149.

Hamdouch Y, Lawphongpanich S (2008) Schedule-based transit assignment model with travel strategies and capacity constraints. *Transportation Res. Part B: Methodological* 42(7-8):663–684.

Hamdouch Y, Ho H, Sumalee A, Wang G (2011) Schedule-based transit assignment model with vehicle capacity and seat availability. *Transportation Res. Part B: Methodological* 45(10):1805–1830.

Han B, Zhou W, Li D, Yin H (2015) Dynamic schedule-based assignment model for urban rail transit network with capacity constraints. *Scientific World J.* 2015:940815.

Koutsopoulos HN, Ma Z, Noursalehi P, Zhu Y (2019) *Transit Data Analytics for Planning, Monitoring, Control, and Information* (Elsevier, New York).

Kroon L, Maróti G, Nielsen L (2015) Rescheduling of railway rolling stock with dynamic passenger flows. *Transportation Sci.* 49(2): 165–184.

Kumar P, Khani A, He Q (2018) A robust method for estimating transit passenger trajectories using automated data. *Transportation Res., Part C Emerging Tech.* 95:731–747.

Kusakabe T, Iryo T, Asakura Y (2010) Estimation method for railway passengers' train choice behavior with smart card transaction data. *Transportation* 37(5):731–749.

Lam SH, Xie F (2002) Transit path-choice models that use revealed preference and stated preference data. *Transportation Res. Record* 1:58–65.

Li W (2014) Route and transfer station choice modeling in the MTR system. Working paper, Massachusetts Institute of Technology, Cambridge, MA.

Luan H, Xia Z (2015) Theorem of existence and uniqueness of fixed points of monotone operators. *Genetic and Evolutionary Computing* (Springer, Berlin), 11–17.

Ma Z, Koutsopoulos HN, Chen Y, Wilson NHM (2019) Estimation of denied boarding in urban rail systems: Alternative formulations and comparative analysis. *Transportation Res. Record* 2673(11): 771–778.

Mo B, Ma Z, Koutsopoulos HN, Zhao J (2020) Capacity-constrained network performance model for urban rail systems. *Transportation Res. Record* 2674(5):59–69.

Mo B, Ma Z, Koutsopoulos HN, Zhao J (2021) Calibrating path choices and train capacities for urban rail transit simulation models using smart card and train movement data. *J. Adv. Transportation* 2021:5597130.

Nazem M, Trépanier M, Morency C (2011) Demographic analysis of route choice for public transit. *Transportation Res. Record* 2217(1): 71–78.

Nguyen S, Pallottino S, Malucelli F (2001) A modeling framework for passenger assignment on a transport network with timetables. *Transportation Sci.* 35(3):238–249.

Nielsen OA (2000) A stochastic transit assignment model considering differences in passengers utility functions. *Transportation Res. Part B: Methodological* 34(5):377–402.

Nocedal J, Wright S (2006) *Numerical Optimization* (Springer Science & Business Media, New York).

Nuzzolo A, Crisalli U, Rosati L (2012) A schedule-based assignment model with explicit capacity constraints for congested transit networks. *Transportation Res., Part C Emerging Tech.* 20(1):16–33.

Ok EA (2011) *Real Analysis with Economic Applications* (Princeton University Press, Princeton, NJ).

Osorio C, Bierlaire M (2013) A simulation-based optimization framework for urban transportation problems. *Oper. Res.* 61(6):1333–1345.

Papke LE, Wooldridge JM (1996) Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *J. Appl. Econometrics* 11(6):619–632.

Prato CG (2009) Route choice modeling: Past, present and future research directions. *J. Choice Modelling* 2(1):65–100.

Regis RG, Shoemaker CA (2005) Constrained global optimization of expensive black box functions using radial basis functions. *J. Global Optim.* 31(1):153–171.

Scarf H (1967) The approximation of fixed points of a continuous mapping. *SIAM J. Appl. Math.* 15(5):1328–1343.

Schmöcker JD, Bell MG, Kurauchi F (2008) A quasi-dynamic capacity constrained frequency-based transit assignment model. *Transportation Res. Part B: Methodological* 42(10):925–945.

Schmöcker JD, Fonzone A, Shimamoto H, Kurauchi F, Bell MG (2011) Frequency-based transit assignment considering seat capacities. *Transportation Res. Part B: Methodological* 45(2):392–408.

Snoek J, Larochelle H, Adams RP (2012) Practical Bayesian optimization of machine learning algorithms. Pereira F, Burges CJ, Bottou L, Weinberger KQ, eds. *Proc. Adv. Neural Inform. Processing Systems* 25:2951–2959.

Song W, Han K, Wang Y, Friesz T, Del Castillo E (2017) Statistical metamodeling of dynamic network loading. *Transportation Res. Procedia* 23:263–282.

Spiess H, Florian M (1989) Optimal strategies: A new assignment model for transit networks. *Transportation Res. Part B: Methodological* 23(2):83–102.

Stasko T, Levine B, Reddy A (2016) Time-expanded network model of train-level subway ridership flows using actual train movement data. *Transportation Res. Record* 2540(1):92–101.

Sun L, Lu Y, Jin JG, Lee DH, Axhausen KW (2015) An integrated Bayesian approach for passenger flow assignment in metro networks. *Transportation Res., Part C Emerging Tech.* 52:116–131.

Sun Y, Xu R (2012) Rail transit travel time reliability and estimation of passenger route choice behavior: Analysis using automatic fare collection data. *Transportation Res. Record* 2275(1):58–67.

Szeto W, Jiang Y (2014) Transit assignment: Approach-based formulation, extragradient method, and paradox. *Transportation Res. Part B: Methodological* 62:51–76.

Szeto W, Jiang Y, Wong KI, Solayappan M (2013) Reliability-based stochastic transit assignment with capacity constraints: Formulation and solution method. *Transportation Res., Part C Emerging Tech.* 35:286–304.

Vanderbei R (1991) Uniform continuity is almost Lipschitz continuity. Technical Report SOR-91-11, Statistics and Operations Research Series, Princeton University, Princeton, NJ.

Wu JH, Florian M, Marcotte P (1994) Transit equilibrium assignment: A model and solution algorithms. *Transportation Sci.* 28(3):193–203.

Xu X, Xie L, Li H, Qin L (2018) Learning the route choice behavior of subway passengers from AFC data. *Expert Systems Appl.* 95:324–332.

Zhao J, Zhang F, Tu L, Xu C, Shen D, Tian C, Li XY, Li Z (2017) Estimation of passenger route choice pattern using smart card data for complex metro systems. *IEEE Trans. Intelligent Transportation Systems* 18(4):790–801.

Zhou F, Xu RH (2012) Model of passenger flow assignment for urban rail transit based on entry and exit time constraints. *Transportation Res. Record* 2284(1):57–61.

Zhu Y (2017) Passenger-to-itinerary assignment model based on automated data. PhD thesis, Northeastern University, Evanston, IL.

Zhu Y, Koutsopoulos HN, Wilson NH (2021) Passenger itinerary inference model for congested urban rail networks. *Transportation Res., Part C Emerging Tech.* (Elsevier) 123:102896.

Appendix A: Passenger Route Choice Model for MTR System

These results are from Li (2014). The C-logit Model formulation is the same as Eq. (5) and Eq. (6). A total number of 31,640 passengers completed the questionnaire. After filtering duplicate responses, 26,996 responses were available. The model results are shown in Table 1. The main explanatory variables are the total in-vehicle time, relative transfer walking time, and number of transfers. All variables are statistically significant with the expected signs. Routes with high in-vehicle time, walking time, and number of transfers are less likely to be chosen by passengers.

Table 1 Route Choice Model Estimation Results

	Estimate	Std. Error	t-value	
In-vehicle time	-0.147	0.011	-13.64	***
Relative walking time	-1.271	0.278	-4.56	***
Number of transfers	-0.573	0.084	-6.18	***
Commonality factor	-3.679	1.273	-2.89	**
$\rho^2 = 0.54$				

***: $p < 0.01$; **: $p < 0.05$.

Appendix B: An explanation of large gradient for the transit network loading

In this section, we want to show that a slight change in β may lead to a large change in SP2(β). Consider a single direction rail line with M stations and a fixed headway H (Figure 1). Every train has a capacity of 1. Assume that under the path choice parameter β, there is one passenger waiting at station 2, and no passengers from station 1 choose to use this line. Hence, the passenger at station 2 can board the first available train. Suppose that under the path choice parameter $\beta + \Delta\beta$, passengers from station 1 start to use the rail line (with proportion Δp). And the total demand for station 1 is q. Therefore, the waiting passenger at station 2 will be able to board the train after being left behind $\Delta p \times q$ times with increased waiting time by $\Delta p \times q \times H$. These values can be arbitrarily large for large values of q and H. Hence, the path exit rate for the passenger at station 2 can change dramatically depending on how many times they have been left behind.

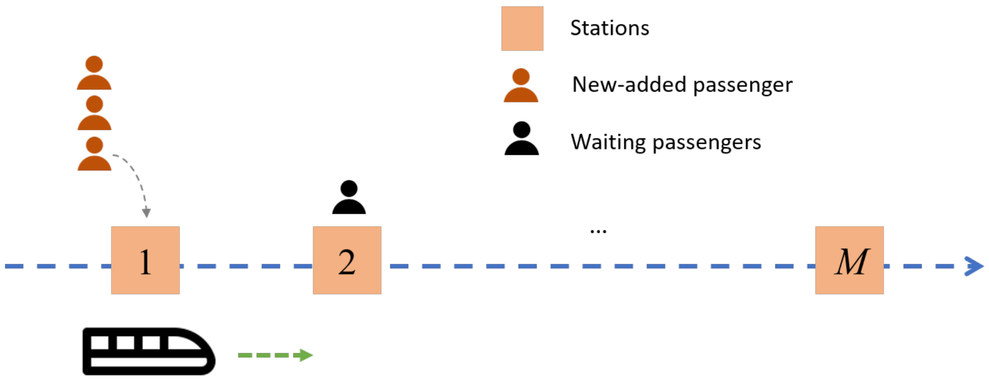


Figure 1 Illustration of large gradients for the transit network loading

Appendix C: Impact of time interval length

In this section, we evaluate the impact of different values of τ (i.e., the time interval length). Specifically, β estimation results for $\tau = 15, 10, 5$ minutes are shown in Table 2. $\tau = 10$ minutes results in a slightly better estimation than that of $\tau = 15$ minutes, which implies that $\tau = 15$ minutes may be slightly larger and this setting ignores some temporal variations. $\tau = 5$ minutes leads to worst results. As mentioned in Section 3.1, the value of τ should be consistent with the information granularity. Clearly, $\tau = 5$ minutes is too fine-grained which makes the system sensitive to observation errors. In the simulation model, the access/egress/transfer walking time are random variables with the mean around 1∼2 minutes and standard deviations similar to the mean. Therefore, the OD entry-exit flows may have random errors at the granularity level of $\tau = 5$ minutes.

Table 2 Impact of time interval length on β estimation

Variable	Synthetic ("True")	Estimates from the proposed model		
		$\tau = 15$ mins	$\tau = 10$ mins	$\tau = 5$ mins
In-vehicle time (β_1)	-0.147	-0.156 (6.1%)	-0.151 (2.8%)	-0.140 (4.7%)
Number of transfers (β_2)	-0.573	-0.544 (5.1%)	-0.534 (6.7%)	-0.504 (12.0%)
Relative walking time (β_3)	-1.271	-1.291 (1.6%)	-1.292 (1.7%)	-1.332 (4.8%)
Commonality factor (β_4)	-3.679	-3.413 (7.2%)	-3.427 (6.9%)	-3.028 (17.7%)
Average absolute relative error	-	5.0%	4.5%	9.8%

[1]: Numbers in the parentheses are the absolute relative error compared to the synthetic β

Appendix D: Extension to other route choice models

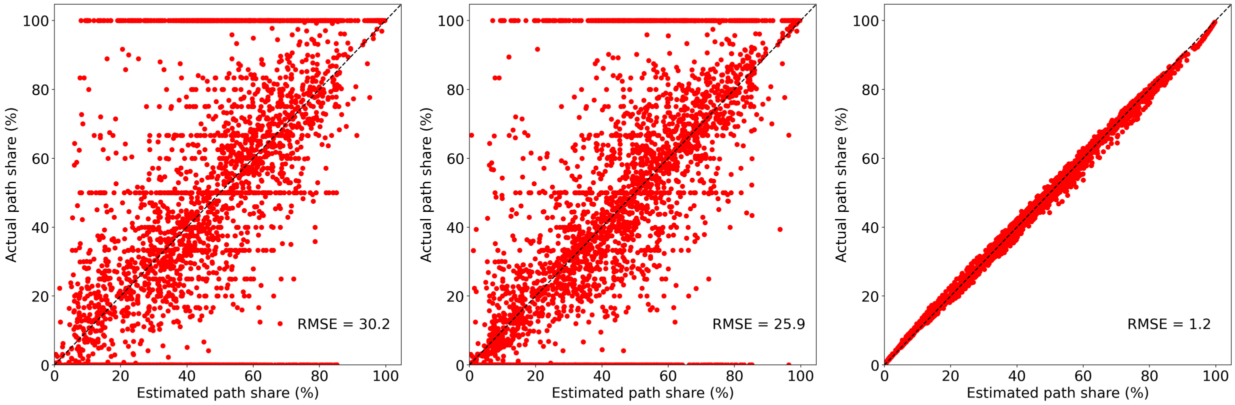
The proposed approach can be extended to other route choice models. The C-logit model can be replaced by any other strictly convex discrete choice model, such as path-size logit. To implement other choice models, we only need to replace sub-problem 1b as follows:

$$\max_{\beta} \sum_{i_m \in \mathcal{N}, j \in \mathcal{S}} q^{i_m,j} \sum_{r \in \mathcal{R}_{i,j}} p_r^{i_m,j} \cdot \log \mathbb{P}[r\,; \beta] \tag{1}$$

where $\mathbb{P}[r\,; \beta]$ is any convex discrete choice models with parameters β. With the strict convexity, all the analysis in this article is still effective. Models such as cross-nested logit and logit kernel models (Ben-Akiva, Ramming, and Bekhor 2004) may be non-convex (or weakly convex). The non-convexity (or weak convexity) makes Lemma 1 invalid because the solution of subproblem 1b may not be unique. However, we can still implement the proposed approach. But we may not have good convergence with convex choice models. One possible modification is to impose the successive average in Algorithm 2. That is, set $\beta^{(k)} = \frac{k-1}{k}\beta^{(k-1)} + \frac{1}{k}\hat{\beta}^{(k)}$, where $\hat{\beta}^{(k)}$ is the estimation of β in iteration k. This may be helpful for convergence.

Appendix E: Effect of subproblem 1 linearization

In this section, we explicitly test the effect of the two procedures to approximately linearize the logit constraints in sub-problem 1 (Section 3.4. Specifically, with the same basic settings as the case study, we solve three problems to get path shares for comparison: 1) The first one is a model similar to subproblem 1a (Eq. 19) but eliminating the ALC (i.e., Constraints 14 and 15). 2) The second is subproblem 1a (i.e., with the ALC). 3) The third model is subproblem 1b (i.e., logit correction) with the results of the second model as input. The path shares are generated using the estimated β. The first two models both use the actual path exit rate $\boldsymbol{\mu}$ as inputs. We compare the output path shares $p_r^{i_m,j}$ in Figures 2a $\sim$ 2c. It is found that adding ALC can improve the path share estimation and decrease the RMSE from 30.2% to 25.9%. After logit correction, the estimated path shares are almost the same as the actual ones (because we use the actual $\boldsymbol{\mu}$ as input). This implies that both techniques for the linearization of subproblem 1 are important to improve the quality of estimated path shares.



(a) Without ALC (b) With ALC (c) After logit correction

Figure 2 Comparison of path shares for the effect of sub-problem 1 linearization

References

Ben-Akiva ME, Ramming MS, Bekhor S, 2004 *Route choice models. Human Behaviour and Traffic Networks*, 23–45 (Springer).

Li W, 2014 *Route and transfer station choice modeling in the MTR system*, working paper.