Estimation of Passenger Route Choice Pattern Using Smart Card Data for Complex Metro Systems

Juanjuan Zhao, Fan Zhang, *Member, IEEE*, Lai Tu, Chengzhong Xu, *Fellow, IEEE*, Dayong Shen, Chen Tian, Xiang-Yang Li, *Fellow, IEEE*, and Zhengxi Li

***Abstract*—Metro systems play an important role in meeting the demand for urban transportation in large cities. The understanding of passenger route choice is critical for public transit management. The wide deployment of automated fare collection (AFC) systems opens up a new opportunity. However, only each trip's tap-in and tap-out time stamp and stations can be directly obtained from AFC system records; the train and route chosen by a passenger are unknown, information necessary to solve our problem. While existing methods work well in some specific situations, they hardly work for complicated situations. In this paper, we propose a solution that needs no additional equipment or human involvement than the AFC systems. We develop a probabilistic model that can estimate from empirical analysis how the passenger flows are dispatched to different routes and trains. We validate our approach using a large-scale data set collected from the Shenzhen Metro system. The measured results provide us with useful input when building the passenger path choice model.**

***Index Terms*—Metro systems, smart card, data mining, intelligent transportation systems, route choice.**

Manuscript received December 16, 2015; revised March 25, 2016; accepted June 25, 2016. This work was supported by the China National Basic Research Program (973 Program, No. 2015CB352400), by the National Natural Science Foundation of China (U1401258, 6160050472, 0326), by Science and Technology Planning Project of Guangdong Province (2015B010129011), by the Basic Research Program of Shenzhen (JCYJ20150630114942313, JCYJ20140610152828686), by Natural Science Foundation of Guangdong Province (2015A030310326, 2016A030313183), by NSF under Grant CMMI 1436786 and CNS 1526638, by Research Program of Shenzhen under Grant JSGG20150512145714248, KQCX2015040111035011 and CYZZ20150403111012661, and by the Fundamental Research Funds for the Central Universities. The Associate Editor for this paper was Z. Ding. *(Corresponding author: Fan Zhang.)*

J. Zhao and F. Zhang are with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China, and also with Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: jj.zhao@siat.ac.cn; zhangfan@siat.ac.cn).

L. Tu is with the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: tulai.net@gmail.com).

C. Xu is with Wayne State University, Detroit, MI 48202 USA, and also with Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: cz.xu@siat.ac.cn).

D. Shen is with the Research Center for Computational Experiments and Parallel Systems, National University of Defense Technology, Changsha 410073, China (e-mail: dayong.shen@nudt.edu.cn).

C. Tian is with State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China (e-mail: tianchen@nju.edu.cn).

X.-Y. Li is with the School of Computer Science and Technology, University of Science and Technology of China, Hefei 230027, China (e-mail: xiangyang.li@gmail.com).

Z. Li is with the Department of Automation, North China University of Technology, Beijing 100144, China (e-mail: lizhengxi@ncut.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2016.2587864

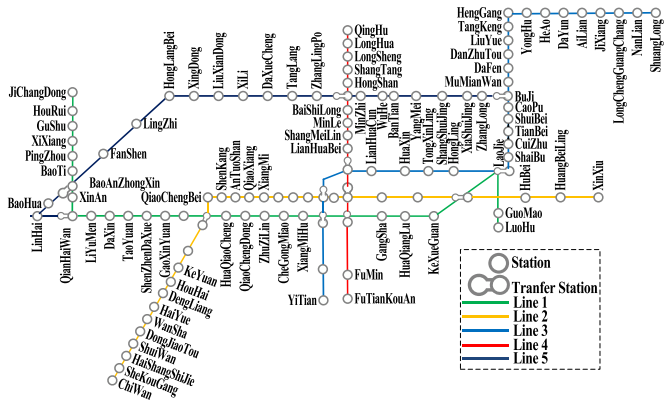


Fig. 1. Map of the Metro of Shenzhen.

I. Introduction

NOWADAYS, metro systems play an important role in meeting the urban transportation demand in large cities. Due to its fast speed, high efficiency, large volume and punctuality, the urban metro has become the first choice of many people. In Shenzhen, China, in mid-June 2015, there were around 3.5 million metro trips every day, which was around one third of the total public traffic. Fig. 1 illustrates the metro operating map of Shenzhen. With further expansion of the metro system, the amount of passengers may increase rapidly. On one hand, the increasing usage of metros can effectively help reduce the traffic pressure on surface roads. On the other hand, it also brings dramatic increasing of passenger demand on metro systems.

The traffic patterns of large metro systems are usually very complex. Under the condition of network operation and seamless transfer in current metro systems, the train and route chosen by a passenger are unknown. It is common to have more than one route between the origin station and the destination station, a.k.a *multi-path* in transportation systems. As shown in Fig. 2(a), there are two routes from station O to station D. This means that for an OD pair with more than one route, we don't know how passengers are distributed over these routes and trains.

This missing information at a fine granularity could be important for both passengers and metro operators. From the operators' point of view, understanding the flow distribution of passengers in the whole metro network is important for improving the service reliability. The potential applications can be a mobile application of trip planning for metro passengers,

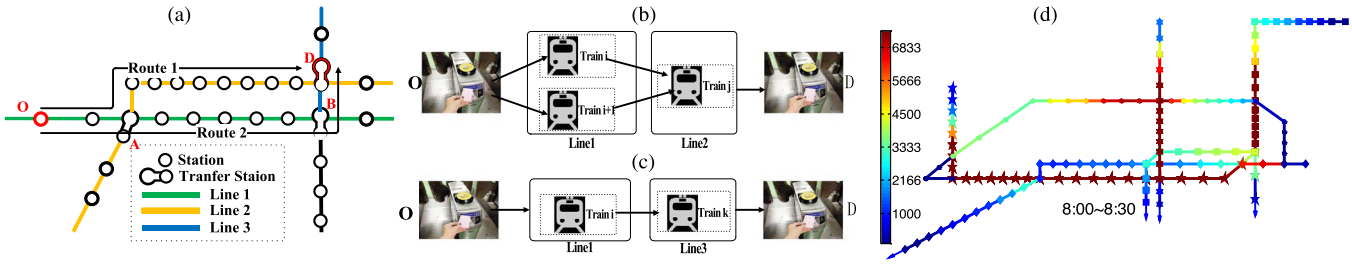


Fig. 2. (a) OD with multiple routes. (b) Trains matching for route 1. (c) Trains matching for route 2. (d) Illustration of traffic monitor application based on the proposed model.

a monitoring system for metro operators, a route suggestion and emergency management system for urban administrators etc. This paper aims to develop a solution to calculate the probability of each route chosen for an OD pair, which can be used to estimate the passengers flow at a granularity of trains of each line, as shown in Fig. 2(d).

Traditional approaches are not scalable. To understand the route choice behavior if passengers, one of traditional methods is to conduct field surveys at train stations, by asking passengers which route they take to reach their destinations. There are limitations of this method: firstly, most surveys are conducted with focus on a part of the passengers at particular locations within a limited time window, hence the results are often limited in diversity, scale and accuracy; secondly, it is both labor-intensive and time-consuming in conducting such surveys.

The wide deployment of Automated Fare Collection(AFC) systems opens up a new opportunity for metro network analysis: the transaction records from AFC can reveal the Origin (O) and the Destination (D) of every passenger's trip, as passengers are required to tap their smart cards or RFID based tickets each time they enter the O station or exit the D station. Passengers' flows can be coarsely demonstrated by OD (origin-destination) pairs. However, AFC records failed to expose the passengers' routes directly. Even in cases that the route of an OD is unique, the AFC records are still not able to show which train a passenger takes. There are too many factors that can affect a passenger's final plan, i.e., trains or train combinations one takes. For example, if the train fails to have enough capacity to accommodate all passengers waiting on a platform, some passengers would have to wait for another train. This phenomenon, known as "travelers left behind" is quite common during rush hours or at large stations. There are already some studies using transaction records from AFC to understand the passengers' route and train choice behavior [1], [2]. Although these methods work well in some specific situations, they don't work for complicated situations, such as the case where there are various "left-behinds" at different stations caused by the imbalance of geographical distribution of passengers. Also, usually the walking time between the charge gate and that platform, and the walking time for transfer between platforms could not be ignored.

In this paper, we propose a solution that needs no additional facility than the trains' operating time table and the AFC records data. By matching a passenger's smart card records with the trains operating time table, the route that he/she might choose can be narrowed down. We develop a probabilistic model that can empirically estimate how the passenger flows are distributed among different routes and trains. As a concrete example in Fig. 2(b) and (c), if a passenger taps-in at station O at time point $T1$ and taps-out at station D at time point $T2$, both Route 1 and Route 2 can be the possible choice after we narrow down the possible plan based on the time table. Our solution is to answer that at what probability each route is chosen by the passengers. For a route like Route 1, which further has multiple possible train combinations that satisfy the time table constraints, we further derive the probability that passengers may choose each plan, i.e., $\{\mathrm{tr}_i, \mathrm{tr}_j\}$ or $\{\mathrm{tr}_{i+1}, \mathrm{tr}_j\}$.

The contributions of this paper include:

- We define two kinds of time-dependent polynomial distributions of the number of trains waited for by passengers. The first is the number of trains that a passenger waits at his/her original station. The other is the number of trains a passenger waits when he/she transits at the transfer station. A set of algorithms are proposed to calculate the parameters of the two distributions.
- We further propose a probabilistic model that can estimate how the passenger flows are distributed among different routes and trains.
- We then deploy the algorithms on a cloud platform and develop supporting modules for the system level solution.
- Finally we validate our approach using a large scale data set collected from the Shenzhen metro system. The measured results provide us with useful inputs when we build the passenger path choice model.

For the rest of this paper, we discuss the related work in Section II. The overview of this study is given in Section III. Section IV discusses the solution in details. We present system design and the algorithm implementation on a cloud platform in Section V. Section VI presents the experimental studies. Finally, Section VII concludes the paper.

II. Related Work

Location-based services have been attracting much attention in a wide range of applications [3]–[5]. Building users route choice model is an important research direction in the field of transportation [6]–[8], which is the basis for traffic management policies-making. Due to the lack of the observation of how probably each route is chosen for an OD pair with multiple routes, most of the past studies focus on building route choice models from empirical perspective. They assume

that all passengers have full knowledge of the transportation when attempting to minimize some objective functions e.g., minimizing their travel time (user equilibrium) or minimizing the total system travel time (system optimum) [9]–[11]. However, those models depend heavily on behavior assumptions and lack in reliable supporting data. Given the dynamic and stochastic nature of transportation systems, the assumption of the passengers' global knowledge is questionable.

Fortunately, the large amount of smart card data in a long period provide us a great opportunity to analyze passengers' transit behavior and evaluate transit service. There were a few previous studies regarding the utilization of smart card data. The literature [12] considered the potential usage of smart card data for travel. The literature [13] analyzed users' travel behavior using data mining technology, which clustered users into four groups according to their temporal travel patterns. Our recent work [14] studied individual passenger's temporal and spatial travel patterns. We found that if a passenger is temporally regular, it is very possible that the passenger is also spatially regular. Besides understanding travel behavior, smart card data have been used to improve public transit services. The study in [15] gave a comprehensive review of using smart card data from different aspects: strategic, tactical and operational. To improve the resilience to service disruptions of metro systems, the study [16] investigated a practical problem about integrating localized bus service with metro network. Using the same data set, three optimization models were formulated to design demand-driven timetables for a single-track metro service [17].

For understanding passengers' flow assignments in metro system, The authors in [1] proposed a method to estimate which trains every smart card holder boarded during his/her journey. This method could be used to estimate trains' occupancies. However, it was also based on some assumptions that may be only available in some limited scenarios: (a) The methodology assumed that all passengers know the train timetable beforehand. When choosing route, they will first choose the plan with minimum total waiting time, then choose the plan with fewer transits. The remaining small percentage of undecided trips was assumed to be assigned to all possible plans with an equal probability. (b) The walking time between the charge gates and that platform, and the walking time for transfer between platforms are ignored, which may lead to mismatching between passengers' tap time and trains' operation time.

The authors in [18] proposed a linear regression model to analyze the individual trajectory during a metro trip, which could be used to estimate the spatial-temporal density inside a metro system. However, their model focused on a single-track scenario that is oversimplified. The study in [19] used a clustering algorithm to infer the route-use patterns of metro passengers from the smart-card data. It confirmed that a Gaussian mixture model worked well in finding the route shares and the mean and variance of travel times for each route of London underground. But the conclusion based on two preconditions. First, the number of routes used by users must be known in advance. Second, the probability distribution function of travel time of each route must be Gaussian. The study in [2] developed an integrated Bayesian approach to infer both network attributes and passenger route choice behavior in a complex metro system, which worked well in some cases that there are lack of train timetable. But a large set of explanatory variables and the probability distribution of these variables need to be calibrated, such as it assumed that all link costs are characterized by independent normal distributions. This is not always true. Taking the phenomenon "left-behind" as example, the imbalance of geographical distribution of passengers leads to the various "left-behinds" in different stations and results that the time cost does not follow normal distributions in a station for different number of trains that need to be waited for. There are other related work of big data based analysis for smart transportation systems [20]–[23], while they are not targeting metro systems.

In sum, the existing methods did not consider the passengers' "left-behind" in detail, which however is one of the main factors affecting us to understanding passengers' path choice behavior. In this paper, we propose a novel approach to calculate the probability of each route used for an OD with multiple routes, it can be used to complicated traffic situation of complex metro network, especially for the situation that "left-behind" occurs often.

III. Overview

A. Data Set

There are two types of data used in this study, smart card transaction data and train operation data. A smart card transaction record is reported when a passenger passes through the entrance or exit gate by tapping smart card, which includes fields id(unique identifier of smart card), s(metro station), t(transaction time) and type(enter or exit). A train operation record is collected when a train arrives at or departs a station, which includes fields sq(train sequence), l(metro line), s(metro station), t(transaction time) and type(arrive or leave).

For a trip x of a passenger, we can observe the trip's beginning time $x.b$, origin $x.o$, end time $x.e$, and destination $x.d$, by joining the tap-in and tap-out tap events together. If the trip needs $i-1$ transfers, we say the trip has i parts. The first part is from the passenger entering metro system to he/she getting off from his/her first boarded train. The last part is from the passenger getting off from the second last train to he/she exiting metro system. If the passenger doesn't need to transfer, then the first part of the trip is also the last part.

B. Notations and Assumptions

Suppose the set of effective routes of an OD pair is R and $R=\{R_1,R_2,\ldots,R_Z\}$. For simplicity, we divide one day into fixed slots with a time interval δ. We set δ to be a half hour. Then one day can be split as $I=\{I_1,I_2,I_3,\ldots,I_{48}\}$. And we assume that given the interval δ, the probability of each route being chosen is stable in each time interval. We further define that the routes being chosen in a specific time slot I_j obeys polynomial distribution with parameter $\boldsymbol{\alpha}_j=\{\alpha_{j,1},\alpha_{j,2},\ldots,\alpha_{j,Z}\}$ where $\sum_{z=1}^{Z}\alpha_{j,z}=1$. Given the train operation table Tab, the set of trips of passengers $X=\{x^1,x^2,x^3,\ldots,x^Q\}$ that begin at time slot I_j, we aim to calculate $\boldsymbol{\alpha}_j$.

For simplicity, we further present some assumptions and notations that will be used in this paper.

- We assume that the time that most passengers spend to walk between the platform and the ODs' entrance/exit is less than the departure interval between two adjacent trains. This assumption is rational, because for most metro system, the distance between gate and platform is not far.
- We assume that most passengers will exit the metro station through the exit gate as soon as possible after getting off the train that reaches her/his destination.

Based on the two assumptions, we can infer that given a trip of a passenger and the route that he/she chooses, the train that he/she boards in the last part of trip can be determined uniquely.

Tapin(ζ): where $\zeta = (s, l, j)$, represents the passengers who enter metro system at time slot j in station s and chooses metro line l.

Transfer(η): where $\eta = (s, l, l', j)$, represents the passengers who transfer from metro line l to l' in transit station s at the time slot j.

To calculate $\boldsymbol{\alpha}_j$ of an OD pair, all effective routes are needed firstly. Then given all effective routes R, and all trips X starting at time slot I_j from station O to D, and train operation table Tab, we use the maximum likelihood function as Equation (1) to calculate $\boldsymbol{\alpha}_j$, where $\Pr(x^q.e|\text{Tab}, x^q.b, R_z)$ is the possibility that a passenger x^q passes through exit gate at time $x^q.e$ on condition of Tab, $x^q.b$ and the route chosen R_z.

$$
\begin{aligned}
&L(X, \text{Tab}, \boldsymbol{\alpha}_j) \\
&= \log \prod_{x^q \in X} \left(\sum_{R_z \in R} ((\alpha_{j,z} \times \Pr(x^q.e|\text{Tab}, x^q.b, R_z))) \right) \\
&= \log \sum_{x^q \in X} \sum_{R_z \in R} ((\alpha_{j,z} \times \Pr(x^q.e|\text{Tab}, x^q.b, R_z))). \quad (1)
\end{aligned}
$$

In practical, the time cost of a trip ($x.e - x.b$) has a certain relation with train operation data. So given a trip of a passenger, we can find all possible plans (train or trains combination) that the passenger may choose for a route by matching two types of data. So $\Pr(x^q.e|\text{Tab}, x^q.b, R_z)$ can be calculated by summing up the probabilities of all plans.

In order to get the probability of each plan being chosen, we first transform the train that a passenger may board into the number of trains being needed to wait for. Then we define that the number of trains waited by passengers of Tapin(ζ) obeys the polynomial distribution with parameter $\boldsymbol{\theta}_\zeta = \{\theta_\zeta(1), \theta_\zeta(2), \ldots, \theta_\zeta(n)\}$, and the number of trains waited by passengers of Transfer(η) obeys the polynomial distribution with parameter $\boldsymbol{\beta}_\eta = \{\beta_\eta(1), \beta_\eta(2), \ldots, \beta_\eta(n)\}$.

From the process of a trip of a passenger, we can infer that $\boldsymbol{\beta}$ is affected by $\boldsymbol{\theta}$. So we can calculate $\boldsymbol{\theta}$ firstly, then $\boldsymbol{\beta}$. As not all the OD pairs have multiple routes, the trips with one route and no transfer can be used to estimate $\boldsymbol{\theta}_\zeta$ because the train chosen is unique. Then considering $\boldsymbol{\theta}_\zeta$ as prior knowledge, the trips with one route and some transfers can be used to estimate $\boldsymbol{\beta}$. Finally, considering both $\boldsymbol{\theta}_\zeta$ and $\boldsymbol{\beta}$ as prior knowledge, $\alpha_{j,z}$ can be estimated by maximizing function (1).

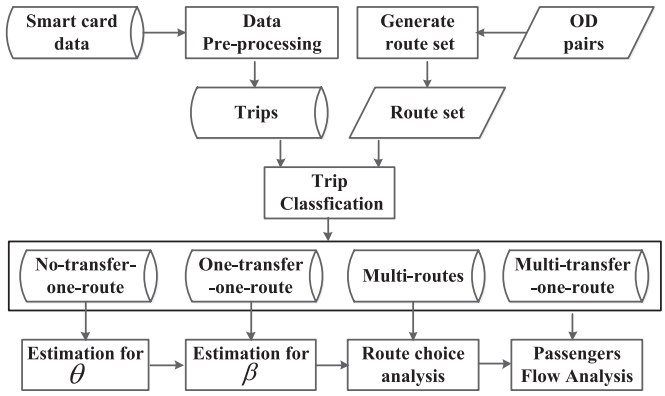


Fig. 3. Processing flowchart.

C. Framework

The framework is illustrated in Fig. 3. The details are given as follows.

For smart card data, we have been finding several kinds of errant data, e.g., missing data, duplicated data and data with logical errors. So in the step of data pre-processing, we conduct a detailed clearing process to filter out errant data on a daily basis. In the step of generating route set, we use the algorithm proposed in [24] to find the k shortest paths of all OD pairs. Then according to the time cost of passengers in practice, we filter some routes that passengers have never used. In the step of trips classification, according to the number of routes and transfers of their ODs, we classify all trips into four groups: No-transfer-one-route, One-transfer-one-route, Multi-transfer-one-route, Multi-routes. In the step of possible plan analysis, we find all possible plans that a passenger may choose by matching smart card data with trains' operation timetable. The trips in No-transfer-one-route and One-transfer-one-route groups are used for estimating $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ respectively. Then considering $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ as prior knowledge, we calculate the probability of each route being chosen for an OD with multiple-routes. Finally as an application, passenger flows are analyzed.

IV. Methodology

A. Finding All Effective Routes for an OD Pair

In this subsection, we use two steps to find all effective routes for an OD pair. The first step is to find all routes for an OD pair. The second step is to filter the routes that have never been used by passengers from these possible routes. We use the algorithm proposed in [24] to find the k shortest routes with efficiency in time $O(m + n \log n + k)$, where n, m are the numbers of the vertices and edges in a digraph respectively. We define the cost of a route as the maximum time cost that contains the minimum of walking time and running time of trains. k is determined in term of the accessibility and complexity of metro system. In practice, not all of the k routes of an OD are used by passengers. In order to filter those routes that passengers never choose, we sort all trips of an OD pair over two months by the time cost. We then filter the top $Y\%$ trips with largest time cost. Although

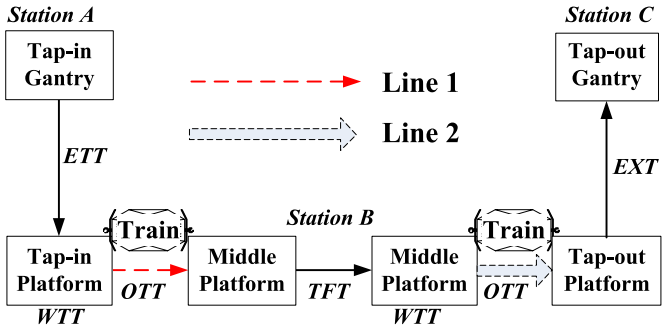


Fig. 4. Segmentation of a one-transfer trip.

most of passengers do not linger too long inside metro system, there are still some passengers showing abnormal travelling behaviors, such as beggars, express logistics worker. Their time cost and travel plan choice may be anomaly. Our recent work [14] found that a reasonable value of Y is 2, which can filter the abnormal passengers with high accuracy. Then we get the largest time cost denoted as $C_{\max}$ of the remaining $1-Y\%$ trips. Finally we filter the routes with time cost longer than $C_{\max}$ from all possible routes. The rest are effective routes. In this paper, if not explicitly pointed out, a route refers to an effective one.

B. Extracting All Possible Plans Chosen by Each Passenger

In this subsection, given a passenger's smart card data and train operation data, we extract all possible plans that can be chosen by the passenger. A general trip of a passenger in metro system can be depicted as 5 steps as shown in Fig. 4: 1) passing through entrance gate and walking to the platform, 2) waiting on the platform for a train, 3) boarding a train and staying on the train until the train reaches the passenger's destination, 4) getting off the train and exiting the metro system. To be noted, if the passenger needs to transfer, before step 4), 5) transit between platforms needs to be considered. So the whole trip duration is composed of: entry time (ETT), wait time (WTT), on train (OTT), transfer time (TFT), and exit time (EXT).

In this paper, we denote the minimal walking time of ETT, TFT, EXT as $\mathrm{ETT}_{\min}(l,s)$, $\mathrm{TFT}_{\min}(l,l',s)$ and $\mathrm{EXT}_{\min}(l,s)$ respectively, where l and l' are metro lines and s is a metro station. The method for calculating the value of $\mathrm{TFT}_{\min}(l,l',s)$ and $\mathrm{EXT}_{\min}(l,s)$ has been given in our previous paper [25].

Let us denote the arrival and departure time of a train tr at station s of metro line l as $T_{\mathrm{arrv}}(l,s,\mathrm{tr})$ and $T_{lv}(l,s,\mathrm{tr})$ respectively. Suppose a passenger x enter metro system at station s. His being able to board the train tr needs to satisfy the following Equation (2)(i). Likewise, if a passenger x exits metro system at station s, his being able to board the train tr before his exiting metro system needs to satisfy the following Equation (2)(ii). For a passenger who need to transit at transfer station s, let us denote the arrival time of a train tr at station s of line l as $T_{\mathrm{arrv}}(l,s,\mathrm{tr})$ and the departure time of another train tr' of another line l' at station s as $T_{lv}(l',s,\mathrm{tr}')$. That the passenger getting off from tr can board train tr' needs to satisfy Equation (2)(iii).

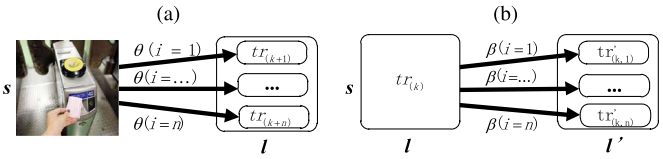


Fig. 5. Number of trains waited by passengers of (a) Tapin(ζ) and (b) Transfer(η).

$$
\begin{aligned}
&(i)x.b+\mathrm{ETT}_{\min}(l,s)\le T_{lv}(l,s,\mathrm{tr}).\\
&(i)x.e-\mathrm{ETE}_{\min}(l,s)\le T_{\mathrm{arrv}}(l,s,\mathrm{tr}).\\
&(iii)T_{lv}(l',s,\mathrm{tr}')-T_{\mathrm{arrv}}(l,s,\mathrm{tr})\ge \mathrm{TFT}_{\min}(l,l',s). \qquad (2)
\end{aligned}
$$

In sum, for a passenger, given each route, we can find all plans chosen during her/his trip by Equation (2).

C. Solution of $\boldsymbol{\theta}_\zeta$ and $\boldsymbol{\beta}_\eta$

In this section, we give the approaches for calculating $\boldsymbol{\theta}_\zeta$ and $\boldsymbol{\beta}_\eta$. As aforementioned, we define that the number of trains waited by the passengers of Tapin(ζ) obeys the polynomial distribution with parameter $\boldsymbol{\theta_\zeta}=\{\theta_\zeta(1),\theta_\zeta(2),\ldots,\theta_\zeta(n)\}$, and the number of trains waited by passengers of Transfer(η) obeys the polynomial distribution with parameter $\boldsymbol{\beta}_\eta=\{\beta_\eta(1),\beta_\eta(2),\ldots,\beta_\eta(n)\}$. We consider the two polynomial distributions $\boldsymbol{\theta}_\zeta$ and $\boldsymbol{\beta}_\eta$ separately. That's because the transfer passengers arrive at transit station almost simultaneously. While the time that the passengers arrive at the origin station is more random. Hence we first solve that given a plan chosen by a passenger, how to transform it into the number of trains that the passenger waits for. Then we give an approach to estimate $\boldsymbol{\theta}_\zeta$ and $\boldsymbol{\beta}_\eta$ using several specific trips.

1) The Number of Trains Waited by Passengers: Given a train boarded by a passengers of Tapin(ζ), in order to transform it into the number of trains the passenger waited for, we divide these passengers of Tapin(ζ) into several groups according to the arrival time of trains. We use tapin(ζ,k) of Tapin(ζ) to represent the passengers who enter the metro system between the departure time of train $\mathrm{tr}_{(k)}$ and the departure time of train $\mathrm{tr}_{(k+1)}$. Suppose for these passengers in tapin(ζ,k), the set of trains that they may board is $\{\mathrm{tr}_{(k+1)},\mathrm{tr}_{(k+2)},\ldots,\mathrm{tr}_{(k+n)}\}$, as shown in Fig. 5(a). n is the maximum number of trains needed to be waited for. Field observations show that the first-come-first-served policy is not applicable in practice. There are many factors affecting which train a passenger eventually gets on, such as the distance between the gate and the platform, walking speed, the number of passengers in the waiting queue. Furthermore, a typical train has six to eight cars with multiple doors available for boarding simultaneously. A wise strategy or good luck in choosing train doors could also lead to an earlier boarding. So the train that a passenger eventually gets on is more likely to be a random variable in practice. Let the probability of train $\mathrm{tr}_{(k+i)}$ boarded by these passengers is $\theta_\zeta(i)$.

Thus the number of trains needed to be waited for (the number of passengers in these trains) obeys to polynomial distribution, where $\sum_{i=1}^{n} \theta_\zeta(i) = 1$.

Likewise, given a train boarded by the passengers of Transfer(η), in order to transform it into the number of trains the passengers waited for, we divide these passengers of Transfer(η) into several groups according to the arrival time of trains of line l. We use transfer(η, k) to represent the passengers who get off train $\mathrm{tr}_{(k)}$ of line l. Suppose for these passengers in transfer(η, k), the set of trains that they transfer is $\{\mathrm{tr}'_{(k,1)}, \mathrm{tr}'_{(k,2)}, \ldots, \mathrm{tr}'_{(k,n)}\}$ as shown in Fig. 5(b). For a passenger, the train that he may get on is also influenced by many factors, such as the distance between platforms, walking speed, the waiting position for train, the number of people in the waiting queue, and so forth. So we see the train that a passenger eventually gets on as a random variable. We assume the probability of the number of trains needed to be waited for is stable in same time slot. We denote the probability of $\mathrm{tr}'_{(k,i)}$ boarded by a passengers of transfer(η, k) as β_η^i. Thus the number of passenger in these trains obeys to polynomial distribution, where $\sum_{i=1}^{n} \beta_\eta^i = 1$.

2) Calculating of $\boldsymbol{\theta}_\zeta$ and $\boldsymbol{\beta}_\eta$: From the process of trips of a passengers, we can infer that $\boldsymbol{\beta}_\eta$ is affected by $\theta_\zeta(i)$ ($\boldsymbol{\theta} \rightarrow \boldsymbol{\beta}$). So we can estimate $\boldsymbol{\theta}_\zeta$ first, then for $\boldsymbol{\beta}_\eta$.

In order to calculate $\theta_\zeta(i)$, we assume that several specific trips of Tapin(ζ) are representative enough to analyze the distribution of the number of trains needed to wait for at an origin station. This is a practical because it is difficult to ascertain the exact train chosen for every passenger during the first part of his trip, especially for complex scene with multiple transfers and multiple routes. However the train chosen for a trip with only one route and no transfer can be inferred. So we classified the trips with only one route as group 1. According to the our assumption in Section III-B, we can know that given the route chosen, the train chosen in the last part of a trip can be determined uniquely. That means the train boarded at the first part of the trip of group 1 can be inferred, because these trips only have one part during their total journeys. And according to our statistics, the volume of trips in group 1 accounts for 30% of the whole. Clearly, those trip are representative enough [26]. So we can use these trips with only one route and no transfer needed to estimate $\boldsymbol{\theta}$.

Similarly we assume several specific trips of Transfer(η) are representative enough to analyze the distribution of the number of trains needed to wait for at transfer station when passengers need to transit. A trip with only one route and one transfer has two parts during its journey. We classified these trips as group 2 in this paper. The train boarded in the last part of trips in group 2 can be inferred uniquely. Though there may be more than one trains passengers boarded in the first part, the possibility for these trains can be known from $\boldsymbol{\theta}$. So we can see $\boldsymbol{\theta}$ as prior knowledge and use trips in group 2 to estimate $\beta_\eta(i)$.

So, we divide all trips into four classes according to the number of routes and transfers of their ODs: No-transfer-one-route, one-transfer-one-route, multi-transfer-one-route, Multi-routes. The passengers in No-transfer-one-route and One-transfer-one-route are used for estimating $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ respectively. Then using the passengers in Multi-routes and considering $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ as prior knowledge, we calculate the probability of each route being chosen for an OD with Multi-routes in the next subsection.

Suppose the set of passengers of Tapin(ζ) in group 1 (one-route-no-transfer) is $X = \{x^{(1)}, x^{(2)}, x^{(3)}, \ldots, x^{(Q)}\}$. Using the approach given in Section IV-B, we can get the number of trains that each passenger waits for in X is $W = \{w^{(1)}, w^{(2)}, w^{(3)}, \ldots, w^{(Q)}\}$ respectively. Suppose the number of passengers waiting i trains is c_i by counting the same digits of W. We use maximum-likelihood estimation to obtain the value of $\boldsymbol{\theta}_\zeta$ by Equation (3a) and (3b).

$$L(X, Tt, \boldsymbol{\theta}_\zeta) = \log \sum_{i=1}^{n} \left(\boldsymbol{\theta}_\zeta(i)\right)^{c_i} \tag{3a}$$

$$\boldsymbol{\theta}_\zeta(i) = \frac{c_i}{\sum_{j=1}^{n} c_j}. \tag{3b}$$

Suppose the set of passengers of Transfer(η) in group 2 is $X = \{x^{(1)}, x^{(2)}, x^{(3)}, \ldots, x^{(Q)}\}$. They may have different original stations and beginning time slots. We use $\zeta^{(q)}$ to represent the first part of the trip of passenger $x^{(q)}$. Given a passenger $x^{(q)}$, suppose the set of plans that the passenger may choose is $p_q = \{p_q^1, p_q^2, \ldots, p_q^{M^q}\}$. The train chosen in the second part of trip can be obtained. A plan p_q^m can be represented as $\{\mathrm{tr}_q^{m,1}, \mathrm{tr}_q^2\}$. The numbers of trains needed to be waited for are $\{w_q^{m,1}, w_q^{m,2}\}$. We use maximum-likelihood estimation to obtain the value of $\boldsymbol{\beta}_\eta$ by function Equation (4). It is difficult to calculate the derivatives of the logarithm of the sum of some formulas in Equation (4a) for maximum. So firstly we convert it to Equation (4b) by applying the Jensen inequality, and then calculate $\boldsymbol{\beta}$ by maximizing the value of right hand side of Equation (4).

$$\begin{aligned} L(X, Tt, \boldsymbol{\theta}, \boldsymbol{\beta}_\eta) &= \log \prod_{q=0}^{Q} \Pr\left(p_q | x^q.b, \boldsymbol{\theta}, \boldsymbol{\beta}_\eta\right) \\ &= \sum_{q=0}^{Q} \log \left(\sum_{m=1}^{M^q} \Pr(w_q^{m,1}, w_q^{m,2} | x^q.b, \boldsymbol{\theta}, \boldsymbol{\beta}_\eta \right) \end{aligned} \tag{4a}$$

$$\begin{aligned} &= \sum_{q=0}^{Q} \log \left(\sum_{m=1}^{M^q} \boldsymbol{\theta}_{\zeta^{(q)}}\left(w_q^{m,1}\right) \times \boldsymbol{\beta}_\eta\left(w_q^{m,2}\right) \right) \\ &\geq \sum_{q=0}^{Q} \left(\sum_{m=1}^{M^q} \boldsymbol{\theta}_{\zeta^{(q)}}\left(w_q^{m,1}\right) \log \boldsymbol{\beta}_\eta\left(w_q^{m,2}\right) \right). \end{aligned} \tag{4b}$$

D. Calculating the Probability of Each Route Being Chosen for anOD Pair

In this subsection, we aim to give an approach to calculate the probability of each route being chosen for an OD pair with multiple effective routes. Suppose the set of effective routes from station O to D is $R = \{R_1, R_2, \ldots, R_Z\}$. We denote the

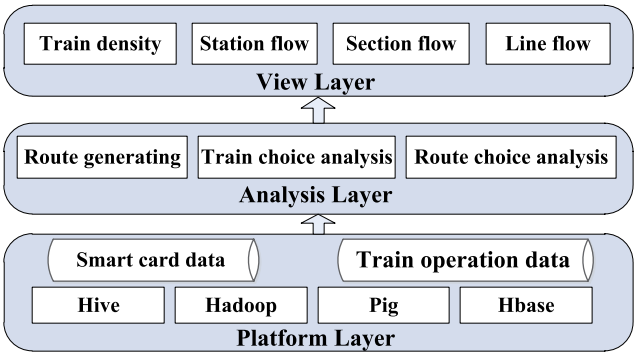


Fig. 6. System architecture.

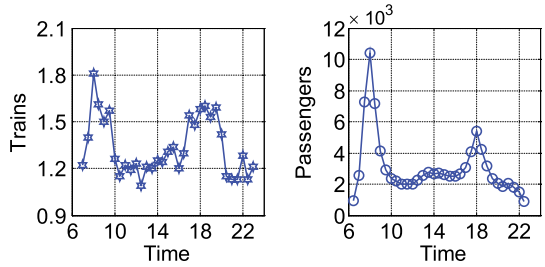


Fig. 7. (a) Average number of trains waited. (b) Sectional flow (LaoJie–DaJuYuan).

probability of route R_z being chosen at time slot I_j is denoted as $\alpha_{j,z}$, where $\sum_{z=0}^{Z} \alpha_{j,z} = 1$.

Suppose the set of the passengers entering metro system during time slot I_j from station O to station D is $X = \{x^1, x^2, x^3, \ldots, x^Q\}$, where Q is the number of passengers. We assume that they are independent. For a passenger x^q in X, given that he chooses route R_z, we can obtain all plans that the passenger may choose during his trip using the approach given in Section IV-B. Denote the set of plans as $p_{q,z}$, where $\Pr(p_{q,z}|x^q.b, R_z, \boldsymbol{\theta}, \boldsymbol{\beta})$ is the possibility that a passenger x^q choose $p_{q,z}$ on condition of $x^q.b$, route chosen R_z, $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. For the same reason, it is difficult to calculate the derivatives of the logarithm of the sum of some formulas in Equation (5a), so we transform it into Equation (5b).

$$L\left(X, Tt, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\alpha}_j\right)$$

$$= \sum_{x^q \in X} \log \sum_{R_z \in R} \left(\alpha_{j,z} \times \Pr(p_{q,z}|x^q.b, R_z, \boldsymbol{\theta}, \boldsymbol{\beta})\right) \quad (5a)$$

$$\geq \sum_{x^q \in X} \sum_{R_z \in R} \left(\alpha_{j,z} \times \log \Pr(p_{q,z}|x^q.b, R_z, \boldsymbol{\theta}, \boldsymbol{\beta})\right). \quad (5b)$$

V. System Implementation

Our algorithm of calculating the probability of each route being chosen for an OD pair with multiple routes is based on a large amount of data. The framework of our system is illustrated in Fig. 6, which has three layers: the platform layer, the analysis layer and the view layer.

The platform layer is mainly used for storage and job processing purpose. Our algorithms need batch processing on a large amount of data. So it is more efficient to run on a parallel platform [27]. We use distributed computing platform Hadoop [28] that was designed for batch processing in big data. It mainly includes two modules, HDFS [29] and MapReduce [30]. HDFS provides high-throughput access to large data. MapReduce is for parallel processing of large data sets. In our platform, we utilize a 34 TB Hadoop Distributed File System (HDFS) on a cluster consisting of 11 nodes, each of which is equipped with 32 cores and 32 GB RAM. To improve retrieving efficiency, some mapReduce based software tools, such as Pig [31], Hbase [32] are used.

The analysis layer running on platform layer is the keystone of our paper. It mainly contains three parts: Route generating for getting all effective plans of each OD pair; Train choice analysis for finding all possible trains that a passenger may get on; Route choice analysis for calculating the probability of each route being chosen for an OD pair with multiple routes. The three parts are based on large volume of data. They are all being translated to a series of MapReduce jobs that run on the distributed environment.

The view layer based on analysis layer performs passenger flow analysis and displays the results to public or transport agencies for strategic planning and management, such as the spatio-temporal passenger flow analysis for all metro lines, trains, sections, and so on.

VI. Case Study

A. Data Set

The data set used in this study is the smart card transaction records and train operation logs in Shenzhen, China. The metro system has 5 metro lines by 2013. The whole data collected from around 4 million smart cards have more than 300 million smart card transaction records, covering 60 consecutive days from June 1, 2015 to July 30, 2015.

B. Left Behind Analysis

Fig. 7(a) shows the average number of trains waited by passengers in station LaoJie of metro line (LuoHu $\sim$ JiChangDong) at the first part of their trips at different time slots of one day. Fig. 7(b) shows the number of passengers passing through station LaoJie (sectional flows of two adjacent sites LaoJie to DaJuYuan) at different time slots of one day. The station Laojie locating in the heart of Shenzhen business district is a transfer station of line 1 and line 3. There are about 12 thousand tap-in passengers and 60 thousand transfer passengers in LaoJie per day. From Fig. 7, we can get that there are a remarkable similarity of the two lines. The cross-correlation of the number of train waited and sectional flow is 0.75 which is larger than 0.7. So we can assume that the more the passengers passing through a station, the more left behind passengers there are in the station.

From Fig. 7(a) we also can get that the phenomena of left behind is varying with time, which is a good indicator of transit service performance and can provide better travel advice

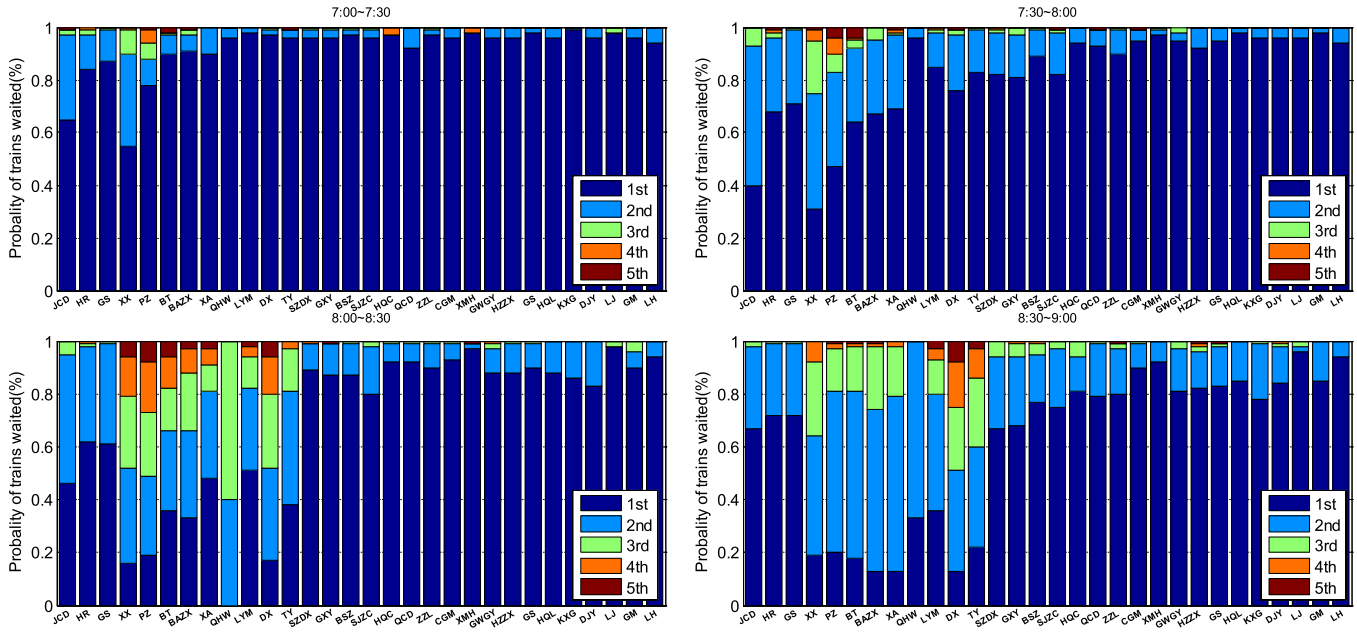


Fig. 8. Probability of the number of trains needed to wait for line 1 (from JCD to Luohu) at four time slots in AM peak travel time.

for users. There are two obvious peak periods 7:00∼9:00 and 18:00∼20:00 in Fig. 7. That is because there are so many passengers who go to work in the morning rush hours and back home in the evening rush hours that the capacity of trains cannot meet the actual requirements. So in rush hours, passengers must wait for more trains. Another point that deserves further explanation is that even during off-peak periods that train may not be crowded, the average number of trains needed to be waited for is bigger than 1.0. That is, not all of passengers get on the first available train. This is understandable because some passengers care about comfort that they anticipate that the next train will have seats available and choose to wait.

Fig. 8 gives the distribution of the number of trains waited by passengers at all stations of metro line 1(JiChangDong-LuoHu) at four time slots (07:00∼07:30, 07:30∼08:00, 08:00∼08:30, 08:30∼09:00) of morning rush hours. In Fig. 8, the bar labeled with "*1st*" means the probability that a passenger needs to wait for one train (gets on the 1st coming train). "*2nd*" means the probability that a passenger needs to wait for two trains, and so on. From Fig. 8, we can get that the left-behind varies in time and space. That is because the distribution of passengers is uneven in time and space as shown in Fig. 9.

Fig. 9 gives all sectional flows of metro line 1 (JiChangDong-LuoHu) at four time slots (07:00∼07:30, 07:30∼08:00, 08:00∼08:30, 08:30∼09:00) of morning peak. We can get that the left-behind is most serious from XX (short of XiXiang) to TY (short of TaoYuan) in 08:30∼08:30, which indicates that the train capacity cannot meet the demand of passengers in these station.

Moreover, station JCD(JiChangDong) is the first station of line 1 from JCD to Luohu, which tell us that all trains arriving at this station are empty. That means that there are more remaining capacity than other stations. But from this figure we get that there are also many passengers in JCD who need to wait for more than one trains. There are two reasons: Firstly, JCD is the

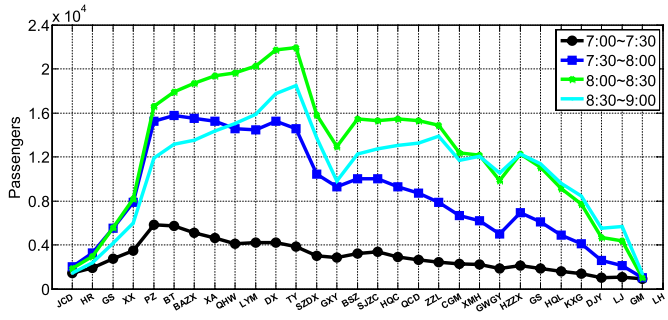


Fig. 9. Sectional flows of line 1 (from JCD to Luohu) at four time slots in AM peak travel time.

closest metro station to Shenzhen airport, where there are a lot of passengers getting off from plane and carry packs of luggage and struggle for a local train. Secondly, for safety and comport, they are more likely to choose the next train with more seats. As JCD is the first station of line 1, passengers prefer a train with seats more than other stations.

C. Route Choice Pattern

In this section, two typical OD pairs of stations were chosen to illustrate our proposed method to calculate the probability of each route being chosen. Fig. 10(a) shows the layout of the two OD pairs.

The first OD pair is BaiShiLong-FuTian that had two effective routes, 1) Take the metro line 4, get off at station ShaoNianGong. Then take metro line 3, get off at the FuTian station. 2) Take the metro line 4, get off at ShiMinZhongXin. Then take metro line 2, get off at the FuTian station. Both of the two routes need one transfer and average time cost of them is nearly the same. The first route is recommended by some mobile apps, such as Baidu map App, Shenzhen metro App

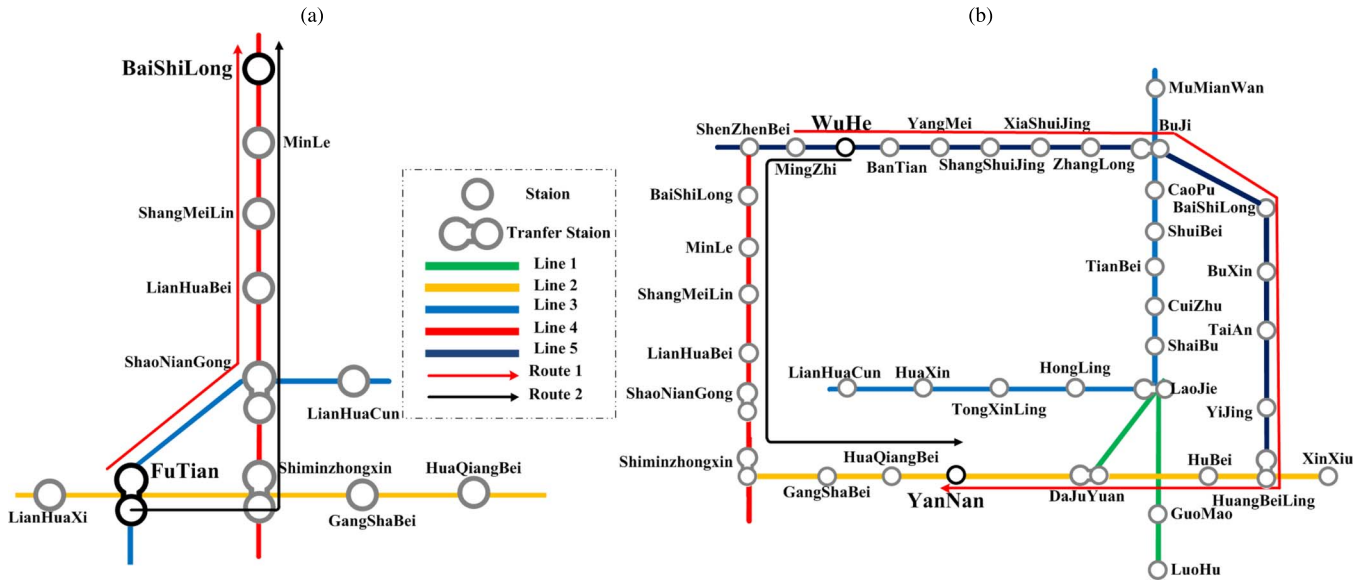


Fig. 10. Layout of two OD pairs.

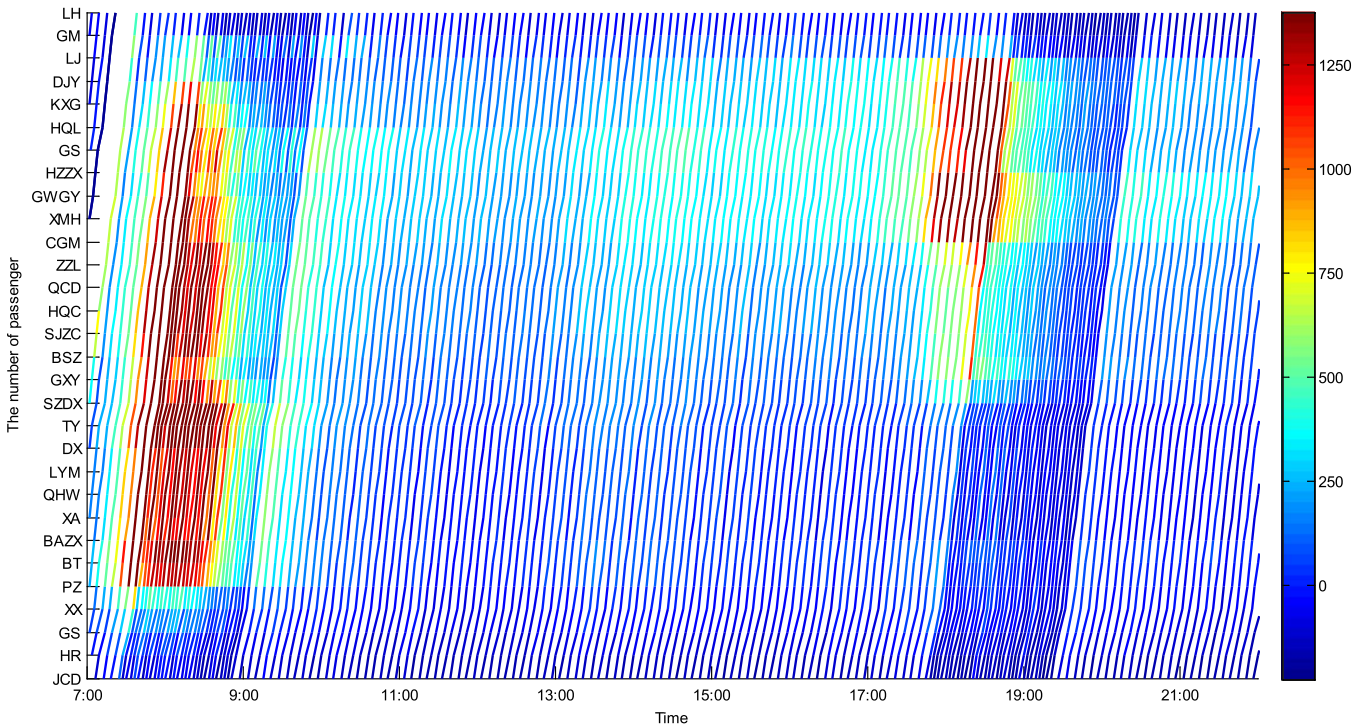


Fig. 11. Spatiotemporal density of all trains.

provided by Shenzhen Metro Group Company. However our experiment results show that the probability of the first route being chosen at rush hours and off peak hours is 31% and 42%, respectively, which is less than the probability of the second route being chosen. The results tell us that the route given by mobile app doesn't always reflect most passengers' real choice. It also provides proof of the walking penalty when the general cost of a path is calculated.

According to a survey about all transfer stations in Shenzhen, ShaoNianGong is one of the transfer stations with longest walking time. The walking time is about 5 minute, which is more than that of ShiMinZhongXin with 2 minute. Our on-site investigations tell us that most of passengers do not prefer to walking two much time when transferring. Some passengers tell us that they do not know the actual walking time in every transfer station. Generally, they will follow the shortest path by some map app in their smartphone, which tell them the first route has less time cost than the second one. Based on that, it is understandable for the passengers' route choice behavior of BaiShiLong-FuTian. Most of passengers at peak period are

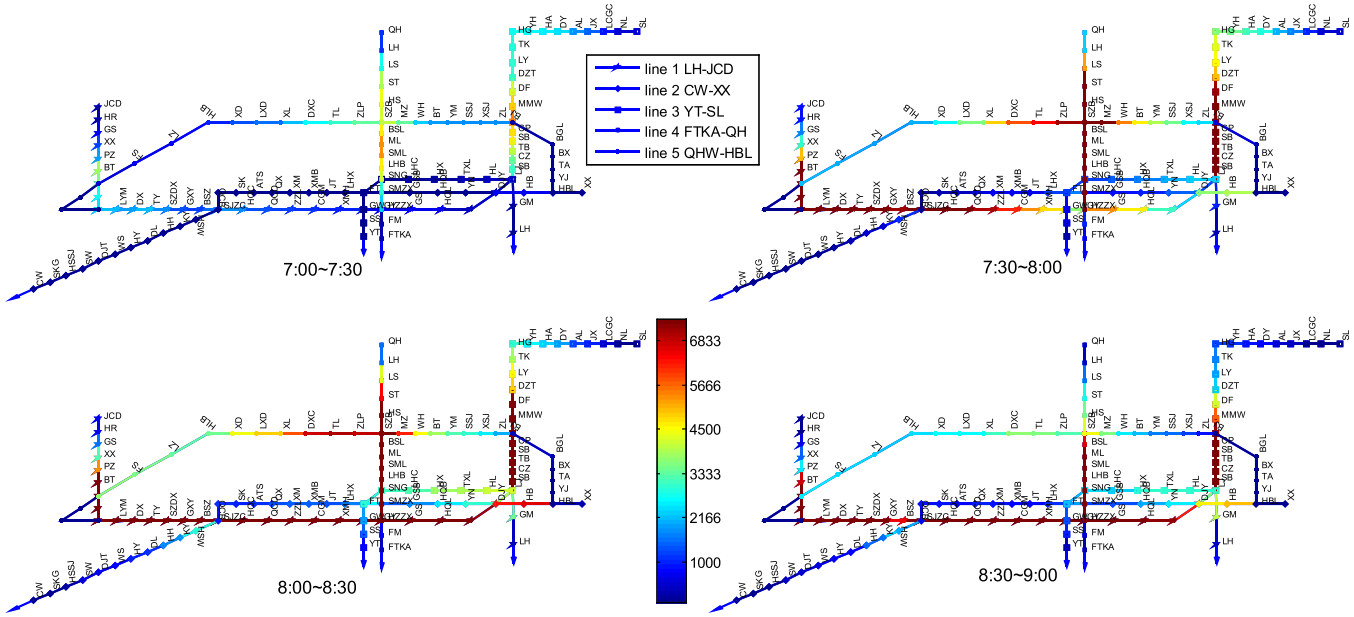


Fig. 12. Metro sectional flow at AM peak travel hours.

local residents. They are more familiar with the metro and know more about the walking time cost than the passengers in off peak period such as visitors. Tourists who have less experience are more likely to rely on mobile apps. So they are more likely to choose to transfer in ShiMinZhongXin.

The second OD pair is WuHe-YanNan. There are also two effective routes as shown in Fig. 10(b). 1) Take metro line 5, get off at HuangBeiLing. Then take metro line 2. 2) Take the metro line 5, get off at ShenZhenBei, then take metro line 4, get off at ShiMinZhongXin, then take metro line 2. The first route costs ten minutes more than the second one. However our analysis results show that there are still 40% passengers choosing the first route. This is because the second route has one more transfers, which is likely to offset the advantage of low time cost. The result provides proof of the transfer penalty when the general cost of a path is calculated.

D. Spatio-Temporal Density Analysis

Spatio-temporal density of all trains of metro line 1(Luohu-Jichangdong) is shown in Fig. 11, where the x axis and y axis represent time and station respectively. Every train starts at the lowest station and finally reaches the highest station in the y axis. Each diagonal line represents a train and covers the information about the train' spatio-temporal density. The color represents the density of passengers. From this figure we can get that there are two peak hours as morning and evening. The density in morning peak is more serious than in evening peak. So the departure intervals of trains in morning peak could be set to be shorter than that in evening peak.

Beside, this spatio-temporal density information can be used for assessing the metro service and forecasting the density of all running train and so on. The sectional flow of whole metro system at four time slots in morning peak is shown in Fig. 12. We can get that 1) The passenger flows are most crowded in 8:00∼8:30 2) The passenger flow is uneven distributed throughout whole metro system. The metro line 1, 3, 4 have more passengers than line 2 and 5. For metro line 1, 3, 4, the densities are more serious in middle than that in both sides. This can help to design schedules for shuttle trains and so on. All these information can improve both passengers' and transportation operators' knowledge on transportation system. For example the information can be further used to improve the service by redesigning timetable, adjusting velocity, and etc. Passengers on the other hand can also plan their trips based on the information.

VII. Conclusion and Discussion

In this paper, we present an approach to calculate the probability of route choices for an OD pair with multiple routes in a complex metro network. In doing so, we find, for each passenger, all possible plans that he/she can choose for each effective route by matching smart card data and train operation logs. We also calculate two kinds of time-dependent polynomial distributions using maximum likelihood estimation. One is the number of trains that a passenger waits for at his/her original station. The other is the number of trains that a passenger waits for when he/she changes at the transfer station. Based on that, we propose a probability model to calculate the probability of each route being chosen for an OD with multi-paths. The approach in this paper is applied to Shenzhen metro system. On-site investigations validate that our algorithm is accurate and can be used to estimate passenger flows.

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments.

REFERENCES

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

[2] L. Sun, Y. Lu, J. G. Jin, D.-H. Lee, and K. W. Axhausen, "An integrated bayesian approach for passenger flow assignment in metro networks," *Transp. Res. C, Emerging Technol.*, vol. 52, pp. 116–131, 2015.

[3] Q. Qu, S. Liu, B. Yang, and C. S. Jensen, "Efficient top-k spatial locality search for co-located spatial web objects," in *Proc. IEEE 15th Int. Conf. MDM*, Brisbane, Australia, Jul. 14–18, 2014, vol. 1, pp. 269–278.

[4] Q. Qu, S. Liu, C. S. Jensen, F. Zhu, and C. Faloutsos, "Interestingness-driven diffusion process summarization in dynamic networks," in *Proc. ECML PKDD*, Nancy, France, Sep. 15–19, 2014, pp. 597–613.

[5] Q. Qu, C. Chen, C. S. Jensen, and A. Skovsgaard, "Space-time aware behavioral topic modeling for microblog posts," *IEEE Data Eng. Bull.*, vol. 38, no. 2, pp. 58–67, 2015.

[6] S. Peeta and A. K. Ziliaskopoulos, "Dynamic traffic assignment: The past, the present, and the future," *Netw. Spatial Econ.*, vol. 1, no. 3/4, pp. 233–266, 2001.

[7] S. Liu, Q. Qu, and S. Wang, "Rationality analytics from trajectories," *Trans. Knowl. Discovery Data*, vol. 10, no. 1, p. 10, 2015.

[8] C. Liu and Q. Qu, "Trip fare estimation study from taxi routing behaviors and localizing traces," in *Proc. IEEE ICDMW*, Atlantic City, NJ, USA, Nov. 14–17, 2015, pp. 1109–1116.

[9] Y. Sheffi, *Urban Transportation Networks: Equilibirum Analysis With Mathematical Programming Methods*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1985.

[10] H. Talaat and B. Abdulhai, "Modeling driver psychological deliberation during dynamic route selection processes," in *Proc. IEEE ITSC*, 2006, pp. 695–700.

[11] S. Nakayama and R. Kitamura, "Route choice model with inductive learning," *J. Trans. Res. Board*, vol. 1725, pp. 63–70, 2000.

[12] M. Bagchi and P. White, "The potential of public transport smart card data," *Transp. Policy*, vol. 12, no. 5, pp. 464–474, 2005.

[13] B. Agard, C. Morency, and M. Trépanier, "Mining public transport user behaviour from smart card data," in *Proc. 12th IFAC Symp. INCOM*, 2006, pp. 17–19.

[14] J. Zhao, C. Tian, F. Zhang, C. Xu, and S. Feng, "Understanding temporal and spatial travel patterns of individual passengers by mining smart card data," in *Proc. IEEE 17th ITSC*, 2014, pp. 2991–2997.

[15] M.-P. Pelletier, M. Trépanier, and C. Morency, "Smart card data use in public transit: A literature review," *Transp. Res. C, Emerging Technol.*, vol. 19, no. 4, pp. 557–568, 2011.

[16] J. G. Jin, L. C. Tang, L. Sun, and D.-H. Lee, "Enhancing metro network resilience via localized integration with bus services," *Transp. Res. E, Logist. Transp. Rev.*, vol. 63, pp. 17–30, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1366554514000039

[17] L. Sun, J. G. Jin, D.-H. Lee, K. W. Axhausen, and A. Erath, "Demand-driven timetable design for metro services," *Transp. Res. C, Emerging Technol.*, vol. 46, pp. 284–299, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0968090X1400182X

[18] L. Sun, D.-H. Lee, A. Erath, and X. Huang, "Using smart card data to extract passenger's spatio-temporal density and train's trajectory of mrt system," in *Proc. ACM SIGKDD Int. Workshop Urban Comput.*, 2012, pp. 142–148.

[19] Q. Fu, R. Liu, and S. Hess, "A Bayesian modelling framework for individual passengers probabilistic route choices: A case study on the London underground," in *Proc. Transp. Res. Board 93rd Annu. Meet. (No. 14-5328)*, 2014, pp. 197–203.

[20] Z. Tian, Y. Wang, C. Tian, F. Zhang, L. Tu, and C. Xu, "Understanding operational and charging patterns of electric vehicle taxis using GPS records," in *Proc. IEEE 17th Int. ITSC*, 2014, pp. 2472–2479.

[21] J. Zhang, X. Yu, C. Tian, F. Zhang, L. Tu, and C. Xu, "Analyzing passenger density for public bus: Inference of crowdedness and evaluation of scheduling choices," in *Proc. IEEE 17th Int. ITSC*, 2014, pp. 2015–2022.

[22] J. Huang, L. Wang, C. Tian, F. Zhang, and C. Xu, "Mining freight truck's trip patterns from GPS data," in *Proc. IEEE 17th Int. ITSC*, 2014, pp. 1988–1994.

[23] Y. Li, C. Tian, F. Zhang, and C. Xu, "Traffic condition matrix estimation via weighted spatio-temporal compressive sensing for unevenly-distributed and unreliable GPS data," in *Proc. IEEE 17th Int. Conf. ITSC*, 2014, pp. 1304–1311.

[24] D. Eppstein, "Finding the k shortest paths," *SIAM J. Comput.*, vol. 28, no. 2, pp. 652–673, 2006.

[25] F. Zhang, J. Zhao, C. Tian, C. Xu, X. Liu, and L. Rao, "Spatio-temporal segmentation of metro trips using smart card data," *IEEE Trans. Veh. Technol.*, vol. 65, no. 3, pp. 1137–1149, Mar. 2015.

[26] *Wikipedia, the free encyclopedia*, Sampling statistics. [Online]. Available: https://en.wikipedia.org/wiki/Samplingstatistics

[27] F.-Y. Wang, "Parallel control and management for intelligent transportation systems: Concepts, architectures, and applications," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 3, pp. 630–638, Sep. 2010.

[28] T. White, *Hadoop: The Definitive Guide*. Sebastopol, CA, USA: O'Reilly, 2012.

[29] D. Borthakur, "HDFS architecture guide," *HADOOP APACHE PROJECT*, 2008. [Online]. Available: https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

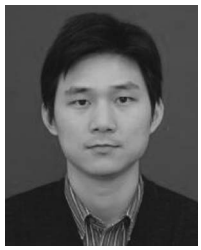
[30] J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[31] A. F. Gates *et al.*, "Building a high-level dataflow system on top of map-reduce: The pig experience," *Proc. VLDB Endowment*, vol. 2, no. 2, pp. 1414–1425, 2009.

[32] L. George, *HBase: The Definitive Guide*. Sebastopol, CA, USA: O'Reilly Media, 2011.

Juanjuan Zhao received the M.S. degree from the Department of Computer Science, Wuhan University of Technology, Wuhan, China, in 2009. She is currently working toward the Ph.D. degree in the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China.

From 2009 to 2012, she was a Research Assistant with Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. Her research interests include cloud computing, big data processing, streaming-data processing, data-fusion technique, big-data-driven systems, and spatiotemporal data mining.



Fan Zhang received the Ph.D. degree in communication and information system from Huazhong University of Science and Technology, Wuhan, China, in 2007.

From 2009 to 2011, he was a Postdoctoral Fellow with University of New Mexico, Albuquerque, NM, USA, and University of Nebraska-Lincoln, Lincoln, NE, USA. He is an Associate Professor with Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. His research topics include big data processing, data privacy, and urban computing.

Lai Tu received the B.S. degree in communication engineering and the Ph.D. degree in information and communication engineering from Huazhong University of Science and Technology, Wuhan, China, in 2002 and 2007, respectively.

From July 2007 to December 2008, he was a Postdoctoral Fellow with the Department of Electronic and Information Engineering, Huazhong University of Science and Technology. From January 2009 to October 2010, he was a Postdoctoral Researcher with the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan. Currently, he is an Associate Professor with the School of Electronic and Information and Communications, Huazhong University of Science and Technology. His research areas include urban computing, human behavior study, mobile computing, and networking.

Chengzhong Xu received the Ph.D. degree from University of Hong Kong, Kowloon, Hong Kong, in 1993.

He is a Professor with the Department of Electrical and Computer Engineering, Wayne State University (WSU), Detroit, MI, USA. He also holds an adjunct appointment with Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China, as the Director of the Institute of Advanced Computing and Data Engineering. His research interest is in parallel and distributed systems and cloud computing. He has published more than 200 papers in journals and conferences.

Dr. Xu serves on a number of journal editorial boards, including IEEE TRANSACTIONS ON COMPUTERS, IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, IEEE TRANSACTIONS ON CLOUD COMPUTING, *Journal of Parallel and Distributed Computing*, and *Science China Information Sciences*. He was the recipient of the Best Paper Nominee of the 2013 IEEE High Performance Computer Architecture and the Best Paper Nominee of the 2013 Association for Computing Machinery High Performance Distributed Computing. He was a recipient of the Faculty Research Award, Career Development Chair Award, and the President's Award for Excellence in Teaching of WSU. He was also a recipient of the Outstanding Overseas Scholar Award of the National Science Foundation of China. For more information, visit http://www.ece.eng.wayne.edu/czxu.

Dayong Shen received the bachelor's and master's degrees in system engineering from National University of Defense Technology, Changsha, China, in 2011 and 2013, respectively, where he is currently working toward the Ph.D. degree in social transportation and social logistics.

His research interests include intelligent scheduling, artificial intelligence algorithm, and parallel social systems. He has rich experience in designing and implementing parallel logistics system projects.

Chen Tian received the B.S., M.S., and Ph.D. degrees from the Department of Electronics and Information Engineering, Huazhong University of Science and Technology, Wuhan, China, in 2000, 2003, and 2008, respectively.

He was previously an Associate Professor with the School of Electronics Information and Communications, Huazhong University of Science and Technology. From 2012 to 2013, he was a Postdoctoral Researcher with the Department of Computer Science, Yale University, New Haven, CT, USA. He is an Associate Professor with State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China. His research interests include data center networks, network function virtualization, distributed systems, Internet streaming, and urban computing.

Xiang-Yang Li received the bachelor's degree from the Department of Computer Science and the bachelor's degree from the Department of Business Management from Tsinghua University, Shenzhen, China, both in 1995, and the M.S. and Ph.D. degrees from the Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA, in 2000 and 2001, respectively.

He previously was a Professor with the Department of Computer Science, Illinois Institute of Technology, Chicago, IL. He currently is a Professor with the School of Computer Science and Technology, University of Science and Technology of China, Hefei, China. He published a monograph entitled *Wireless Ad Hoc and Sensor Networks: Theory and Applications.* He coedited several books, including, *Encyclopedia of Algorithms.* His research interests include wireless networking, mobile computing, security and privacy, cyberphysical systems, and algorithms.

Dr. Li is an Editor of several journals, including IEEE TRANSACTION ON MOBILE COMPUTING and IEEE/ACM TRANSACTION ON NETWORKING. He has served many international conferences in various capacities, including the ACM International Conference on Mobile Computing and Networking, the ACM International Symposium on Mobile Ad Hoc Networking and Computing, and the IEEE International Conference on Mobile Ad hoc and Sensor Systems IEEE MASS. He is an ACM distinguished Scientist. He and his students were the recipients of the five Best Paper Awards (IEEE GlobeCom 2015, IEEE HPCCC 2014, ACM MobiCom 2014, COCOON 2001, IEEE HICSS 2001) and one Best Demo Award (ACM MobiCom 2012).

Zhengxi Li received the B.S. degree from the Department of Electrical Automation, Beihua University, Jilin, China, in 1976.

He is currently a Doctoral Supervisor, a Professor, and the Vice President with North China University of Technology, Beijing, China. His research interests cover intelligent traffic control and management, control theory and control engineering, and electric drive technology.