Hindawi
Journal of Advanced Transportation
Volume 2018, Article ID 2710608, 13 pages
https://doi.org/10.1155/2018/2710608

Research Article

Using Smart Card Data Trimmed by Train Schedule to Analyze Metro Passenger Route Choice with Synchronous Clustering

Wei Li,[1,2] **Qin Luo**,[2,3] **Qing Cai**,[4] **and Xiongfei Zhang**[3]

[1]*Key Laboratory of Optoelectronic Devices and Systems of Ministry of Education and Guangdong Province, College of Optoelectronic Engineering, Shenzhen University, Shenzhen, China*
[2]*Shenzhen Key Laboratory of Urban Rail Transit, Shenzhen University, Nanshan Ave 3688, Shenzhen, China*
[3]*College of Urban Traffic and Logistics, Shenzhen Technology University, Lantian Road 3002, Shenzhen, China*
[4]*Department of Civil, Environment and Construction Engineering, University of Central Florida, Orlando, Florida 32816, USA*

Correspondence should be addressed to Qin Luo; luoqin82@126.com

Received 23 November 2017; Revised 13 March 2018; Accepted 29 April 2018; Published 24 July 2018

Academic Editor: Francesco Corman

The metro passenger route choice, influenced by both train schedule and time constraints, is important to metro operation and management. Smart card data (Automatic Fare Collection (AFC) data in metro system) including inbound and outbound swiping time are useful for analysis of the characteristics of passengers' route choices in metro while they could not reflect the property of train schedule directly. Train schedule is used in this paper to trim smart card data through removing inbound and outbound walking time to/from platforms and waiting time. Thus, passengers' pure travel time in accord with trains' arrival and departure can be obtained. Synchronous clustering (SynC) algorithm is then applied to analyze these processed data to calculate passenger route choice probability. Finally, a case study was conducted to illustrate the effectiveness of the proposed algorithm. Results showed the proposed algorithm works well to analyze metro passenger route choice. It was shown that passenger route choice during both peak period and flat period could be clustered automatically, and noise data are isolated. The probability of route choice calculated through SynC algorithm can be used to revise traditional model results.

1. Introduction

Metro passenger route choice is vitally important to metro operation and management, such as passenger flow distribution and metro tickets clearing. It can provide useful data to help enhance train schedules to make full use of the train capacity. However, the metro passenger behavior is totally different from the car user behavior. The former one is largely influenced by both metro network structure and train schedule while the latter one is mostly decided by users themselves. On one hand, different metro network structures will lead to different route choices. For example, passengers would like to select those routes with few transfers. On the other hand, the train schedule will also influence passenger behaviors. Coordinated transit line could reduce passengers' waiting time in transfer stations. The routes with coordinated transit line should be more attractive than those without coordinated transit line.

So far, many scholars have modeled, analyzed, and studied the problem of passenger route choice behavior within private transportation, such as Kato et al. [1]. Unlike private transportation, metro trains are operated according to the train schedule, leading metro passengers' traveling to be restricted to the schedule. Therefore, traditional methods used in private transportation are not applicable for analyzing metro passenger behavior. Hence, the researchers tried to adopt some technologies widely used in metro transportation into the metro passenger behavior analysis. Among them, AFC (Automatic Fare Collection) system can collect these smart card data about passenger swipe inbound and outbound time of stations, which is useful for analyzing passenger behavior. A lot of research has been done to analyze passenger route choice based on smart card data. However, passengers with different walking time and waiting time may select the same route as metro trains' arrival and departure are dispersed. Hence, passengers walking time to/from platforms

and waiting time on platforms which were included in the smart card data should be useless for the analysis of passenger route choices.

This paper aims to propose a new method to analyze metro passenger route choice over travel periods based on smart card data and train schedule. Firstly, smart card data are trimmed using train schedule to eliminate walking time to/from platforms and waiting time. Then synchronous clustering algorithm, a kind of cluster algorithm, is applied to analyze passenger route choice based on these preprocessed data. Finally, a case study is carried out on the Shanghai metro network to validate the proposed algorithm.

2. Literature Review

Traditional methods on passenger behavior can be classified by Wardrop Law (Liu et al. [2]) as nonequilibrium model and equilibrium model (Smith et al. [3]). They believed that passengers' trip preference depends on travel time perception while individuals' perceptions are different. Some scholars put forward the stochastic user equilibrium model (stochastic user equilibrium (SUE)) to describe the problem. A simulation method was used to realize random users equilibrium model, and experiments were carried out in a large scale urban rail transit network (Kato et al. [1]). With the continuous expansion of parameter types and network sizes, SUE model has been becoming more and more complex for the reality (Thomas [3], Cascetta [4]). However, some scholars found that the traditional models may have some defects when they are applied in metro transportation. The main reason is that passengers' travel routes are affected by metro train schedule; that is to say, metro passengers' arrival and departure are limited to trains' arrival and departure. Thus the applicability of these traditional models is questioned.

The AFC system has been put into application in many metro systems worldwide. AFC system can record these data including passenger inbound swiping time, outbound swiping time, and some other related information. These data are useful in analyzing the passengers' route behaviors in metro. Pelletier [6] divided the usage of smart card data into three categories, long-term planning service, short-term planning service, and operation planning service. For example, swipe card data can be used to forecast the passenger flow OD matrix (Munizaga and Palma [7, 8]), to deal with demand analysis (Morency et al. [9]), to carry on operation and management of rail transit planning (Utsunomiya et al. [10]), etc.

Specifically, smart card data are getting more attention and more research has been made recently. Chan [11] put forward two research ideas based on London metro transit Oyster card data: one was to estimate the OD traffic matrix and the other was to build the metro transit service reliability matrix. This is the first time to use historical card data to make metro transit service quality evaluation. The main application of smart card is to analyze passenger travel behavior. For example, Kusakabe et al. [12] proposed a method to predict the specific trains that passengers choose to ride by using a vast number of long-term history swipe data and parameters. Zhu et al. [13] proposed a method to calibrate the metro passenger behavior model using the AFC data with the genetic algorithm and parameter estimation combining technology. Zhu et al. [14] presented a methodology for assigning passengers to individual trains using both smart card data and AVL data from train tracking systems; it can estimate the probability of the passenger boarding each feasible train and the probability distribution of the number of trains a passenger is unable to board due to capacity constraints. Ma et al. [15] developed a data mining method to identify the spatiotemporal commuting patterns of Beijing public transit riders using transit smart card data. Hong et al. [16] proposed a methodology for assigning passenger flows on a metro network based on Automatic Fare Collection (AFC) data and realized timetable. Briand et al. [17] analyzed the behavioral habits of public transport passengers using a real dataset of smart card data covering a period of five years. Faroqi et al. [18] investigated the relationship between passengers' spatial and temporal characteristics with a novel passenger-based perspective using smart card data. It is implemented for four-day smart card data including 80,000 passengers in Brisbane, Australia. Similarly, Zhu et al. [19] presented an integrated framework for estimating individual passenger's train choices through a data-driven approach with real timetable and Automatic Fare Collection (AFC) data. Besides, smart card data can also be used for estimation or prediction. For example, Hörcher et al. [20] presented a comprehensive method to estimate the user cost of crowding in terms of the equivalent travel time loss with large scale smart card, in a revealed preference route choice framework. Zhao et al. [21] developed a methodology for predicting daily individual trip making and trip attributes using transit smart card data, and the methods are tested using transit smart card data of 10,000 users in London. Also, smart card data are used to make metro train schedule. Zhang et al. [22] proposed a novel method to optimize the skip-stop scheme for bidirectional metro lines using the time-dependent passenger demand extracted from smart card data, so that the average passenger travel time can be minimized.

Some recent studies have made some progress on analyzing passenger behavior based on smart card data, part of which are useful for realistic size networks. The specific focus of this paper is to propose a method specifically aimed at using a small number of parameters, so that it can be easily used for large scale networks. Hence, this paper uses data analysis methods, i.e., ***cluster algorithm***, to analyze the passenger route choice behaviors on metro networks. The cluster algorithm is a method of multivariate statistical analysis. Data are classified according to individual characteristics so that the data in the same category have the highest homogeneity. On the other hand different categories should have relatively higher heterogeneity. The cluster algorithm aims to analyze and mine the intrinsic structure and rules of given data [23, 24]. In the process of data clustering, the clustering algorithm can automatically divide data points into different sets according to the attributes. These data with similar attributes are divided into the same set, while these data points with different attributes are divided into different sets [25]. Clustering algorithms can be divided into several types: clustering algorithms based on division (i.e., K-means),

clustering algorithms based on density (i.e., DBSCAN and OPTICS), affinity propagation clustering algorithm (affinity propagation (AP) algorithm), synchronous clustering algorithm (SynC algorithm), etc.

K-means algorithm is the most widely used clustering algorithm based on division. It has been nearly 60 years since it was proposed [26]. However, the biggest shortcoming of the K-means algorithm is to select the initial K value and the value of the selected K data points since the initial value may lead the convergence of the K-means algorithm to different results. Hence, many scholars proposed other new clustering algorithms, among which AP algorithm is one kind of typical clustering algorithms [27]. AP clustering algorithm does not need to specify the number of clusters in advance. ***Synchronous clustering algorithm*** (SynC algorithm) [28, 29] is another kind of clustering algorithm of which initial values are not sensitive. The main idea of synchronous clustering is that each data point is regarded as an independent individual, and similar individuals automatically get together to form clustering collections. Due to the characteristics of synchronous clustering algorithm, this algorithm has many advantages; for example, (1) the algorithm does not require given cluster centers in advance, (2) the algorithm is not sensitive to the initial value, and (3) the algorithm can well avoid noise interference data.

However, to our best knowledge, no studies adopted the SynC algorithm to analyze metro passenger route choices with smart card data trimmed by train schedules. Hence, taking the advantages of the synchronous clustering algorithm (SynC) into consideration, this paper adopts the SynC algorithm to analyze metro passenger behavior.

3. Methodology

3.1. Basic Assumptions. Some necessary assumptions and elements are firstly described as follows:

(1) All passengers' behaviors are assumed to be reasonable, and passengers would not stay in stations for a long time. But there are always some unreasonable data which spend a very long time or an extremely short time during given OD pairs. This proposed algorithm will regard these data as noise data in the dataset.

(2) Train congestion is not considered in data preprocessing. It means passengers can ride the first arriving train after they reach platforms.

(3) All trains are operated according to the train schedule strictly.

3.2. Definition of Train Schedule and Smart Card Data

3.2.1. Train Schedule. The metro train schedule contains necessary information of all trains running on the network, like train codes, arrival and departure time of trains at each station, etc. Figure 1 shows an example of a train schedule used by a metro line in Shanghai. Each red line represents a planned operation train.

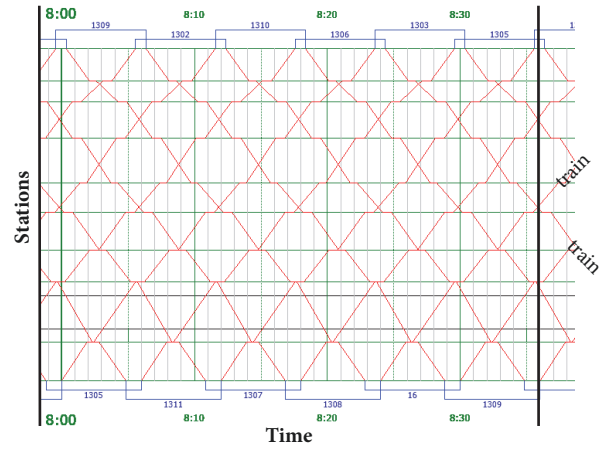


Figure 1: An example of the train schedule.

The definition of train schedule is described below: metro line is defined as $L = \{1, 2, \ldots, l, \ldots, N\}$, and the station collection on line l is $S_l = \{1, 2, \ldots, i, \ldots, M\}$. Then, station $S_{l,i}$ represents the station i in line l; $S_{l,i}(A^j_{l,i}, D^j_{l,i})$ defines the arrival time $A^j_{l,i}$ and departure time $D^j_{l,i}$ of train j at $S_{l,i}$. Thus, the trajectory of train j is described as $\{\forall i \in l \mid S_{l,i}(A^j_{l,i}, D^j_{l,i})\}$, and the network train schedule can be described as $T = \{\forall j, l, i \mid S_{l,i}(A^j_{l,i}, D^j_{l,i})\}$.

3.2.2. Smart Card Data. AFC system can record the original station (O is used in this paper), destination station (D is used in this paper), and their corresponding inbound and outbound time. These swiping data can be used to obtain the detailed passenger flow demand. Table 1 shows some examples of entry and exit swiping card data recorded by the AFC system, like card number, swiping date, inbound station code, inbound swiping time, outbound station code, outbound swiping time, etc.

Smart card data (AFC data) are defined as $OD(n, T^{(si)}, S^{(in)}, T^{(so)}, S^{(out)})$, in which n is the card ID, $T^{(ci)}$ is the inbound swiping time, $T^{(co)}$is the outbound swiping time, $S^{(in)}$ is the O station, and $S^{(out)}$ is the D station.

3.2.3. Passenger Travel Process on Metro. Figure 2 shows the metro passenger travel process. It displays typical metro passenger traveling, which mainly contains passengers' swiping card at entry gates, walking to platforms, waiting for coming trains, riding trains (transfer if it has), and finally walking out of station. As shown in the figure, symbol definition includes walking cost time (entry walking time, $T^{(in)}$), waiting cost time (waiting time on platforms, $T^{(w)}$), travel cost time (in-vehicle time, $T^{(v)}$), and walking out of station cost time (exit walking time, $T^{(out)}$). If a passenger makes a transfer, the additional transfer walking cost time (transfer walking time, $T^{(t)}$) and transfer waiting cost time (waiting time, $T^{(tw)}$) are required.

Table 1: Samples of smart card data.

Date	Card ID	O station	Inbound Time	D station	Outbound Time
2014-11-17	1416107917	0248	09:15:45	1056	09:36:00
2014-11-17	1282520204	0751	09:20:00	0727	09:36:17
2014-11-17	0934484109	1060	09:13:56	0248	09:36:22
2014-11-17	1069233288	0411	09:22:54	0750	09:36:41
.....					

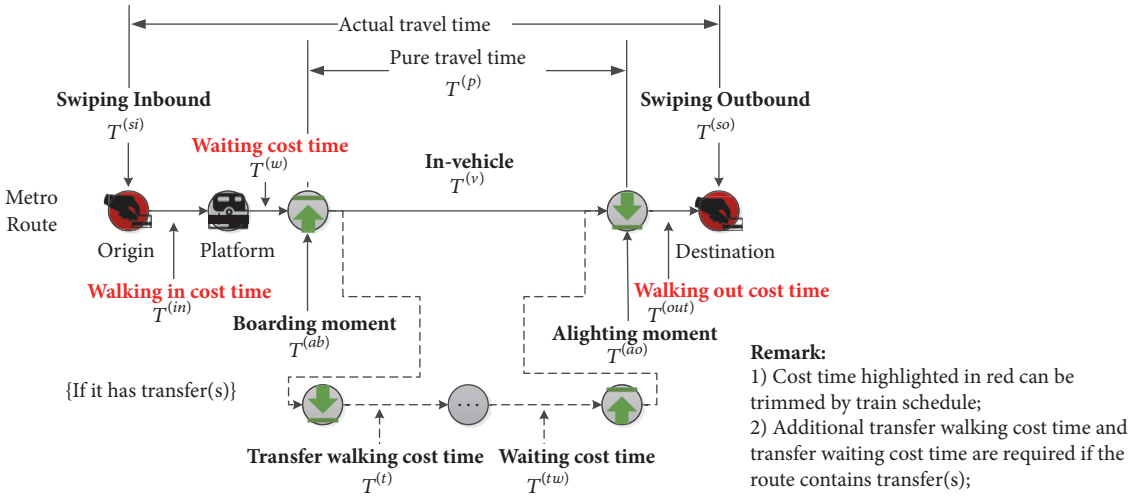


Figure 2: Passenger trip diagram by metro transit.

Here, $T^{(si)}$(inbound swiping time) is defined as the moment passengers swipe in stations. $T^{(so)}$(outbound swiping time) is defined as the moment passengers swipe out of stations. The difference between $T^{(si)}$ and $T^{(so)}$ is the passengers' actual travel time during metro. Besides, $T^{(ab)}$ (actual board time) is defined as the actual moment when passengers board trains, while $T^{(ao)}$ (actual alight time) refers to the actual moment when passengers alight trains. Then, the pure travel time (pure travel time, $T^{(p)}$) is the difference between $T^{(ab)}$ and $T^{(ao)}$. It is obvious that the values of $T^{(ab)}$ and $T^{(ao)}$ are limited to train arrival, which is related to the train schedule.

3.3. AFC Data Trimmed by Train Schedule. The passengers' travel time by metro (actual travel time is used in this paper) can be obtained from the difference between the inbound swiping time and the outbound swiping time from smart card data. Obviously, the actual travel time could be different in one OD pair if passengers select different route. When the difference of route travel time between OD pairs is large, passenger's selected route can be easily decided based on the travel time. However, smart card data contains inbound and outbound walking time and waiting time, which are useless information. Since trains' arrival at stations is dispersed, some passengers with different walking time may take the same trains. That is to say, some passengers may take the train just after they arrive at platforms, while some passengers may wait for a long interval for a train they just miss. Thus, the travel time without waiting time and walking time at O station and D station can present more useful information than the travel time with waiting and walking time.

We could use train schedule to trim smart card data by removing walking and waiting time at O stations and walking time at D stations. The trimmed result can be used in cluster algorithm, subsequently. Figure 3 shows some passenger travel time before and after using AFC data trimming algorithm. It can be seen that the original AFC data are out of order, while these data after trimming are orderly. The pure travel time could reflect some discrete characteristics of train arrival and departure.

The method to determine passengers' actual boarding and alighting time is shown in Figure 4. First, for each AFC data, its inbound station is set as $S^{(in)} = S_{l,i}$, and its inbound time is set as $T^{(si)}$. Find train j based on the following equation after searching all trains which run pass $S_{l,i}$ in order:

$$D_{l,i}^{j-1} \le T^{(si)} \le D_{l,i}^{j} \tag{1}$$

It means that passengers can ride train j to their destinations or transfer stations. Thus the possible actual board time $T^{(ab)}$ is

$$T^{(ab)} \longleftarrow D_{l,i}^{j} \tag{2}$$

Similarly, the actual alighting time can be obtained in the same way. Its outbound station is set as $S^{(out)} = S_{l,i'}$, while its outbound time is set as $T^{(so)}$. Find train j' with the following

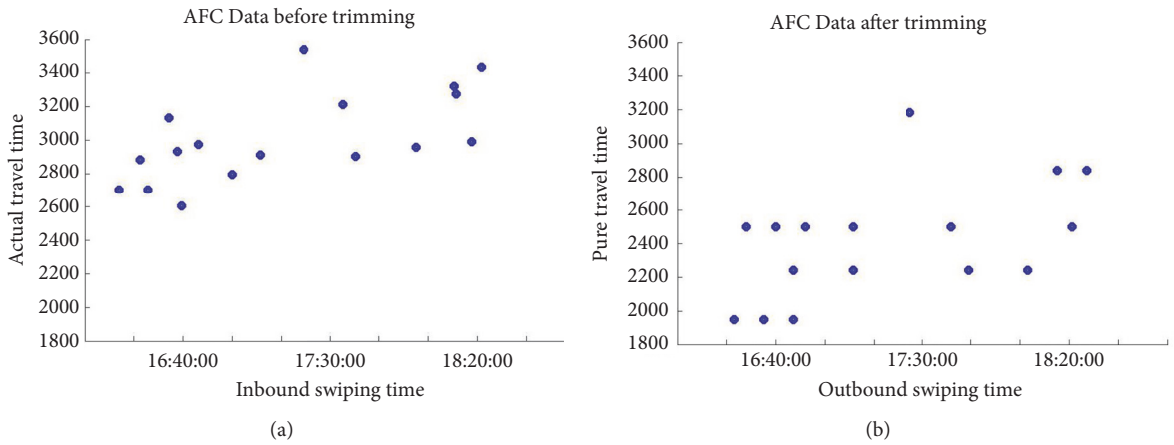


Figure 3: A sample of AFC data before and after trimming.

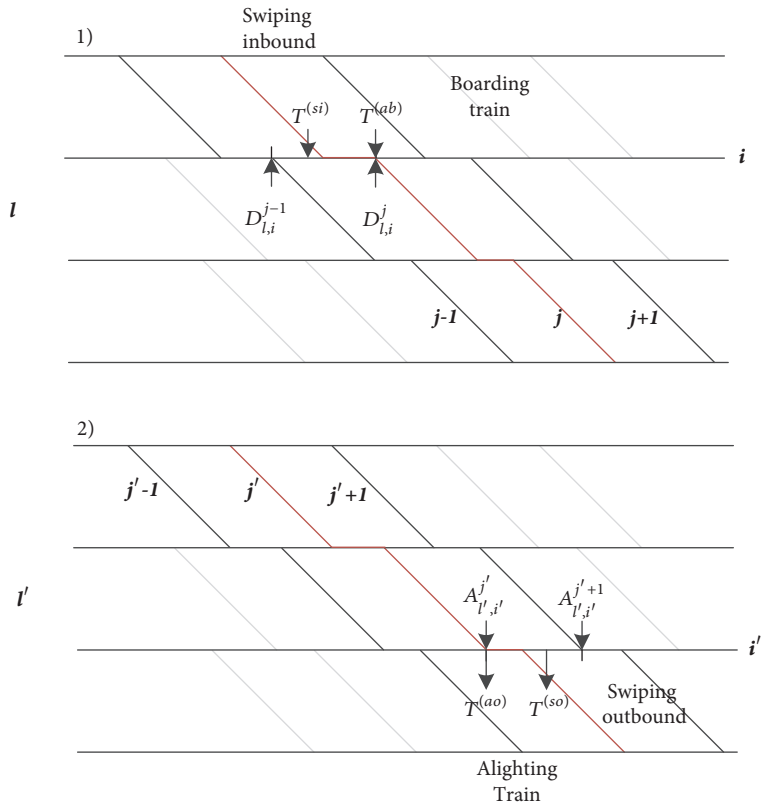


Figure 4: Determination of passengers' actual boarding (1) and alighting (2) time.

equation after searching all trains which run pass $S_{l',i'}$ in reverse order:

$$A^{j'}_{l',i'} \le T^{(so)} \le A^{j'+1}_{l',i'} \tag{3}$$

Thus the possible actual board time $T^{(ao)}$ is

$$T^{(ao)} \longleftarrow A^{j'}_{l',i'} \tag{4}$$

It should be noted that a least walking time is needed to enter into or exit from the platform by gates. The minimum time constraint ε is considered in $T^{(ab)}$ and $T^{(ao)}$ as follows:

$$D^{j-1}_{l,i} \le T^{(si)} + \varepsilon \le D^{j}_{l,i} \tag{5}$$

$$A^{j'}_{l',i'} \le T^{(so)} + \varepsilon \le A^{j'+1}_{l',i'} \tag{6}$$

Therefore, the pure travel time can be acquired by

$$T^{(p)} = T^{(ao)} - T^{(ab)} \tag{7}$$

3.4. SynC Algorithm. Based on the pure travel time, this paper applies SynC algorithm analysis to process these data. This part presents how to use the SynC algorithm to analyze metro passenger route choice.

3.4.1. Data Normalization. Before cluster, the data need to undergo normalization since data points may have different scales and dimensions which will affect the effectiveness of clustering algorithm. Data normalization is firstly adopted to make data fall into a certain range. This paper wants to make inbound swiping time and pure travel time into the same certain range to carry on the cluster.

Z-score normalization is used in this paper to carry on data normalization, which is based on the mean and standard deviations of attribute values. The advantage of Z-score normalization is that it does not need to compute the maximum and minimum values of the data set and has good effects on the normalization of outliers. Its formula is

$$\widetilde{v_i} = \frac{v_i - \overline{v}}{\sigma} \tag{8}$$

where $\overline{v}$ is the mean value of attribute value, and σ is the standard deviation of attribute values.

3.4.2. Synchronous Clustering Algorithm (SynC Algorithm). The main idea of SynC algorithm is to regard each data point as an individual, and the similar points would get clustered. The procedure of the algorithm is shown in Figure 5: firstly, data points are independent and move close to their similar data points, as shown in Figure 5(a)); secondly more and more data points will gather together to the one with same attribute, as shown in Figure 5(b)); finally, all similar data points are clustered together to form a cluster center, while some noise data are automatically isolated, as shown in Figure 5(c)).

Some equations should be given in SynC algorithm.

Definition 1 (domain distance ϵ). It means the maximum distance from the given point.

Definition 2 (Nb_ϵ (the ϵ collection of data point x)). Let x be a data point of data set D; Nb_ϵ means the data whose distance from x is smaller than ϵ:

$$Nb_\epsilon = \{y \in D \mid dist(y, x) \le \epsilon\} \tag{9}$$

where $dist(y, x)$ is the distance between data points x and y.

Definition 3 (Kuramoto Amplitude of data point x). Let x_i be the ith dimension of data point x. After it is influenced by other points in Nb_ϵ, the Kuramoto Amplitude of data point x_i can be described as

$$\frac{d_{x_i}}{dt} = \omega_i + \frac{S}{|Nb_\epsilon|} \sum_{y \in Nb_\epsilon(x)} \sin(y_i - x_i) \tag{10}$$

where ω can be ignored in this cluster algorithm, and S is a constant (equal to 1 in this part). Finally, the Kuramoto Amplitude can be rewritten as

$$x_i(t+1) = x_i(t) + \frac{1}{|Nb_\epsilon|} \sum_{y \in Nb_\epsilon(x)} \sin(y_i - x_i) \tag{11}$$

where t is the time step, and $t = 0$ represents the initial state.

Definition 4 (synchronous coordination parameter). It represents the degree of synchronous coordination of all data points in the data set at the current time step:

$$r_c = \frac{1}{N} \sum_{i=1}^{N} \sum_{y \in Nb_\epsilon(x)} e^{-(y_i - x_i)} \tag{12}$$

It can be seen that synchronous coordination parameter of the data set will increase gradually when more data points gather together. And after the parameter does not change for a long time, the data set achieves convergence within Nb_ϵ. It reaches a local synchronized status. Finally, when all data points gather together ($r_c \longrightarrow 1$), it reaches a global synchronized status.

Definition 5 (optimal domain distance ϵ). It means the cluster result is the best when ϵ is equal to a certain value. The optimal distance can be determined according to the SynC algorithm [28]:

$$M^j = \underset{j}{\operatorname{argmin}} L\left(D, M^j\right) \tag{13}$$

where M^j is the jth cluster center of the given data; argmin_j is the function that can calculate j which leads the value of $L(D, M^j)$ to be minimum.

$L(D, M^j)$ can be computed by following equations:

$$L\left(D, M^j\right) = L(M) + L(D \mid M) \tag{14}$$

$$L(M) = \sum_{i=1}^{K} \sum_{j=1}^{|C_i|} \log_2\left(\frac{N}{|C_i|}\right) + \sum_{i=1}^{K} \frac{p_i}{2} \log_2\left(\frac{N}{|C_i|}\right) \tag{15}$$

$$L(D \mid M) = -\sum_{i=1}^{K} \sum_{x \in C_i} \log_2\left(pdf(x)\right) \tag{16}$$

where K is the number of cluster centers; C_i is the ith cluster set; $|C_i|$ is the number of data points in C_i; p_i is the data dimension; $pdf(x)$ is the probability of data point x which belongs to C_i.

Therefore, the steps of synchronization clustering algorithm (SynC algorithm) are described as follows, while the flowchart of SynC algorithm is shown in Figure 6:

(1) Initial time step is set as $t = 0$, and all data points are regarded as independent cluster center.

(2) Set domain distance ϵ, and calculate Nb_ϵ of all data points.

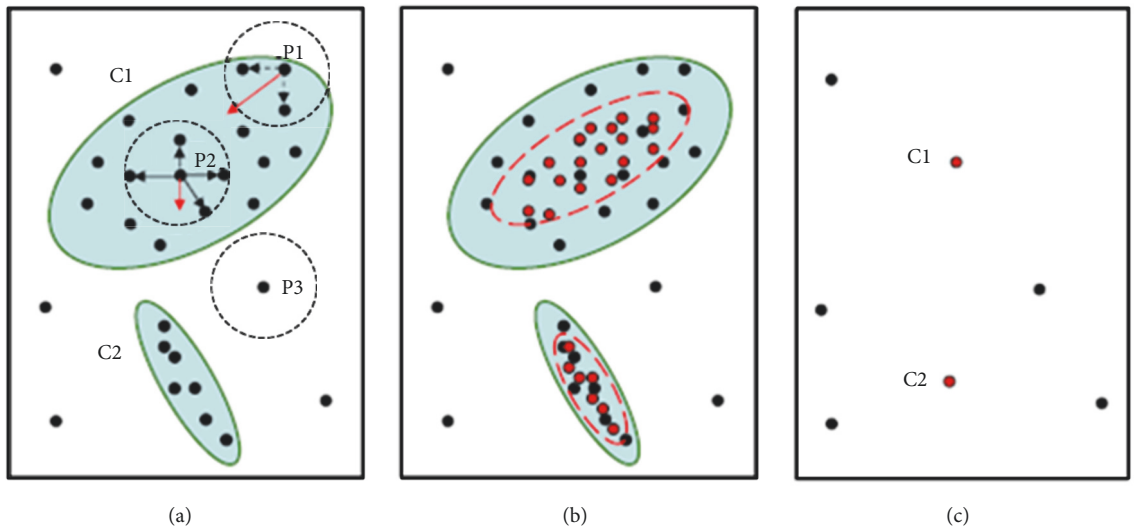


Figure 5: Sketch of synchronous clustering (SynC) algorithm process [28].

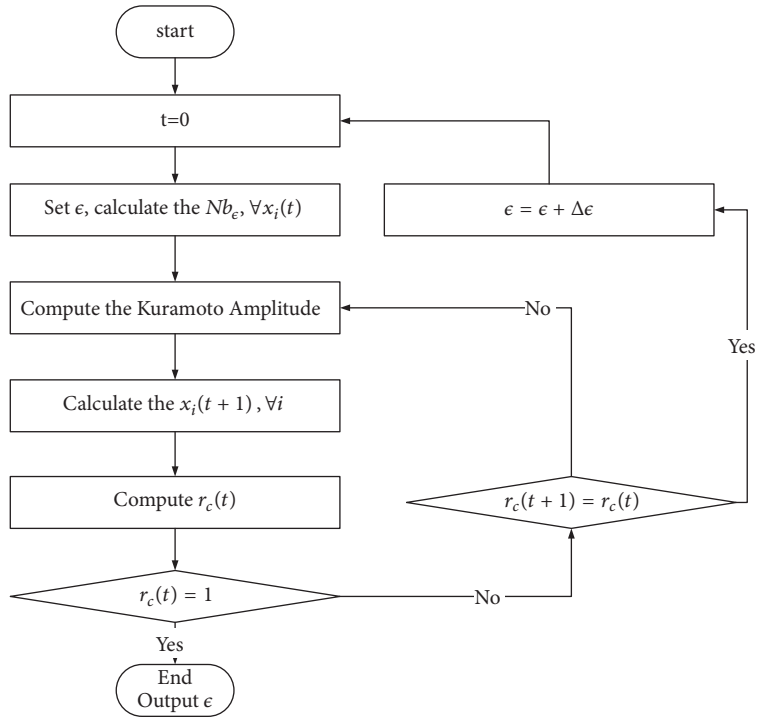


Figure 6: Flowchart of SynC algorithm.

(3) Compute the Kuramoto Amplitude of all data points using Nb_ϵ, and data points of $x_i(t+1)$ can be calculated when it moves to next time step ($t = t+1$).

(4) Compute the synchronous coordination parameter r_c of this data set at this time step.

(5) If $r_c = 1$, then it reaches a global synchronized status, the algorithm ends and the optimal domain distance ϵ can be computed. If this is not the case, the algorithm moves to step (6).

(6) If r_c remains the same ($r_c(t+1) = r_c(t)$), then it reaches a local synchronized status. Let $\epsilon = \epsilon + \Delta\epsilon$, $t = 0$, move to step (2), and start a new cluster. Otherwise, move to step (3) and continue this cluster.

4. Case Study

To evaluate the proposed algorithm of smart card data trimming and SynC, a real-life metro network (the Shanghai Metro system, shown in Figure 7) with a large number of

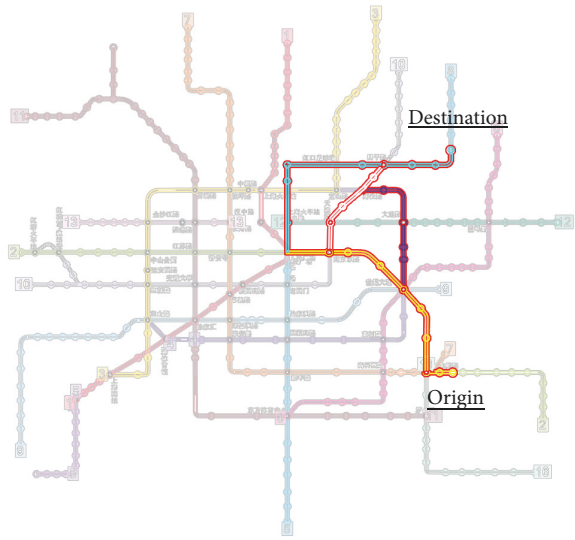


Figure 7: Shanghai metro network.

lines and stations is presented as a case study application. The network consists of 14 transit lines and each has an upstream direction and a downstream direction. There are totally 289 stations in the network, of which 42 stations are transfer stations. Jinke Road Station and Huang Xing Road Station are selected as O and D station in this case. Jinke Road Station is surrounded by working companies, while Huangxing road is located in the residential area. It leads to the fact that there are larger passenger flows in the OD pair during the evening peak.

4.1. Calculation Process

(1) OD Pair. Jinke Road Station (station code 0254) in line 2 is taken as O station and Huangxing Road Station (station code 0844) in line 8 is taken as D station. Whole week data from November 11, 2016, to November 15, 2016, are selected in this case, which had 199 data records in total. (The AFC data were obtained from the Shanghai Metro Company.)

(2) Train Schedules and AFC Data Trimming. To make the case study easy to program, the planned train schedule instead of actual schedule is used. And the planned train schedule using at weekday during November 2016 is applied in the case study, and all trains are assumed to operate according to the train schedule strictly. Train schedule is used to trim AFC data to obtain the pure travel time by removing entry/exit walking time and waiting time according to the proposed AFC data trimmed method. The process is shown in Table 2.

(3) Data Normalization. The inbound swiping time is selected as the X axis of cluster data set and the pure travel time is selected as the Y axis. However, due to the different dimensions of data points, data normalization is needed to get a better cluster result. The normalization example of the data points is shown in Table 3.

(4) Clustering Process. C#.Net programming language is applied to program coding to achieve the algorithm. Figure 8 shows the process of SynC algorithm. The X axis is inbound time after data normalization while the Y axis is pure travel time after data normalization. And two horizontal lines in each figure represent morning and evening peak period, respectively. Each part in Figure 8 represents a local synchronized status in SynC algorithm. At the first part, each data point is regarded as a cluster center/centroid. The data points automatically get together in local synchronized status, leading centroids to be merged slowly in the following parts. It can be seen that, with the clustering process, data points gradually merge to form cluster centers, and noisy data are isolated obviously at the same time, when reaching the optimal domain distance as (13)-(16). The final result is shown in Figure 9. Point color refers to the cluster they belong to. The more the data points of the same color, the higher the passenger flow this route has. Passenger route selection probability during both peak and flat period is easy to obtain with the result.

4.2. Algorithm Analysis.

The cluster algorithm applies the pure travel time which removes entry/exit walking time and waiting time using train schedule. Some comparative analyses are made in this part. Figure 10 shows the cluster results using both AFC data with trimming (Figure 10(a)) and AFC data without trimming (Figure 10(b)) by train schedule. It is indicated that the trimming results could present metro travel time characteristics clearly while the no-trimming results present passenger travel time disorderly. Thus, pure travel time trimmed by train schedules could represent some discrete characteristics of metro transportation since it could take train schedules into consideration.

4.3. Result Analysis.

Table 4 shows cluster results by the distinction of early peak, flat peak, and evening peak. This table shows passengers preference on route choice with different periods. And clusters with small passenger flow are regarded as noisy. Table 5 shows the route list of this OD pair in traditional model used in Shanghai Metro Company [5]. The candidate route sets are generated according to *the K-short algorithm* with route expected travel time, and the selection probability of each route is calculated by logistics model. It contains some possible routes that passengers may choose to follow and the corresponding selection probability of each route. This table is very important to metro operation since it is used to calculate the passenger flow distribution of the whole network. Also the allocation to each metro line is decided by the line passenger flow computed by the traditional model results.

The routes in Table 5 are used to link the cluster centers in Table 4 according to the comparison of travel time. Take Table 5 (route list) as a contrast; the following results can be summarized from Table 4 (cluster result):

(1) There are mainly two routes during morning peak period. About 60% of passengers choose the route with a long time but less transfer (Route No. 3 in Table 5), and 40% of passengers choose the route with

Table 2: Process of AFC data trimming.

Card ID	Inbound time	Outbound time		Card ID	Actual board time	Actual alighting time
2914517671	07:18:57	08:15:27		2914517671	07:24:53	08:11:50
1119250823	07:21:09	08:17:47	Trimming	1119250823	07:24:53	08:16:50
2663388071	07:22:15	08:20:28	⟶	2663388071	07:24:53	08:14:20
3525029280	10:09:43	11:04:26		3525029280	10:14:10	10:59:02
1454658848	16:27:52	17:15:59		1454658848	16:31:50	17:11:19
	.....				.....	

Table 3: Data normalization.

ID	Inbound time	Pure travel time		ID	Inbound time	Pure travel time
1	07:24:53	2817		1	-2.260	-0.382
2	07:24:53	3117	Normalization	2	-2.260	-0.268
3	07:24:53	2967	⟶	3	-2.260	-0.057
4	10:14:10	2692		4	-1.516	-0.652
5	16:31:50	2369		5	0.143	-1.352
	.....				.....	

Table 4: Result of synchronous clustering algorithm (Sync).

Clusters	Normalized X axis	Alighting time∗	Normalized Y axis	Pure travel time∗	Passenger flow (Data count)	Ratio	Remark
Morning Peak							
1	-2.165	7:46:28	0.225	3097.2	17	58.6%	
2	-2.055	8:11:22	-1.051	2508.0	12	41.4%	
Flat Period							
3	-1.301	11:03:05	-0.232	2886.0	9	36.0%	
4	-1.065	11:56:50	0.737	3333.7	3	12.0%	Noisy
5	-0.719	13:15:37	1.769	3810.3	2	8.0%	Noisy
6	-0.404	14:27:06	-0.328	2841.9	11	44.0%	
Evening Peak							
7	0.464	17:44:50	1.138	3519.0	19	13.1%	
8	0.474	17:46:59	3.923	4805.2	2	1.4%	Noisy
9	0.485	17:49:36	-0.033	2978.3	88	60.7%	
10	0.763	18:52:50	-0.947	2555.8	36	24.8%	

Table 5: Route list of the OD pair according to traditional model [5].

Route ID	Origin	Destination	Pass Line	Travel time	Selection probability
1	0254	0844	2-10-8	2509	50%
2	0254	0844	2-4-10-8	2621	33%
3	0254	0844	2-8	2981	17%

a short time (Route No. 1 in Table 5). This result is not similar to Table 5. It is a bit surprising that not all passengers selected the route with the shortest travel time. The possible reasons for selecting the route with a longer travel time but less transfer during peak period are that passengers may want to avoid station congestion. Station congestion may lead to the fat that they miss the first arrival train because of not enough space in vehicle and too many passengers on the platform. Thus, passengers may think the transfer could take them more time in their trips during morning peak.

(2) As shown in Table 4, the difference between cluster 3 and cluster 6 is small; thus these two clusters can actually be considered the same one. After linking clusters to routes, we could find that most passengers choose Route No. 1 or Route No. 2, while few choose Route 3 during flat period, which is in line with the result of Table 5. The results can reveal the fact that passengers do not expect shorter travel time but expect more comfortable service instead during flat period. Besides, many noisy points could be found during flat periods, which is relative long travel time with less passenger flow in metro system. It means

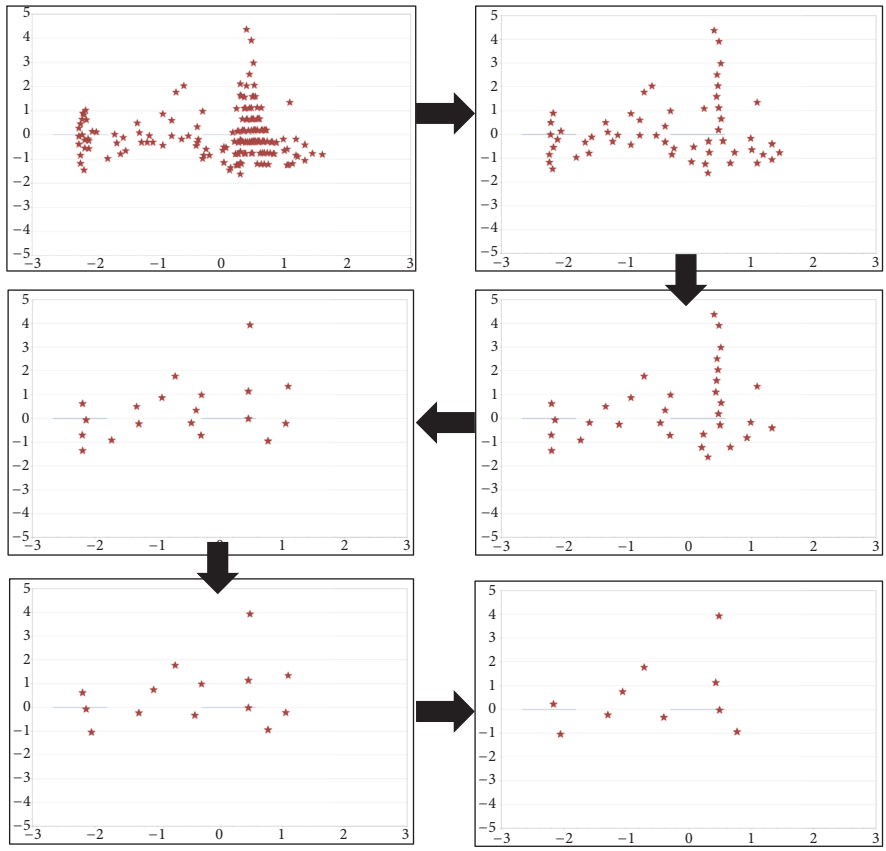


Figure 8: Process of synchronous clustering (SynC) algorithm.

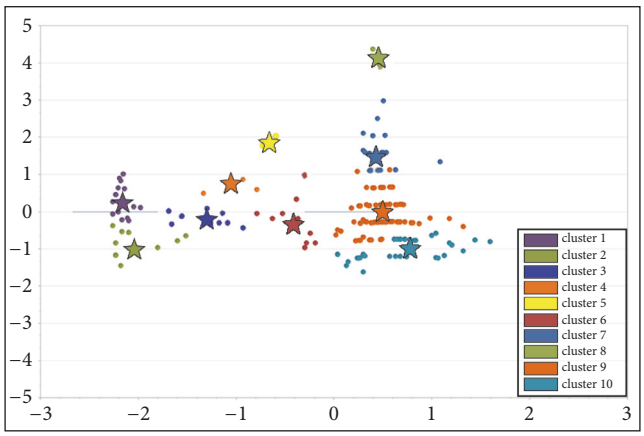


Figure 9: Result of synchronous clustering (Sync) algorithm.

passengers may not be in a hurry on their trips during flat periods.

(3) Passenger flow mainly occur during evening peak from Jinke road to Huangxing road, accounting for more than 70% of the total passenger flow on a whole day. The result shows that passengers are not sensitive to travel time during evening peak. A small number of passengers (about 25%) select Route No. 1 and Route No. 2, while the majority of the passengers (13%+61%) choose Route No. 3, which is different from the results in Table 5. It may be because that passenger prefer to travel with less transfers during evening peak.

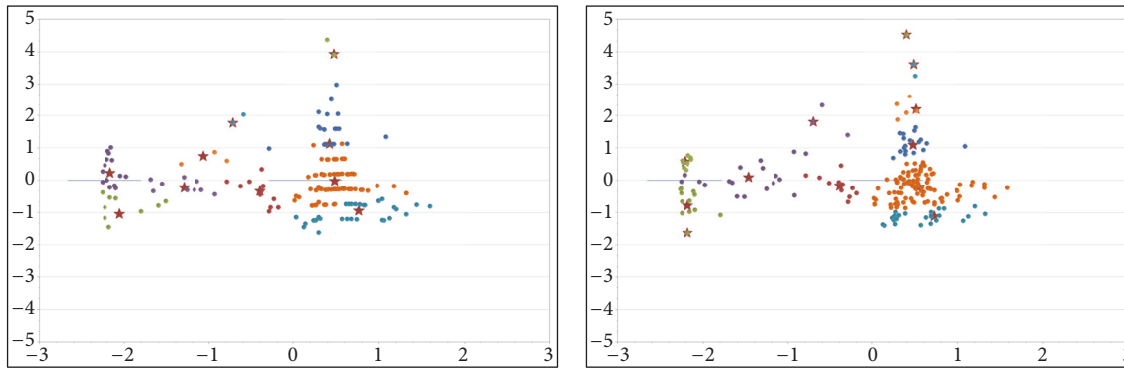
Metro passenger route choices could be various for their travel time and their inbound time, especially for peak and flat period. In this case study, passengers are more likely to choose Route No. 3 (with less transfer) during morning and evening peak while passengers are more likely to choose Route No. 1 or Route No. 2 (with less travel time) during flat period. Passengers' route choices may be influenced by both their travel moment and travel cost time.

It should be noted that smart card data with only a week range are used in the case study. The passenger route selection probability would be more reliable with more smart card data. Therefore, the result of the algorithm can be used to revise traditional model results like those of Table 5.

4.4. Algorithm Extension. The proposed algorithm can be applied to other OD pairs easily on metro network. But there are two limitations in the algorithm of AFC data trimmed by train schedule when it is used for other OD pairs. The one is that passengers can choose any transit line to finish their trips when origin station or destination station is a transfer station. The other one is that passengers can choose either upstream trains or downstream trains of the transit line to finish their trips when origin station or destination station

Table 6: Applicability of the proposed algorithm in Shanghai metro network.

Situation ID	Situation Name	OD pairs count	OD pairs percent	Passenger flow count	Passenger flow percent
		119918	100%	5313949	100%
1	OD pairs with unclear routes	23390	19.50%	606245	11.41%
2	OD pairs with similar travel time	7193	6.00%	98003	1.84%
3	OD pairs with small passenger flow	25737	21.46%	26810	0.50%
4	Noise data (Same in and out)	4112	3.43%	95712	1.80%
5	applicable OD pairs	53233	49.61%	4520205	84.44%



(a) (b)

Figure 10: Comparison of cluster results after and before data trimming.

is a normal station. These two kinds of OD pairs are called *unclear routes OD pairs*. The algorithm of AFC data trimmed by train schedule cannot be applied to these two kinds of OD pairs. The main reason is that passengers' walking in cost time or walking out cost time is not able to be removed from AFC data since it is not clear which transit line or which upstream/downstream trains of metro schedule should be chosen.

There are also two limitations in the SynC algorithm when it is used for other OD pairs. One is that it is useless to apply the algorithm into the OD pairs whose travel time of routes is similar. Routes cannot be clustered by these similar travel times. The other is that the SynC algorithm is useless when the passenger flow is very low between the OD pairs. The cluster algorithm cannot work with such less data.

Shanghai metro network is applied to discuss the applicability of the proposed algorithm, as shown in Table 6. The date of selected data is November 15, 2016. There are 119,918 OD pairs in this network and 5,313,949 passengers traveling on that day. Some interesting findings can be obtained as follows:

(1) There are about 20% OD pairs (23,390, 19.50%) in which passengers can ride more than one line to their destinations. These OD pairs contain 1) both upstream and downstream lines which are both feasible routes and 2) many routes which are feasible in original/destination stations when they are transfer ones; thus the proposed algorithm of AFC data trimming by train schedule cannot be used in this type of OD pairs, accounting for 11.41% passenger flow (606,245).

(2) There are about 6.00% OD pairs (7,193) having similar travel time. The SynC algorithm cannot use these data to cluster distinct points.

(3) However, more than 20% OD pairs (25,737) contain less than 5 passengers on a whole day. Such small passenger flow is useless for cluster. But there are only 26,810 passengers (accounting for 0.50%) traveling though these OD pairs.

(4) AFC data also contains some noisy data; for example, passengers may swipe in and swipe out from same stations. There are about 4,112 OD pairs (3.43%), accounting for passenger flow of 95,712 (1.80%). These data are useless for the analysis of passenger behavior among metro system.

(5) Therefore, besides those above data, the proposed algorithm can be used in about 50% OD pairs (53,233, 49.61%) to cluster passenger travel routes, while more than 85% AFC data (passenger flow, 4,520,205, 84.44%) can be used in clustering. That is to say, most

AFC data are useful for the analysis of passenger route behavior.

5. Conclusion

This paper studied metro passenger route choice with train schedule and cluster algorithm. On the basis of AFC data, the algorithm of AFC data trimmed by train schedule was proposed to obtain pure travel time. The results were then used in synchronous clustering algorithm to analyze the passenger route choice (selection probability) under time constraints. Then, a case study by using Shanghai metro data was conducted to validate the proposed algorithm. It was indicated that the probability of route choice can be calculated through SynC algorithm in different periods, and thus the algorithm can be used to revise traditional model results. The proposed algorithm can help to analyze passenger route preference with smart card data without traditional methods which contains a large number of parameters. And the passenger route preference would be relatively accurate with more smart card data.

However, there are some limitations in the proposed method which needs further research. (1) The journey time of different routes over different periods should be different. The travel time in Table 5 in the results of the paper has theoretical values, which are calculated by train section running time and passenger transfer cost time. The results do not consider congestion in trains and variable train operation headway in the calculation. Thus, further research should be made in the determination of dynamic travel time of passenger routes over periods. For example, congestion data in the train carriages and on the platforms of stations, which can be acquired by passenger flow detection devices based on image recognition, are useful for calculating dynamic journey time of passenger routes over periods. (2) Besides, how to link these clustering results to these travel routes automatically needs a further study, in order to make the data process more complete.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This paper is supported by the Research Projects of the Social Science and Humanity on Young Fund of the Ministry of Education under Grant no. 15YJCZH108 and the Research Projects of Natural Science Foundation of Guangdong Province under Grant no. 2015A030310341.

References

[1] H. Kato, Y. Kaneko, and M. Inoue, "Comparative analysis of transit assignment: evidence from urban railway system in the Tokyo Metropolitan Area," *Transportation*, vol. 37, no. 5, pp. 775–799, 2010.

[2] Y. Liu, J. Bunker, and L. Ferreira, "Transit userś route-choice modelling in transit assignment: a review," *Transport Reviews*, vol. 30, no. 6, pp. 753–769, 2010.

[3] R. Thomas, "Traffic assignment techniques," 1991.

[4] E. Cascetta, *Transportation Systems Analysis: Models and Applications*, Springer, 2009.

[5] R.-H. Xu, Q. Luo, and P. Gao, "Passenger flow distribution model and algorithm for urban rail transit network based on multi-route choice," *Journal of the China Railway Society*, vol. 31, no. 2, pp. 110–114, 2009.

[6] M.-P. Pelletier, M. Trépanier, and C. Morency, "Smart card data use in public transit: a literature review," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 4, pp. 557–568, 2011.

[7] M. A. Munizaga and C. Palma, "Estimation of a disaggregate multimodal public transport Origin-Destination matrix from passive smartcard data from Santiago, Chile," *Transportation Research Part C: Emerging Technologies*, vol. 24, pp. 9–18, 2012.

[8] M. Munizaga, F. Devillaine, C. Navarrete, and D. Silva, "Validating travel behavior estimated from smartcard data," *Transportation Research Part C: Emerging Technologies*, vol. 44, pp. 70–79, 2014.

[9] C. Morency, M. Trépanier, and B. Agard, "Measuring transit use variability with smart-card data," *Transport Policy*, vol. 14, no. 3, pp. 193–203, 2007.

[10] M. Utsunomiya, J. Attanucci, and N. Wilson, "Potential uses of transit smart card registration and transaction data to improve transit planning," Transportation Research Record 1971, Transportation Research Board of the National Academies, Washington, DC, USA, 2006.

[11] J. Chan, *Rail transit OD matrix estimation and journey time reliability metrics using automated fare data*, Massachusetts Institute of Technology, 2007.

[12] 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.

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

[14] Y. Zhu, H. N. Koutsopoulos, and N. H. M. Wilson, "A probabilistic Passenger-to-Train Assignment Model based on automated data," *Transportation Research Part B: Methodological*, vol. 104, pp. 522–542, 2017.

[15] X. Ma, C. Liu, H. Wen, Y. Wang, and Y.-J. Wu, "Understanding commuting patterns using transit smart card data," *Journal of Transport Geography*, vol. 58, pp. 135–145, 2017.

[16] L. Hong, W. Li, and W. Zhu, "Assigning Passenger Flows on a Metro Network Based on Automatic Fare Collection Data and Timetable," *Discrete Dynamics in Nature and Society*, vol. 2017, 2017.

[17] A.-S. Briand, E. Côme, M. Trépanier, and L. Oukhellou, "Analyzing year-to-year changes in public transport passenger behaviour using smart card data," *Transportation Research Part C: Emerging Technologies*, vol. 79, pp. 274–289, 2017.

[18] H. Faroqi, M. Mesbah, and J. Kim, "Spatial-temporal similarity correlation between public transit passengers using smart card data," *Journal of Advanced Transportation*, vol. 2017, 2017.

[19] W. Zhu, W. Wang, and Z. Huang, "Estimating train choices of rail transit passengers with real timetable and automatic fare

collection data," *Journal of Advanced Transportation*, vol. 2017, Article ID 5824051, 12 pages, 2017.

[20] D. Hörcher, D. J. Graham, and R. J. Anderson, "Crowding cost estimation with large scale smart card and vehicle location data," *Transportation Research Part B: Methodological*, vol. 95, pp. 105–125, 2017.

[21] Z. Zhao, H. N. Koutsopoulos, and J. Zhao, "Individual mobility prediction using transit smart card data," in *Transportation Research Part C: Emerging Technologies*, vol. 89, pp. 19–34, 2018.

[22] P. Zhang, Z. Sun, and X. Liu, "Optimized Skip-Stop Metro Line Operation Using Smart Card Data," *Journal of Advanced Transportation*, 2017.

[23] K.-L. Du and M. N. S. Swamy, "Clustering i: Basic clustering models and algorithms," in *Neural Networks and Statistical Learning*, pp. 215–218, Springer, London, UK, 2014.

[24] K.-L. Du and M. N. S. Swamy, "Clustering II: Topics in clustering," in *Neural Networks and Statistical Learning*, pp. 259–297, Springer, London, UK, 2014.

[25] G. Gan, C. Ma, and J. Wu, *Data Clustering: Theory, Algorithms, and Applications*, ASA-SIAM Series on Statistics and Applied Probability, SIAM American Statistical Association, Philadelphia, Pa, USA, 2007.

[26] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.

[27] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007.

[28] C. Böhm, C. Plant, J. Shao, and Q. Yang, "Clustering by synchronization," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD-2010*, pp. 583–592, USA, July 2010.

[29] X. Chen, "A new clustering algorithm based on near neighbor influence," *Expert Systems with Applications*, vol. 42, no. 21, pp. 7746–7758, 2015.