Manuscript

Calibrating Route Choice for Urban Rail System: A Comparative Analysis Using Simulation-based Optimization Methods

Baichuan Mo
Research Assistant
Department of Civil and Environmental Engineering
Massachusetts Institute of Technology
77 Massachusetts Ave, Cambridge, MA 20139
Tel: +1 857-999-5906; Email: baichuan@mit.edu

Zhenliang Ma, Ph.D. (Corresponding Author)
Assitant Professor
Department of Civil and Environmental Engineering
Monash University
Clayton, VIC 3800, Australia
Tel: +61 175109768; Email: mike.ma@monash.edu

Haris N. Koutsopoulos, Ph.D.
Professor
Department of Civil and Environmental Engineering
Northeastern University
Boston, MA 02115, United States
Tel: +1 617-373-6263; Email: h.koutsopoulos@northeastern.edu

Jinhua Zhao, Ph.D.
Associate Professor
Department of Urban Studies and Planning
Massachusetts Institute of Technology
77 Massachusetts Ave, Cambridge, MA 20139
Tel: +1 617-253-7594; Email: jinhua@mit.edu

Word count:5,853 words text + 3 table × 250 words (each) = 6,603 words

Submission Date: July 31, 2019

ABSTRACT
This paper proposed a simulation-based optimization framework to identify the route choices patterns with an event-driven transit network loading model. Five optimizers of three main brunches of SBO methods are applied in this paper for comparative analysis, which includes Nelder-Mead Simplex Algorithm (NMSA), Mesh Adaptive Direct Search (MADS), Simultaneous Perturbation Stochastic Approximation, Bayesian Optimization (BYO) and Constrained Optimization using Response Surfaces (CORS). We use the real-world metro system of Hong Kong Mass Transit Railway (MTR) as the testbed. The results show the response surface methods (BYO and CORS) have the fastest convergence speed. They can also reach the lowest objective function value. Specially, the CORS method has the best performance over other SBO techniques.

Keywords: Simulation-based optimization, Route choice estimation, Smart card data

INTRODUCTION

Motivation

Urban rail systems are important components of the urban transportation system. Given their high reliability and large capacity, urban rail services attract high passenger demand, but this can also lead to problems such as overcrowding, disturbances and disruptions, which dramatically decrease level of service and impose negative effects on passengers. To maintain service reliability and develop efficient response strategies, it is crucial for operations to understand the passenger demand and flow patterns in the urban rail network.

The implementation of a transit assignment (simulation) model for metro systems provides a powerful instrument for network performance monitoring, which enables operators to characterize the level of service and make decisions accordingly. A typical simulation model requires Origin-Destination (OD) matrix and route choice fractions as input. Thanks to the wide deployment of automated fare collection (AFC) systems, the OD demand can be directly observed from recorded transactions. However, obtaining the corresponding route choices remains a challenge for both industry and academia. Once route choices are obtained, operators can easily leverage the assignment model to identify the network performance, then adjust the operation strategies accordingly to relieve congestion, improve efficiency, etc..

Related work

The traditional way to calibrate route choices is on-site survey, for which researchers ask people's real-world route choices and estimated a choice model to construct the route choice fractions. However, these survey-based methods are time-consuming and labor-intensive, limiting their ability to real-world practice. In addition, the results are only valid for current network and can't be used to test counter-factual scenarios and hard to update when network changes. To overcome these disadvantages, many new methods based on AFC data has been proposed.

AFC systems are designed to conveniently charge passengers who use the metro system. When passengers tap in or tap out in the system with a smart card, the exact locations and time of the transactions will be recorded, which provide rich information for analyzing passenger behaviors. On the context of route choice estimation, the AFC data-based methods can be categorized into two groups: the route-identification methods (*1–3*) and the parameters-inference methods (*4–7*). The former studies aimed to identify the exact route chosen by each user. The route attributes are used to evaluate how likely a passenger's trip from their observed origin to their observed destination was taken along any possible route. While the latter studies formulated probabilistic models to describe passengers' decision-making behaviors. Bayesian inference is usually used to estimate the corresponding parameters and thus derive the route choice fractions. Despite using different methods, the key component for those AFC data-based studies are similar. They all attempted to match the model-derived journey time with the observed journey time from AFC data. However, a shortcoming for these studies is that the denied-boarding phenomenon is not well addressed (*8*). In a congested metro system, passengers are likely to be denied boarding due to the limited capacity of trains. Denied boardings will increase the passengers' waiting time on the platform, thus increasing their total travel time. It is possible that the journey time for a longer-distance route without denied boarding is close to the journey time of a shorter-distance route with multiple denied boardings, which makes the two routes indistinguishable for the purely journey time-based methods (*9*).

To address the denied-boarding phenomenon in route choice estimation problems, it is important to incorporate the transit assignment model with the information of network topology and train operation schedule (*8*). By setting the network loading criteria of the transit assignment model, we can naturally incorporate the denied-boarding phenomenon in the congested network, which not only considers the denied boarding itself, but also the internal correlation of denied-boarding probability among different stations (*8*). However, as known in the literature, the schedule-based dynamic transit assignment is a complicated problem without direct closed form (*10*). A typical way to deal with the non-analytic problem is the simulation-based (black-box) optimization (SBO) methods. SBO methods are designed to solve optimization problems where the objective function and its derivatives are difficult and expensive to evaluate, which have been widely used to solve the problems of congestion pricing (*11, 12*), traffic signal control (*13–16*), transit scheduling (*17*), ride sharing (*18*), supply chain management (*19*), liner shipping (*20*) and more. Particularly, in the domain of route choice estimation, Cheng et al. (*21*) developped a SBO method to calibrate the day-to-day route choice for road traffic systems with license plate recognition data. However, the SBO methods have not been applied to urban rail system for route choice calibration problem. In general, there are three classes of methods for the SBO, including the direct search method, the gradient-based method, and the response surface (meta-model) method (*13*). None of the previous studies has compared the effectiveness of different SBO methods on the route choice estimation.

Paper objectives and organization

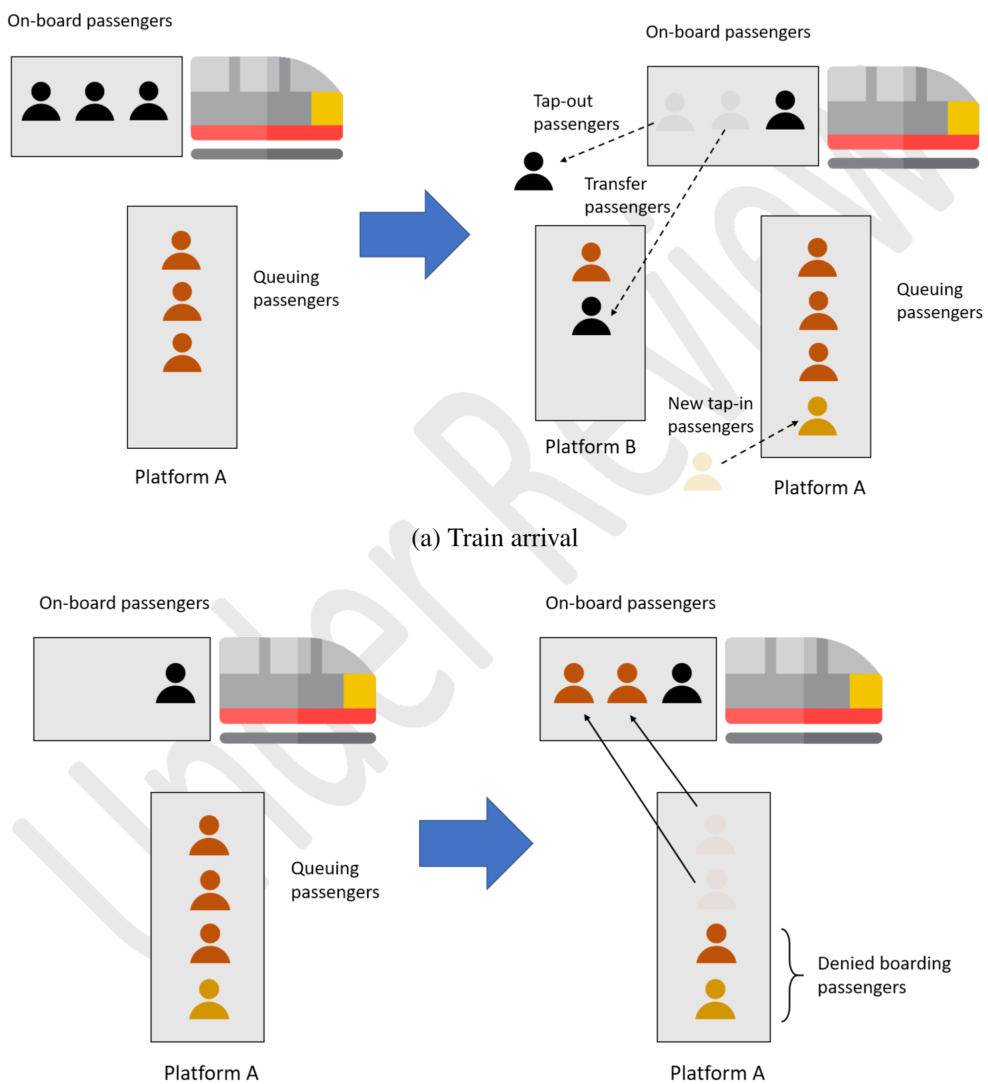
In this study, we proposed a SBO framework to identify the route choices patterns with an event-driven transit network loading (TNL) model (*22*). Five optimizers of three main brunches of SBO methods are applied in this paper for comparative analysis, which includes Nelder-Mead Simplex Algorithm (NMSA), Mesh Adaptive Direct Search (MADS), Simultaneous Perturbation Stochastic Approximation, Bayesian Optimization (BYO) and Constrained Optimization using Response Surfaces (CORS). The paper focuses on SBO techniques with good short-term performance. That is, we compare the all the SBO methods within a tight computational budget. The computational budget is defined as a limited number of simulation runs. Such techniques respond to the needs of transportation practitioners by allowing them to address problems in a practical manner. We use the real-world metro system of Hong Kong Mass Transit Railway (MTR) as the testbed. The results show the response surface methods have the fastest convergence speed. They can also reach the lowest objective function value. Specially, the CORS method has the best performance over other SBO techniques.

This remainder of this paper is organized as follows: In Section 2, we propose the modeling framework, which contains several components including mode assumption, problem definition and description of all SBO methods. We apply the proposed framework on the Hong Kong MTR network as a case study in Section 3. The quantitatively model comparison based on the synthetic AFC data are conducted. Finally, we conclude our study, summarize our main findings and discuss future research directions in Section 4 .

METHDOLOGY

Transit network loading model

To perform the SBO methods, we first need a simulation engine which can take the decision variables as input and output the data for calibration. In this study, the transit network loading (TNL) model is used as the simulation engine. TNL models aim to assign passengers over a transit network given the (dynamic) OD entry demand and route choice. The schedule-based TNL models are more appropriate for this study because they can capture the detailed travel behaviors in the network (e.g. queuing, transferring, boarding and alighting) and thus is closer to reality than the frequency-based TNL model (*23*). In this study, we applied an event-driven schedule-based TNL model (*8, 24*) as the simulator. The model takes OD entry demand (tap-in passengers), route choice parameters, time table and infrastructure information (e.g. train capacity, walking time distribution) as input, and output the passengers' tap-out time, train load, waiting time, and any other network indicators of interests.



(a) Train arrival

(b) Train departure

FIGURE 1: Event-based Transit Network Loading Model Structure

Figure 1 illustrates the structure of the model. Three objects are defined: train, waiting queue (on platform), and passengers. An event is defined as a train arrival at, or departure from, a station. Events are ordered chronologically. New and transferring passengers join the waiting queue on the platform and board a train based on the FIFB criteria. The number of boarding passengers depends on the available train capacity.

The assignment model works by generating a train event list (arrivals and departures) based on the timetable or the actual train movement data from AVL, and then sequentially processing the ordered events until all events are processed for the time period of interest. To process an individual event,

- If the event is an arrival (Figure 1a), the train offloads passengers and updates its state (e.g. train load and in-vehicle passengers). From the alighting passengers, those who transfer at that station, are assigned to the waiting queues on the corresponding transfer platforms (e.g passengers transferring to platform B in the graph). Passengers who tap-out at this station will be removed from the system. New tap-in passengers who entered the station between two events are added into the queue. Then, the waiting queue objects for these platforms are updated accordingly.
- If the event is a departure (Figure 1b), passengers board trains based on a FIFB boarding priority rule. If the on-board passengers reached the train capacity, the remaining passengers at the platform will be denied boarding. Finally, the state of the train (train load and in-vehicle passengers) and the waiting queue at the platform are updated accordingly.

Problem definition

Consider a general urban rail network within a specific time period T, which is represented as $G = (S, A)$, where S is the set of stations and A is the set of directed links. We divide T into several time intervals with equal width τ. Denote the set of all time intervals as $\mathscr{T} = \{1, 2, ..., T/\tau\}$. We define a concept called *Time-space (TS) node* as i_m, where $i \in S$ and $m \in \mathscr{T}$. i_m represents the state of station i during time interval m. Considering an OD pair (i, j), we denote the *OD entry flow* as $q^{i_m,j}$, which represents the number of passengers entering at station i during time interval m and exiting at station j. $q^{i_m,j}$ is the OD demand input for TNL model. Another variable related to OD entry flow is the *OD exit flow*, denoted as q^{i,j_n}, which represents the number of passengers exit at station j during time interval n with origin i. q^{i,j_n} is the output of TNL model. Importantly, the ground truth q^{i,j_n} can be obtained from the AFC data. Therefore, q^{i,j_n} can be used to calibrate the route choice.

We denote the ground truth OD exit flow as $\tilde{q}^{i,j_n}$. Then $\sum_{i,j_n}(q^{i,j_n} - \tilde{q}^{i,j_n})^2$ (square difference) can be one term of the objective function which we want to minimize. However, q^{i,j_n} only captures the information of exit demand volume. The entry time information is not included. It is possible that the model can output similar q^{i,j_n} but the flows may come from different entry time compared to the ground truth. To capture the entry time information, we introduce another parameter named *journey time distribution* (JTD). For all passengers belong to q^{i,j_n}, we denote the JTD for origin i, destination j and exit time interval n as $p_{i,j_n}(x)$. $p_{i,j_n}(x)$ can be calculated by the kernel density method (*25*) given the passengers' journey time samples. Therefore, we can output $p_{i,j_n}(x)$ from the TNL model. And the ground truth JTD can also be obtained from AFC data, which is represented by $\tilde{p}_{i,j_t}(x)$. We can formulate the difference of two distributions as Kullback-Leibler (KL) divergence (D_{KL}), that is:

$$D_{\mathrm{KL}}(p_{i,j_n}(x)||\tilde{p}_{i,j_n}(x)) = \int_x p_{i,j_n}(x) \cdot \log \frac{p_{i,j_n}(x)}{\tilde{p}_{i,j_n}(x)} \mathrm{d}x. \tag{1}$$

To avoid the unstable estimation of $p_{i,j_t}(x)$, only the OD pairs with more than 50 passengers exit in a specific time interval are considered for D_{KL} calculation (i.e. $q^{i,j_n} > 50$). Denote the set of corresponding OD pairs and exit time intervals as $\mathscr{D}$, where $\mathscr{D} = \{(i, j_n) \mid q^{i,j_n} > 50\}$. Then, $\sum_{i,j_n \in \mathscr{D}} D_{\text{KL}}(p_{i,j_t}(x) || \tilde{p}_{i,j_t}(x))$ can be another item in the objective function. Therefore, we can formulate the route choice estimation problem as the following.

$$
\begin{aligned}
\min_{\beta} \quad & w_1 \sum_{i,j_n} (q^{i,j_n} - \tilde{q}^{i,j_n})^2 + w_2 \sum_{i,j_n \in \mathscr{D}} D_{\text{KL}}(p_{i,j_n}(x) || \tilde{p}_{i,j_n}(x)) \\
\text{s.t.} \quad & q^{i,j_n} = \text{TNL}(\beta, q^{i_m,j}, \theta) \qquad \forall i, j_n, \\
& p_{i,j_n}(x) = \text{TNL}(\beta, q^{i_m,j}, \theta) \qquad \forall i, j_n \in \mathscr{D}, \\
& L_\beta \leq \beta \leq U_\beta
\end{aligned} \tag{2}
$$

where β is the parameters of route choice model; L_β and U_β are the predefined lower bound and upper bound of β. These variables and parameters will be defined and explained in the model assumption section. w_1 and w_2 are the weights to balance the scale of two terms. θ is the external parameters for the TNL model, including time table (or AVL data), transit network typology, access/egress/transfer time, and train capacity. Since the TNL model has no analytic form, Eq. 2 is a bound-constrained SBO problem. In the following section, we will show five different algorithms which belongs to three categories of SBO methods to solve this problem.

Model assumption

Two major assumptions are made for the model and are presented below. First, we assume route choice behavior can be formulated as a C-logit model (*26*), which is an extension of multinomial logit (MNL) model. The choice fraction of route r for OD pairs (i, j) in time interval m can be formulated as below.

$$
p_r^{i_m,j} = \frac{\exp(\beta_X \cdot X_{r,m} + \beta_{CF} \cdot CF_r)}{\sum_{r' \in \mathscr{R}_{ij}} \exp(\beta_X \cdot X_{r',m} + +\beta_{CF} \cdot CF_{r'})} := \frac{\exp(\beta Y_{r,m})}{\sum_{r' \in \mathscr{R}_{ij}} \exp(\beta Y_{r',m})}, \tag{3}
$$

where $X_{r,m}$ is the vector of attributes for route r in time interval m, which include in-vehicle time, number of transfers, transfer walking time and map distance. $\mathscr{R}_{ij}$ is the route set for OD pair (i, j), where $r \in \mathscr{R}_{ij}$. CF_r is the commonality factor of route r which measures the degree of similarity of route r with the other routes of the same OD. β_X and β_{CF} are the corresponding coefficients to be estimated. For simplicity, we define the β and $Y_{r,m}$ as the combination of the two terms in the utility function. The CF_r can be expressed as following.

$$
CF_r = \ln \sum_{r' \in \mathscr{R}_{ij}} \left(\frac{L_{r,r'}}{L_r L_{r'}}\right)^{\gamma}, \tag{4}
$$

where $L_{r,r'}$ is the number of common stations of route r and r'. L_r and $L_{r'}$ are the number of stations for route r and r', respectively. γ is a positive parameter which is assigned to 5 based on empirical studies (*27*). Typical MNL models assume alternatives are independent and irrelevant (IIA). When two different routes have overlapping segments, this assumption will not hold. C-logit model can address the route overlapping problem with the correction term CF, which is widely used in modeling route choices (*28*). Also, it remains the formulation of MNL, which is practical and easy to compute.

For the purpose of parameters inference, we also assume we have a reasonable boundary for all parameters β. The boundary can be obtained by from the prior knowledge and previous survey results. Denote the upper bound as U_β and lower bound as L_β, where U_β and U_β are both vectors with the same cardinality of β. Then we have

$$L_\beta \leq \beta \leq U_\beta, \tag{5}$$

which is added into the constraints in the SBO problem (Eq. 2).

Another set of assumptions are related to the network loading criteria. We assume the following rules.

- When loading a train, passengers waiting at the platform are loaded based on a First-In-First-Board (FIFB) principle.
- Every train has a strict physical capacity. When on-board passengers reach the capacity, the remained passengers will be denied boarding and wait in the platform for next available train.
- All transit services arrive on time. Timetable is sufficiently reliable and can be considered as deterministic (*29*). This assumption can be relaxed when the automated vehicle location (AVL) data is available, which can provide the ground-truth train arrival and departure time.
- The distribution of access walking time, egress walking time and transfer walking time are known.
- The platform has infinite capacity to serve waiting passengers

All these network loading criteria have been incorporated into the TNL model.

Simulation-based optimization algorithms

In general, there are three classes of methods for the SBO, including the direct search method, the gradient-based method, and the response surface method (*13, 30*). Direct search can be defined as the sequential examination of trial points generated by a certain strategy. Then it compares the function values of these trial points directly without approximating derivatives. These methods remain attractive as they are easy to describe and implement. More importantly, it is suitable for objective functions where gradients does not exist everywhere. Gradient-based approaches (or stochastic approximation method) attempt to optimize the function values using estimated gradient information. These methods aim to imitate the steepest descent methods in derivative-based optimization. Finite difference schemes can be used to estimate gradients. Response surface methodology is useful in the context of continuous optimization problems. It focuses on learning input-output relationships to approximate the underlying simulation by a pre-defined functional form (also known as a meta-model or surrogate model). This functional form can then be used for optimization leveraging powerful derivative-based optimization techniques.

In this study, we applied five different algorithms belonging to these three classes of SBO methods to solve the aforementioned route choice estimation problem. The summary of all algorithms is described in Table 1. The details of all algorithms will be introduced in the following.

Nelder-Mead Simplex Algorithm

The Nelder-Mead Simplex Algorithm (NMSA) is a simplex method for finding a local minimum of the objective function Nelder and Mead (*36*). NMSA in n dimensions maintains a set of $n+1$ test points arranged as a *simplex*. Denote the initial value of β as β^{ini}, the initial simplex sets $\{\beta_0, \beta_1, ..., \beta_n\}$ is generated as:

TABLE 1: Algorithms Summary

Category	Algorithms	Source
Direct search	Nelder-Mead Simplex Algorithm (NMSA)	Gao and Han (*31*)
	Mesh Adaptive Direct Search (MADS)	Abramson et al. (*32*)
Gradient-based	Simultaneous Perturbation Stochastic Approximation (SPSA)	Spall et al. (*33*)
Response surface	Bayesian Optimization (BYO)	Snoek et al. (*34*)
	Constrained Optimization using Response Surfaces (CORS)	Regis and Shoemaker (*35*)

$$\beta_i = \begin{cases} \beta^{\text{ini}} & \text{if } i = 0 \\ \beta^{\text{ini}} + \sigma \cdot e_i & \text{otherwise} \end{cases} \tag{6}$$

where e_i is the unit vector in the i th coordinate, σ is the step-size which is equal to 0.05 in this study (*31*). Based on the initial simplex, the model will evaluate the objective function for each test point, in order to find a new test point to replace one of the old test points, and so the technique progresses. The new candidate can be generated through simplex centroid reflections, contractions or other means depending on the function value of the test points. The process will generate a sequence of simplex, for which the function values at the vertices get smaller and smaller. The size of the simplex is reduced and finally the coordinates of the minimum point are found.

Four possible operations: reflection, expansion, contraction, and shrink are associated with the corresponding scalar parameters: α_1 (reflection), α_2 (expansion), α_3 (contraction) and α_4 (shrink). In this study, we set the value of these parameters as $\{\alpha_1, \alpha_2, \alpha_3, \alpha_4\} = \{1, 2, 0.5, 0.5\}$. The algorithm is implemented by the Python `scikit-learn` package. Since NMSA is designed for unconstrained problem, we ignore the boundary of β for this algorithm, which turns out not to affect the results because the searching trajectories of β are found all located in the boundary. More details regarding the NMSA can be referred to Gao and Han (*31*).

Mesh Adaptive Direct Search

The MADS algorithm is a directional direct search framework for nonlinear optimization (*37*). Briefly, MADS seeks to improve the current solution by testing points in the neighborhood of the current point (the incumbent). The neighborhood points are generated by moving one step in each direction from the incumbent on an iteration-dependent mesh. Each iteration of MADS comprises of a SEARCH stage and an optional POLL stage. The SEARCH stage evaluates a finite number of points proposed by the searching strategy (e.g. moving one step around from current point). Whenever the SEARCH step fails to generate an improved mesh point, the POLL step is invoked. The POLL step conducts local exploration near the current incumbent, which also intends to find an improved point on the mesh. Once an improved point is found, the algorithm will update the current point and construct a new mesh. According to Audet and Dennis Jr (*37*), the mesh size parameters will approach zero as the number of iteration approaches to infinity, which demonstrates the convergence of MADS algorithm.

In this paper, we use a variant of MADS method called ORTHO-MADS, which leverages a special orthogonal positive spanning set of polling directions. More details regarding the algorithm can be found in Abramson et al. (*32*). The NOMAD 3.9.1 (*38*) with Python interface is adopted for the MADS algorithm application. The hyper-parameters are tuned based on the NOMAD user guide. The direction type is set as orthogonal, with $N+1$ directions generated at each poll, where N is the number of estimated parameters (i.e. $N = |\beta|$). Latin Hyper-cube search is not applied.

Simultaneous Perturbation Stochastic Approximation

SPSA is a descent method for finding local minimum. It approximates the gradient with only two measurements of the objective function, regardless of the dimension of the optimization problem, which is also called second-order SPSA. Denote the objective function in Eq. 2 as $Z(\beta)$. The estimated route choice parameters in the k th iteration is denoted as $\beta^{(k)}$. Then one iteration for the SPSA is performed as

$$\beta^{(k+1)} = \beta^{(k)} - a_k \cdot \hat{\nabla} Z(\beta^{(k)}) \tag{7}$$

where

$$\tilde{\nabla} Z(\beta^{(k)}) = \frac{Z(\beta^{(k)} + c_k \Delta_k) - Z(\beta^{(k)} - c_k \Delta_k)}{2 c_k \Delta_k} \tag{8}$$

$$a_k = \frac{a}{(k+1+A)^{\alpha}} \tag{9}$$

$$c_k = \frac{c}{(k+1)^{\gamma}} \tag{10}$$

Δ_k is a random perturbation vector, whose elements are obtained from a Bernoulli distribution with the probability parameter equal to 0.5. $\{\alpha, \gamma, a, c, A\}$ are tuned as $\{0.602, 0.101, 0.01, 0.03, 0.1 \times$ maximum number of iterations$\}$ in this study according to the numerical tests and the empirical studies (*39*).

Bayesian Optimization

BYO aims to constructs a probabilistic model for the objective function and then exploits this model to determine where to evaluate the objective function for next step. The philosophy of BYO is to use all of the information available from previous evaluations, instead of simply relying on local gradient and Hessian approximations, which is expected to find the minimum of difficult non-convex functions with relatively few evaluations.

To perform the Bayesian optimization, we must choose a prior distribution for the objective function values, and an acquisition function, which is used to determine the next point to evaluate. In this study, we choose the Gaussian process prior due to its flexibility and tractability. As for the acquisition function, we numerically compare three popular criteria including probability of improvement (POI), expected improvement (EI), and upper confidence bound (UCB) (*34*). Finally the EI criterion is used in this route choice estimation problem due to its best performance. More details regarding the BYO can be found in Snoek et al. (*34*).

Constrained Optimization using Response Surfaces

CORS is a response surface methods for global optimization. In each iteration, it updates the response surface model based on all previously probed points, and selects the next point to evaluate.

The principles for next points selection are: (a) finding new points that have lower objective function value, and (b) improving the fitting of response surface model by sampling feasible regions where little information exists. Hence, the next point is selected by solving the minimization problem of current response surface function subject to constraints that the next point should be more than a certain distance from all previous points (*35*).

Any algorithms follow the CORS framework requires two components: (a) a scheme for selecting an initial set of points for objective function evaluation, and (b) a procedure for globally approximating the objective function (i.e. a response surface model). In this study, the initial sampling is conducted by the Latin hypercube methods, with the initial sampling number equal to 0.1 $\times$ the total number of function evaluations. The radial basis function (RBS) is used as the response model. For the subsequent sampling, a modified version of CORS algorithm with space re-scaling is used. Details about the algorithm can be found in Regis and Shoemaker (*35*) and Knysh and Korkolis (*40*).

CASE STUDY AND NUMERICAL RESULTS

For the purpose of model illustration, we apply the proposed modeling framework on Hong Kong MTR network.

Hong Kong MTR Network

MTR is the operator of Hong Kong urban rail network, which provides services for the urbanized areas of Hong Kong Island, Kowloon, and the New Territories. The system currently consists of 11 lines with 218.2 km (135.6 mile) of rail, serving 159 stations including 91 heavy rail stations and 68 light rail stops. It uses a smart card fare-payment system named Octopus, which serves over 5 million trips on an average weekday. For the urban heavy rail lines, trip transactions are recorded when passengers enter and exit the system, giving the information of the tap-in and tap-out stations and timestamps. The map for the Hong Kong MTR system is shown in Fig 2. In this study, we remove the airport express and light rail transit services. Passengers who enter the metro system from these two services need to re-tap-in. So the remaining network is still a closed system with full OD information.

Case Study Settings

We use AFC data on March 16th (Thursday), 2017 for the model application. The route sets for each OD pair are obtained from the MTR operation team. For route choice behavior, we assume the following attributes can quantify route utitilies: (a) total in-vehicle time, (b) number of transfer times, (c) total transfer walking time, (d) total map distance, and (e) the commonality factor (Eq. 4). Since the evening peak is the most congested period for MTR system and it is highly interested by the transportation agency, we only consider the period from 18:00 to 19:00 for model application. For simplicity, we assume the route choice fractions are static during this hour, which means only one set of β is used. We set the weights in the objective function of Eq. 2 as $w_1 = 1$ and $w_2 = 50$ to balance the scale of two terms. β^{ini} is set as $(L_\beta + U_\beta)/2$. The system parameters θ of TNL model are summarized in Table 2.

Since the real-world route choice information are usually unavailable, it is common to quantitatively validate models with synthetic data. To generate the synthetic data, we first extract the OD entry flow ($q^{i_m,j}$) from the real-world AFC records. A group of route choice parameters are chosen based on a previous empirical study (*41*), which are assumed to be the "true" passenger

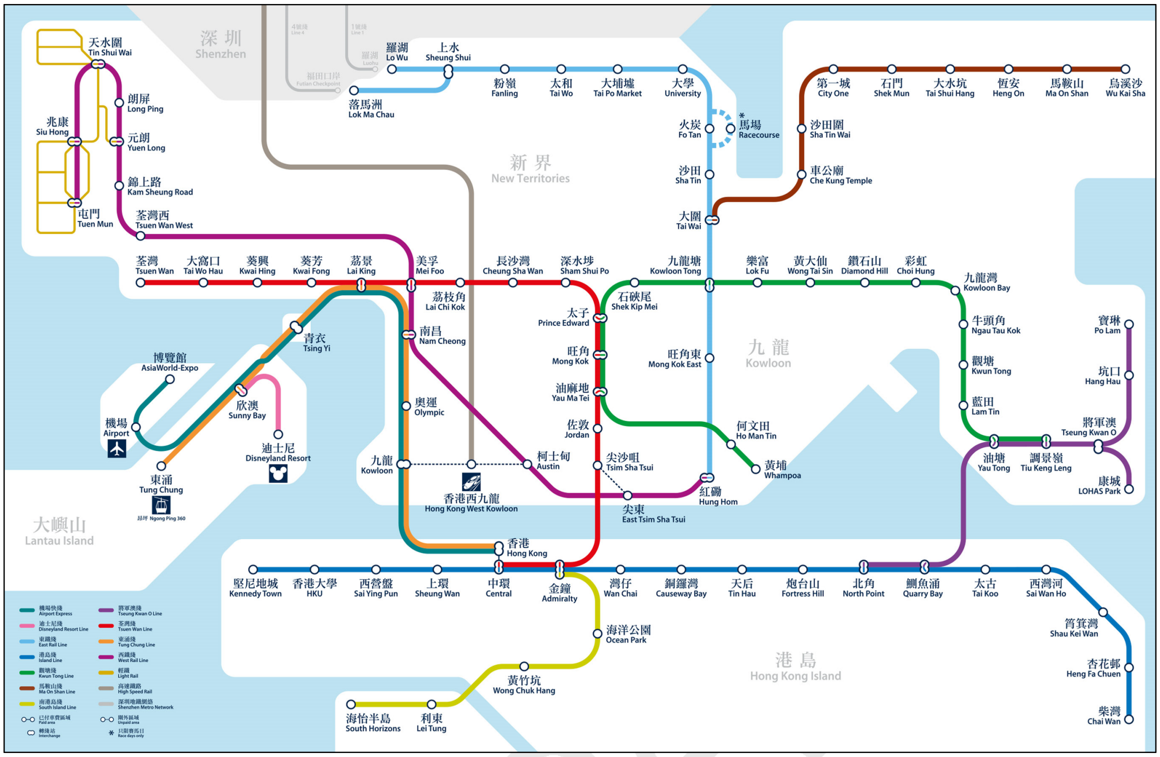


FIGURE 2: Hong Kong MTR Metro System Map

travel behavior parameters (called synthetic β hereafter). Then, we use the network loading model (*27*) with the true OD entry flow and the synthetic β as input to simulate the travel of passengers in the system, and record people's tap-in and tap-out time. The records of all people's tap-in and tap-out time are treated as *synthetic AFC data*. For model validation, we can apply the proposed model to the synthetic AFC data and compare the estimated β and synthetic β. The value of synthetic data can be found in Table 3. We set $L_\beta = [-0.1, -1, -1, -0.1, -1]$ and $U_\beta = [0, 0, 0, 0, 0]$ accordingly. The sequence of vector elements is same to the sequence in Table 3. To compare different algorithms, a fixed computational budget, 100 function evaluations, is applied to all algorithms. All algorithms except for NMSA (deterministic algorithm) are replicated for 10 times (with different random seed) to decrease the impact of randomness. As this paper targets on comparative analysis of SBO algorithms, we use a fixed random seed for the TNL model to eliminate the random error brought from simulation. This means the objective function value of the synthetic β will be 0, which is helpful to focus on the performance of different SBO methods.

Numerical results

The convergence results of all algorithms are depicted in Figure 3. Each point represents the average value of all replications. The standard deviations are shown by vertical lines. We found that all algorithms can converge to a reasonable stage with relatively small objective function given the limited number of function evaluations.The response surface methods (BYO and CORS) have the fastest convergence speed. They can also reach the lowest objective function value, especially for the CORS algorithm. This is corresponding to previous point of views regarding the SBO methods in the transportation domains (*13, 21*). In terms of algorithm stability, NMSA is a deterministic

TABLE 2: Summary of TNL External Parameters θ

Variables	Description	Source
Acess/Egress walking time	Platform-specific with mean and deviation	MTR field measurement
Transfer walking time	Platform-specific with mean and deviation	MTR field measurement
Time table	True time table for the test date, which also includes the number of cars for each train. Future research can use AVL data to get real-world train arrival and departure information	MTR operation team
Capacity	230 passengers per car.	MTR service standard
Warm up and Cool down	60 minutes warm up and cool down time	Mo et al. (*8*)

algorithm, which is not affected by the randomness. All other algorithms have big uncertainty at the first half iterations. As the number of function evaluations increase, the standard deviations will decrease and the results become stable. MADS is the most unstable algorithm in this study. This may be due to that MADS can probably converge to some non-stationary points (*42*). As for response surface methods, although BYO and CORS have similar performance in terms of final objective function, CORS is much more stable than BYO. This is reasonable becasue BYO is based on probabilistic models, where function evaluations contain more uncertainty.

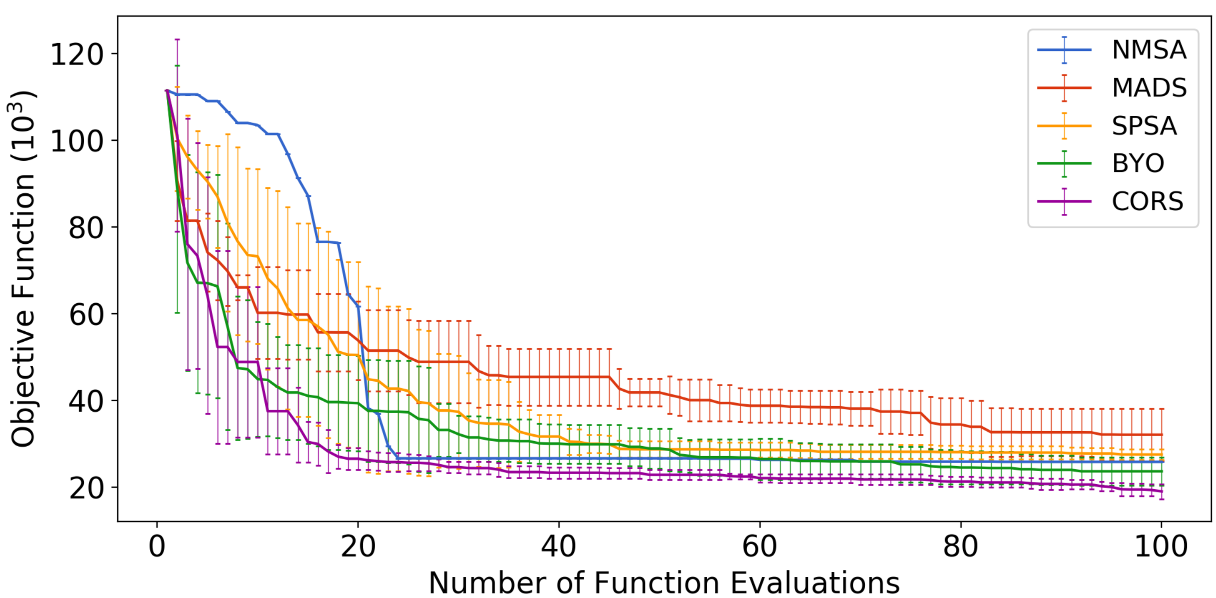


FIGURE 3: Convergence Results of SBO Algorithms.

Table 3 compares the best estimated parameters of different algorithms with the synthetic ones. Despite some algorithms can reach similar objective function value, they can output different results of β. For example, NMSA has good estimation of coefficients of in-vehicle time, number of transfers, but bad estimation in the commonality factor. SPSA also shows the similar properties. While CORS and BYO has the good estimation in all coefficients. Overall, the results demonstrate all algorithms can output reasonable estimation of route choice parameters, which means the proposed SBO framework using AFC data is effective.

TABLE 3: Optimal β Estimation Results

Variable Name	Synthetic	Estimated				
		NMSA	MADS	SPSA	BYO	CORS
In-vehicle time	-0.0663	-0.0656	-0.0542	-0.0693	-0.0623	-0.0645
Number of transfers	-0.438	-0.430	-0.310	-0.301	-0.464	-0.445
Transfer walking time	-0.183	-0.143	-0.180	-0.213	-0.170	-0.184
Map distance	-0.0767	-0.0639	-0.100	-0.0946	-0.0792	-0.0739
Commonality factor	-0.941	-0.676	-0.900	-0.676	-0.948	-0.969
Objective function	0	25795.9	24447.22	25092.22	17551.51	16300.0

CONCLUSION AND DISCUSSION

In this paper, we proposed a SBO framework to estimate the route choice behaviors in metro systems using AFC data. Five different algorithms which cover three main brunches of SBO methods are applied and compared in this study. The advantage of this framework lies in the incorporation of TNL model in route choice estimation, which can consider the internal correlation of denied boarding probabilities among different stations. We applied the proposed framework on the Hong Kong MTR network, and compared the performance of different algorithms, which illustrates the effectiveness of our methods.

The developments in this paper have been focused on a general framework, while the models and examples we presented in this paper still have some limitations. First, we only validate the framework on synthetic AFC data, which ignores the noise and some uncertainties in the real-world. This is caused by the absence of real-world route choice information. Future research can collect the real-world route choice data to conduct more realistic model validation. Second, we imposed a strong assumption on route behavior modeling, that is, only one set of β is applied for the whole network. The real-world route choice behaviour may be more diverse and heterogeneous. Future research can cluster different OD pairs with different β based on the corresponding passengers' characteristics. Third, the estimated parameters in the case study are simplified without considering the temporal dynamics. In practice, to obtain time dependent parameters, we can divide the data set into slices and apply the framework on each of them.

Nevertheless, the overall framework has demonstrated good capability in route choice estimation. As an important component for assignment model, route choice behaviors are always interesting to operators. Applying the inference results on assignment model (*27*), link flow profile can be estimated in temporal scale. The results could also be used to identify critical transfer stations/platforms, providing valuable information to operators to better design and operate the whole metro system.

AUTHOR CONTRIBUTION STATEMENT

The authors confirm contribution to the paper as follows: study conception and design: B. Mo, Z. Ma, H.N. Koutsopoulos, J. Zhao; data collection: B. Mo, Z. Ma, H.N. Koutsopoulos, J. Zhao; analysis and interpretation of results: B. Mo, Z. Ma, H.N. Koutsopoulos; draft manuscript preparation: B. Mo, Z. Ma. All authors reviewed the results and approved the final version of the manuscript.

ACKNOWLEDGEMENTS

The authors would like to thank Hong Kong Mass Transit Railway (MTR) for their support and data availability for this research. The authors also thank Mr. Nate Bailey for reviewing the paper.

REFERENCES

1. Kusakabe, T., T. Iryo, and Y. Asakura, Estimation method for railway passengers' train choice behavior with smart card transaction data. *Transportation*, Vol. 37, No. 5, 2010, pp. 731–749.
2. Zhou, F. and R.-h. Xu, Model of passenger flow assignment for urban rail transit based on entry and exit time constraints. *Transportation Research Record*, Vol. 2284, No. 1, 2012, pp. 57–61.
3. Kumar, P., A. Khani, and Q. He, A robust method for estimating transit passenger trajectories using automated data. *Transportation Research Part C: Emerging Technologies*, Vol. 95, 2018, pp. 731–747.
4. Sun, Y. and R. Xu, Rail transit travel time reliability and estimation of passenger route choice behavior: Analysis using automatic fare collection data. *Transportation Research Record*, Vol. 2275, No. 1, 2012, pp. 58–67.
5. Sun, L., Y. Lu, J. G. Jin, D.-H. Lee, and K. W. Axhausen, An integrated Bayesian approach for passenger flow assignment in metro networks. *Transportation Research Part C: Emerging Technologies*, Vol. 52, 2015, pp. 116–131.
6. Zhao, J., F. Zhang, L. Tu, C. Xu, D. Shen, C. Tian, X.-Y. Li, and Z. Li, Estimation of passenger route choice pattern using smart card data for complex metro systems. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 18, No. 4, 2017, pp. 790–801.
7. Xu, X., L. Xie, H. Li, and L. Qin, Learning the route choice behavior of subway passengers from AFC data. *Expert Systems with Applications*, Vol. 95, 2018, pp. 324–332.
8. Mo, B., Z. Ma, H. Koutsopoulos, and J. Zhao, Assignment-based Path Choice Calibration for Metro Systems Using Smart Card Data, 2019, working paper.
9. Zhu, Y., *Passenger-to-itinerary assignment model based on automated data*. Ph.D. thesis, Northeastern University, 2017.
10. Song, W., K. Han, Y. Wang, T. Friesz, and E. Del Castillo, Statistical metamodeling of dynamic network loading. *Transportation research procedia*, Vol. 23, 2017, pp. 263–282.
11. Chen, X. M., C. Xiong, X. He, Z. Zhu, and L. Zhang, Time-of-day vehicle mileage fees for congestion mitigation and revenue generation: A simulation-based optimization method and its real-world application. *Transportation Research Part C: Emerging Technologies*, Vol. 63, 2016, pp. 71–95.
12. He, X., X. Chen, C. Xiong, Z. Zhu, and L. Zhang, Optimal time-varying pricing for toll roads under multiple objectives: a simulation-based optimization approach. *Transportation Science*, Vol. 51, No. 2, 2016, pp. 412–426.

13. Osorio, C. and M. Bierlaire, A simulation-based optimization framework for urban transportation problems. *Operations Research*, Vol. 61, No. 6, 2013, pp. 1333–1345.
14. Osorio, C. and K. Nanduri, Urban transportation emissions mitigation: Coupling high-resolution vehicular emissions and traffic models for traffic signal optimization. *Transportation Research Part B: Methodological*, Vol. 81, 2015, pp. 520–538.
15. Osorio, C. and K. Nanduri, Energy-efficient urban traffic management: a microscopic simulation-based approach. *Transportation Science*, Vol. 49, No. 3, 2015, pp. 637–651.
16. Chong, L. and C. Osorio, A simulation-based optimization algorithm for dynamic large-scale urban transportation problems. *Transportation Science*, Vol. 52, No. 3, 2017, pp. 637–656.
17. Zhang, W. and W. Xu, Simulation-based robust optimization for the schedule of single-direction bus transit route: The design of experiment. *Transportation Research Part E: Logistics and Transportation Review*, Vol. 106, 2017, pp. 203–230.
18. Cardin, M.-A., Y. Deng, and C. Sun, Real options and flexibility analysis in design and management of one-way mobility on-demand systems using decision rules. *Transportation Research Part C: Emerging Technologies*, Vol. 84, 2017, pp. 265–287.
19. Noordhoek, M., W. Dullaert, D. S. Lai, and S. de Leeuw, A simulation–optimization approach for a service-constrained multi-echelon distribution network. *Transportation Research Part E: Logistics and Transportation Review*, Vol. 114, 2018, pp. 292–311.
20. Dong, J.-X. and D.-P. Song, Container fleet sizing and empty repositioning in liner shipping systems. *Transportation Research Part E: Logistics and Transportation Review*, Vol. 45, No. 6, 2009, pp. 860–877.
21. Cheng, Q., S. Wang, Z. Liu, and Y. Yuan, Surrogate-based simulation optimization approach for day-to-day dynamics model calibration with real data. *Transportation Research Part C: Emerging Technologies*, Vol. 105, 2019, pp. 422–438.
22. Mo, B., Z. Ma, H. Koutsopoulos, and J. Zhao, Network Performance Model for Urban Rail Systems, 2019, working paper.
23. Yao, X., B. Han, D. Yu, and H. Ren, Simulation-based dynamic passenger flow assignment modelling for a schedule-based transit network. *Discrete Dynamics in Nature and Society*, Vol. 2017, 2017.
24. Ma, Z., H. N. Koutsopoulos, Y. Chen, and N. H. Wilson, Estimation of denied boarding in urban rail systems: alternative formulations and comparative analysis. *Transportation Research Record*, 2019, p. 0361198119857034.
25. Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in Python. *Journal of machine learning research*, Vol. 12, No. Oct, 2011, pp. 2825–2830.
26. Cascetta, E., A. Nuzzolo, F. Russo, and A. Vitetta, A modified logit route choice model overcoming path overlapping problems. Specification and some calibration results for interurban networks. In *Proceedings of The 13th International Symposium On Transportation And Traffic Theory*, 1996.
27. Mo, B., Z. Ma, H. Koutsopoulos, and J. Zhao, Network Performance Model for Urban Rail Systems, 2019, working paper.
28. Prato, C. G., Route choice modeling: past, present and future research directions. *Journal of choice modelling*, Vol. 2, No. 1, 2009, pp. 65–100.

29. Hamdouch, Y. and S. Lawphongpanich, Schedule-based transit assignment model with travel strategies and capacity constraints. *Transportation Research Part B: Methodological*, Vol. 42, No. 7-8, 2008, pp. 663–684.

30. Amaran, S., N. V. Sahinidis, B. Sharda, and S. J. Bury, Simulation optimization: a review of algorithms and applications. *Annals of Operations Research*, Vol. 240, No. 1, 2016, pp. 351–380.

31. Gao, F. and L. Han, Implementing the Nelder-Mead simplex algorithm with adaptive parameters. *Computational Optimization and Applications*, Vol. 51, No. 1, 2012, pp. 259–277.

32. Abramson, M. A., C. Audet, J. E. Dennis Jr, and S. L. Digabel, OrthoMADS: A deterministic MADS instance with orthogonal directions. *SIAM Journal on Optimization*, Vol. 20, No. 2, 2009, pp. 948–966.

33. Spall, J. C. et al., Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE transactions on automatic control*, Vol. 37, No. 3, 1992, pp. 332–341.

34. Snoek, J., H. Larochelle, and R. P. Adams, Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, 2012, pp. 2951–2959.

35. Regis, R. G. and C. A. Shoemaker, Constrained global optimization of expensive black box functions using radial basis functions. *Journal of Global optimization*, Vol. 31, No. 1, 2005, pp. 153–171.

36. Nelder, J. A. and R. Mead, A simplex method for function minimization. *The computer journal*, Vol. 7, No. 4, 1965, pp. 308–313.

37. Audet, C. and J. E. Dennis Jr, Mesh adaptive direct search algorithms for constrained optimization. *SIAM Journal on optimization*, Vol. 17, No. 1, 2006, pp. 188–217.

38. Audet, C., S. Le Digabel, and C. Tribes, NOMAD user guide. *Rapport technique*, 2009.

39. Gomez-Dans, J., *A Simultaneous Perturbation Stochastic Approximation optimisation code in python*. `https://github.com/jgomezdans/spsa#spsa`, 2012.

40. Knysh, P. and Y. Korkolis, Blackbox: A procedure for parallel optimization of expensive black-box functions. *arXiv preprint arXiv:1605.00998*, 2016.

41. Li, W., Route and Transfer Station Choice Modeling in the MTR System, 2014, working paper.

42. Abramson, M. A. and C. Audet, Convergence of mesh adaptive direct search to second-order stationary points. *SIAM Journal on Optimization*, Vol. 17, No. 2, 2006, pp. 606–619.