Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 6698942, 11 pages
https://doi.org/10.1155/2021/6698942

*Research Article*

# OD Matching of Metro IC Card Data Based on Analysis Function

**Cheng Ding**,[1] **Cheng Wang**,[1] **Xinyi Wang**,[1] **Yueer Gao**,[2] **Yongxin Liao**,[1] **and Jianwei Chen**[3]

[1]*College of Computer Science and Technology, Huaqiao University, Xiamen 361021, China*
[2]*College of Architecture, Huaqiao University, Xiamen 361021, China*
[3]*Department of Mathematics and Statistics, San Diego State University, San Diego, CA 92182, USA*

Correspondence should be addressed to Cheng Ding; dingcheng@stu.hqu.edu.cn

Received 19 November 2020; Revised 6 June 2021; Accepted 10 June 2021; Published 21 June 2021

Academic Editor: Filippo Cacace



In the case of passengers taking the subway many times in a short time, missing cards in and out of the station, and staying in the subway station for a long time, the previous table join method cannot accurately set the time threshold parameters and correctly match the OD pairs of passengers. In order to solve these problems, an OD matching method based on analysis function is proposed in this paper. *LAG ()* is an analytic function in Oracle which allows you to access the row at a given offset prior to the current row without using a self-join. Metro IC card dataset stores the card swiping records of passengers entering and leaving the subway station every time. In this method, the dataset is sorted in ascending order according to the card number and card swiping time, and then, the lag function of Oracle is used to take the offset of the upper line of card ID, transaction date, transaction time, in and out sign, and station ID. Finally, the matching process is completed according to the OD conditions of card number, time, and inbound and outbound sign fields. This method does not need to set a time threshold and so as to deal with the situation where passengers stay too long in the subway station. The OD matching results on in and out IC swiping cards dataset in April and May 2019 of passengers of Xiamen Metro Line verify that analysis function method has better OD matching, missing swiping identification accuracy, and effect compared to the table join method.

## 1. Introduction

With the rapid development of economy, rail transit has become the development trend of the city. Passenger flow analysis is the basis of safe and reliable operation of urban rail transit. The prediction and accurate grasp of passenger flow characteristics and evolution law can provide decision-making basis for making scientific organization plans such as departure interval and departure frequency. The passenger flow analysis is based on the most original data. The improvement of management and service level of urban rail system depends on the comprehensive grasp, analysis, and application of metro travel data. The matching and calculation of origin destination (OD) matrix is the most important and basic step. When passengers take the subway, they swipe the card when they enter or leave the station. The two card swiping records will be stored in the same table. Every inbound swipe card record will have an outbound swipe card corresponding to it. However, in reality, due to human reasons or equipment failure, the card reading data in and out of the station cannot be completely corresponding. As shown in Figure 1, the comparison of the card reading data in Xiamen from March to August shows that there is a certain deviation between the amounts of card swiping in and out of the station each month, which are not exactly equal. In order to correctly match all the inbound information and outbound information and record them in the same line and separate the missing records of the in and out stations and store them in different tables, it is necessary to match and calculate the OD of the metro. This paper proposes an OD matching algorithm for metro IC card data based on analysis function. It completes the OD matching of metro IC card data and the identification of wrong data and completes the experimental verification with the data from April to May in Xiamen.

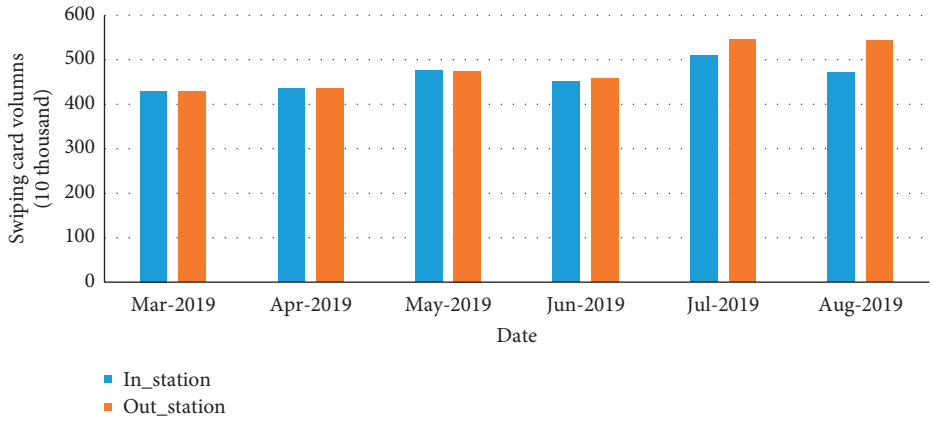


Figure 1: Comparison of IC card data of Xiamen metro from March to August in 2019.

Compared with the method based on table join, this paper proposes an OD matching algorithm based on analysis function for metro IC card data analysis.

The primary contributions of this paper can be summarized as follows:

(1) An OD matching method based on analysis function is proposed for the data mining and OD matching of metro IC card data.

(2) We analyze the differences, advantages, and disadvantages of OD matching methods based on table join and analysis function.

(3) We conduct a theoretical analysis of the characteristics of the OD matching method based on analysis function.

(4) We use the calculation results of OD matching method based on analysis function; part of data is artificially eliminated to construct dataset, and then OD matching based on table join and analysis function is used according to the constructed dataset. The effectiveness of the two algorithms is verified by comparing the analysis results.

At present, many scholars have done a lot of research on metro IC card data. IC card data records the card information of passengers. Every time a passenger swipes the card in and out of the station, a record will be saved in the system; that is to say, the data of two swiping cards in and out of the station are separated. There is important information in the record, such as card ID, time, inbound or outbound sign, and station ID. Based on the IC card data, Shin [1] carried out the metro OD matrix processing. Instead of matching each person's OD, Shin counted the OD matrix according to the inbound/outbound sign and station ID in the field, which is not specific to individuals. Shin did not consider calculating the OD match of each swipe record and was only concerned about the statistics. Chen et al. [2] analyzed the high-dimensional and multivariate metro data through the random matrix theory and then predicted the abnormal data of card swiping. Through the analysis of IC card data, Yu et al. [3, 4] classified the abnormal OD data of passengers into two categories: passenger anomaly and system anomaly. The data they used in the experiment was generated directly from the AFC system, and they did not point out the principle of OD matching of card swiping data in and out of the station. In Zhiyuan et al.'s work [5], with high visualization frameworks, the massive data of passenger flow in Shanghai metro network is highly graphical in time-space, which is processed from four aspects: the network, line, station, and section. The metro card data is also from AFC system. Li et al. [6] extracted traffic origins and destinations (OD) information of travelers from the multisource data and used the extracted data for traffic zone division. Finally, a multimode traffic forecasting model was established on this basis. Moon et al. used Seoul smart card data to measure the traffic convenience of each region by considering the actual experience of passengers in the travel network [7]. Kim et al. used Seoul's bus IC card data to introduce a sticky index to quantify the user's preference range and to measure the degree of habitual behavior of individuals and bus routes [8]. Li et al. proposed a framework for extracting potential paths from smart card data. Taking Beijing as an example, spatial clustering algorithm was used to provide a basis for customizing bus routes [9–12]. Lee et al. used the data of Seoul's bus smart card to evaluate the transfer efficiency of bus and subway transfer stations and put forward improvement strategies. The results showed that the evaluation results of DEA model for transfer efficiency are reasonable [13]. Ha and Lee used metro smart card data from Seoul to classify travel modes into working and nonworking trips to study changes in urban activities and spatial structure [14]. Zhou et al. classified and explored the functions of metro station area according to the smart card data of Wuhan metro system and passenger travel data [15]. Zhang et al. combined the smart card data of Guangzhou metro and proposed a new combination algorithm to evaluate the route choice of subway passengers [16]. Wang et al. analyzed the relationship between travel patterns and urban functional structure based on metro passenger travel data in Beijing, Shenzhen, and London [17]. Zhao et al. developed a probability model based on the smart

card data of Shenzhen metro to estimate how passenger flows are allocated to different routes and trains from empirical analysis [18]. Lin et al. [19] proposed an automated multistage method for inferring the time variable in various components of a metro network. They evaluated the proposed method for a route planning application, using smart card data from Singapore, and compared the estimated results with ground truth values.

Yang et al. [20] considered the optimization problem for timetables in subway systems. Subway passenger flow forecasting models were presented for peak-hour flow by Pan et al. [21], for special event occurrences by Ni et al. [22], and for the Beijing subway using spatiotemporal correlations by Wang and Cai [23]. Chen et al. analyzed the spatiotemporal characteristics of multimode travelers by combining the taxi FCD, the metro IC card data, and the GPS trajectories of Mobike and proposed a binomial logit model (BNL) to estimate mode choices for both peak and off-peak periods. The metro IC card data they used already contained the in and out information after OD matching [24]. Sun and Guan proposed measuring the metro network vulnerability from the perspective of line operation. Passenger flow distribution and redistribution were simulated for different disruption scenarios based on all-or-nothing assignment rule [25].

The above research is based on the metro IC card to complete the problem research, and this paper aims to achieve the complete matching of passengers' OD only from the metro IC card data, find out the abnormal data of missing card swiping in or out of the station, and construct a standard dataset to verify.

The rest of the paper is organized as follows: Section 2 overviews the metro IC card data and OD matching method. Section 3 describes the OD matching algorithm based on analysis function. In Section 4, we describe the experimental dataset construction and present qualitative and quantitative results.

## 2. Data and Methods

The symbols used are defined as follows:

$N_{ic}$ represents the amount of IC card data in a dataset

$N_{od}$ represents the number of OD pairs in a dataset

$N_{err}$ represents the number of records of wrong connection caused by multiple subway ride records of a passenger in 5,400 seconds

$N_{et}$ represents the number of OD pairs whose time difference exceeds the threshold of 5,400 seconds

$k_1$ represents the ratio of the number of records of wrong connection caused by multiple subway ride records of a certain passenger within 5,400 seconds

$k_2$ represents the proportion of the records of passengers who have not successfully matched OD due to a subway trip time of more than 5,400 seconds

$N_{in}$ represents the amount of inbound IC card data in a dataset

$N_{out}$ represents the amount of outbound IC card data in a dataset

$N_{mi}$ represents the amount of outbound card swiping corresponding to the missing inbound card in a dataset

$N_{mo}$ represents the amount of inbound card swiping corresponding to the missing outbound card in a dataset

*2.1. Introduction to the Composition of the Data Dictionary.* Taking Xiamen metro IC card data as an example, *TICKET_ID* stands for transaction card number, *TXN_DATE* stands for transaction date, *TXN_TIME* stands for trading time, *TICKET_MAIN_TYPE* stands for card type, *TRANS_CODE* represents the type of transaction (boarding and alighting), in which 7 represents inbound and 8 represents outbound, and *TXN_ STATION_ ID* represents the current card swiping subway station ID. The specific data are shown in Table 1.

*2.2. Formal Description of Metro OD Matching.* A passenger's complete subway journey is recorded twice with a card: one for entering the station and the other for leaving the station. According to the records in Table 1, this paper studies how to correctly match the OD pairs when two card swiping records exist in the same table and there are two different records, which are recorded as $N_{od}$; Table 2 shows 145 information of OD pairs; and we find out the outbound data of missing corresponding inbound card swiping and the inbound data of missing corresponding outbound card swiping, which are recorded as $N_{mi}$ and $N_{mo}$ and stored in Tables 3 and 4 respectively.

*2.3. Difficulties of the Problem.* There are three cases in the metro IC card record:

(1) Passengers swipe cards normally when entering and leaving the station

(2) Passengers swipe the card when they enter the station but do not swipe the card when exiting the station or the data is lost

(3) Passengers swipe the card when they leave the station but do not swipe the card when entering the station or the data is lost

Cases 2 and 3 are invalid data, which cannot form a complete passenger boarding and alighting record and need

Table 1: Metro IC card data.

| TICKET_ID | TXN_DATE | TXN_TIME | TICKET_MAIN_TYPE | TRANS_CODE | TXN_STATION_ID |
|---|---|---|---|---|---|
| 000****001 | 20190427 | 51688 | 20 | 7 | 108 |
| 000****001 | 20190427 | 51964 | 20 | 8 | 112 |
| 001****450 | 20190427 | 52803 | 20 | 7 | 109 |
| 001****450 | 20190427 | 54023 | 20 | 8 | 103 |
| 000****020 | 20190427 | 65244 | 20 | 7 | 105 |

Table 2: Card data of metro entry and exit station matching obtained by problem solving.

| TICKET_ID | STATION_ID_IN | TRANS_DATE | ENTERTIME | STATION_OUT | EXITTIME | MONEY |
|---|---|---|---|---|---|---|
| 101811200****9F9 | 106 | 20190916 | 07 : 35 : 06 | 108 | 07 : 44 : 08 | 2.00 |
| 101811200****029 | 112 | 20190916 | 07 : 36 : 16 | 101 | 07 : 58 : 31 | 4.00 |
| 101811200****694 | 114 | 20190916 | 08 : 45 : 44 | 102 | 09 : 04 : 39 | 5.00 |
| 101811200****B44 | 105 | 20190916 | 08 : 34 : 36 | 109 | 08 : 45 : 28 | 2.00 |

Table 3: Only outbound card swiping and no corresponding inbound card swiping data sheet.

| TICKET_ID | TRANS_DATE | JYSJZ | TRANS_CODE | STATION_ID |
|---|---|---|---|---|
| 801231904****914 | 20190916 | 07 : 25 : 02 | 8 | 114 |
| 801231905****336 | 20190916 | 07 : 26 : 17 | 8 | 101 |
| 801231905****914 | 20190916 | 07 : 45 : 34 | 8 | 110 |
| 801765100****636 | 20190916 | 08 : 32 : 11 | 8 | 109 |

Table 4: Only inbound card swiping and no corresponding outbound card swiping data sheet.

| TICKET_ID | TRANS_DATE | JYSJZ | TRANS_CODE | STATION_ID |
|---|---|---|---|---|
| 801231904****445 | 20190916 | 07 : 25 : 02 | 7 | 109 |
| 801231904****219 | 20190916 | 07 : 26 : 17 | 7 | 112 |
| 801231905****645 | 20190916 | 07 : 45 : 34 | 7 | 106 |
| 801231905****098 | 20190916 | 08 : 32 : 11 | 7 | 105 |

to be eliminated in the process of OD matching algorithm. However, only the data in case 1 can be matched successfully. Of course, there are few cases in cases 2 and 3, and most of the subway data belong to the scope of case 1.

*2.4. Introduction of Metro OD Matching Method Based on Table Join.* The method of table join is to connect the data of card swiping in and out according to certain conditions to form a complete record of passengers getting on and off the train. Its flow chart is shown in Figure 2, and the steps are as follows:

(1) Data preparation. Select the fields of card number, transaction date, transaction time, station in and out sign, and station ID from the original IC card data, and change the transaction time into seconds, such as 7 : 35 a.m. to 44,100 seconds.

(2) Table join. According to the field *TRANS_CODE*, the original table is divided into inbound table *A* and outbound table *B*, with *TRANS_CODE* = 7 in table *A* and 8 in table *B*. The records of *A* and *B* are connected according to the following conditions:

(a) *TICKET_ID* of table *A* = *TICKET_ID* of table *B*
(b) *TXN_DATE* of table *A* = *TXN_DATE* of table *B*
(c) *TXN_TIME* of table *A* < *TXN_TIME* of table *B*
(d) *TXN_TIME* of table *B* − *TXN_TIME* of table *B* < 5,400; the difference between the outbound trading time and the inbound trading time is less than or equal to 5,400 seconds, that is, one and a half hours

*2.5. Existing Problems.* Based on the table join method, 5,400 seconds is taken as the threshold value, and the whole process (swiping card at the station, waiting for the train, getting on the train, arriving at the destination, and swiping the card at the exit) does not exceed 5,400 seconds (1.5 hours):

(1) There is no actual basis for selecting 5,400 seconds. In reality, there is a record of subway travel time

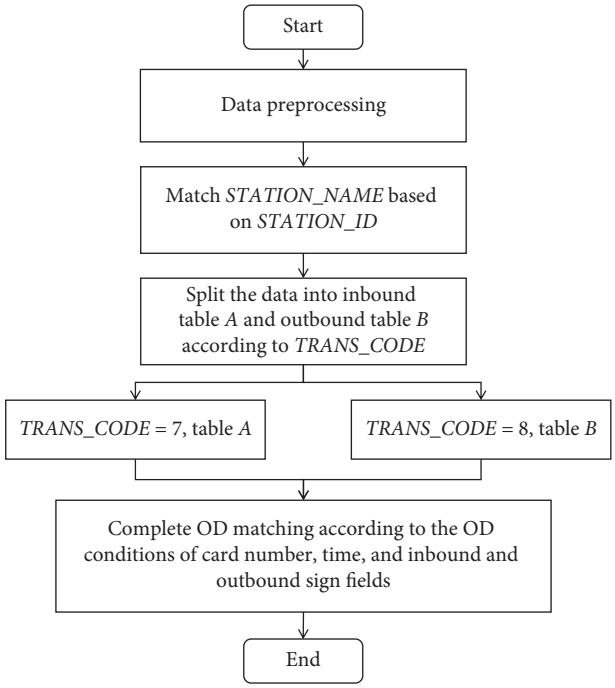


Figure 2: OD matching flow chart of metro IC card data based on table join.

exceeding 5,400 seconds, which leads to data loss when exceeding the threshold value

(2) When there are multiple subway trips in 5,400 seconds, there will be more wrong connections

## 3. Introduction to the Metro OD Matching Method Based on Analysis Function

*3.1. Introduction to Oracle LAG () Function.* Oracle *LAG ()* [26] is an analytic function that allows you to access the row at a given offset prior to the current row without using a self-join:

```
LAG (expression [, offset] [, default])
OVER (
  [query_partition_clause]
  order_by clause
)
```

The usage of the lag function is shown above. *expression* is a scalar expression evaluated against the value of the row at a given offset prior to the current row. *offset* is the number of rows that you want to backward from the current row. The default is 1. For *default*, if the offset goes beyond the scope of the partition, the function returns the default. If you omit default, then the function returns *NULL*. The *query_partition_clause* divides rows into partitions to which the *LAG ()* function is applied. By default, the function treats the whole result set as a single partition. The *order_by_clause* specifies the order of the rows in each partition to which the *LAG ()* function is applied. Similar to the *LEAD ()* function, the *LAG ()* function is very useful for calculating the difference between the values of current and previous rows.

*3.2. Introduction to the Metro OD Matching Method Based on Analysis Function.* The metro OD matching algorithm based on the analysis function mainly uses the lag function of Oracle to replace the table join. In this method, the dataset is sorted in ascending order according to the card number and card swiping time, and then the lag function of Oracle is used to take the offset of the upper line of five related fields. Finally, the matching process is completed according to the OD conditions of card number, time, and inbound and outbound sign fields. Its flow chart is shown in Figure 3, and the steps are as follows:

(1) Select the fields of card number, transaction date, transaction time, inbound and outbound flag, and station ID from the original data table, and change the transaction time into seconds.

(2) For *TICKET_ID*, *TXN_DATE*, *TXN_TIME*, *TRANS_CODE*, and *TXN_STATION_ID*, these five fields are grouped by ID, according to *TXN_DATE*, *TXN_TIME* for sorting to get the *LAG ()* function and set the offset to 1. The field names are *PRE_TICKET_ID*, *PRE_TXN_DATE*, *PRE_TXN_TIEM*, *PRE_TRANS_CODE*, and *PRE_TXN_STATION_ID*.

(3) The correct OD pairs were selected according to the four following conditions:

(a) $TICKET\_ID = PRE\_TICKET\_ID$.
(b) $TXN\_DATE = PRE\_TXN\_DATE$.
(c) $TRANS\_CODE = 8$ *AND* $PRE\_TRANS\_CODE = 7$.
(d) $TXN\_TIME > PRE\_TXN\_TIME$.

Taking the metro IC card data in May 2019 as an example, the OD matching SQL statement based on analysis function is as follows:

The following SQL statement uses the *LAG ()* function to get the offsets of the five fields for the metro IC card data in May 2019. "metro201905" is a table of metro IC card data in April 2019, with the structure shown in Table 1:

```
CREATE TABLE metro201905_t1 AS
SELECT t. TICKET_ID,
  t. TXN_DATE,
  t. TXN_TIME,
  t. TRANS_CODE,
  t. STATION_ID,
  LAG (t. TICKET_ID, 1, NULL) over (partition by t. TICKET_ID
  ORDER by t. TICKET_ID, t. TXN_DATE, t. TXN_TIME) AS PRE_TICKET_ID,
  LAG (t. TXN_DATE, 1, NULL) over (partition by t. TICKET_ID
```

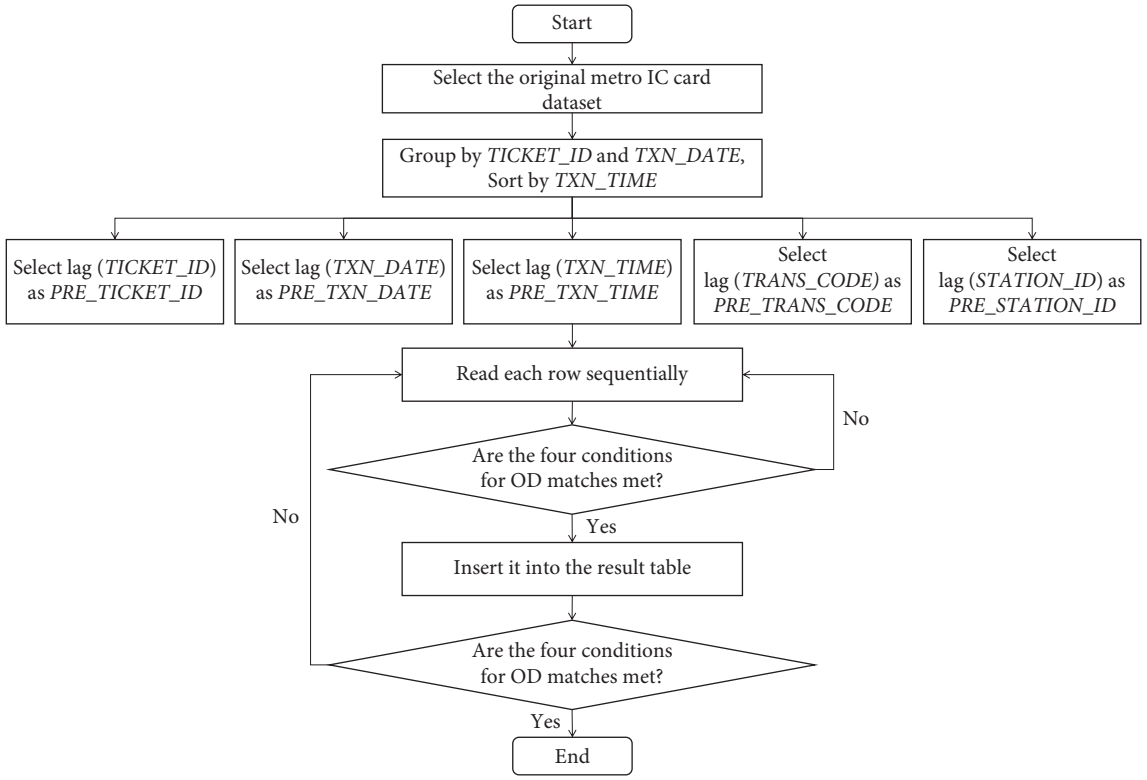


Figure 3: Metro OD matching process based on analysis function.

*ORDER by t. TICKET_ID, t. TXN_DATE, t. TXN_TIME) AS PRE_TXN_DATE,*

*LAG (t. TXN_TIME, 1, NULL) over (partition by t. TICKET_ID*

*ORDER by t. TICKET_ID, t. TXN_DATE, t. TXN_TIME) AS PRE_TXN_TIME,*

*LAG (t. TRANS_CODE, 1, NULL) over (partition by t. TICKET_ID*

*ORDER by t. TICKET_ID, t. TXN_DATE, t. TXN_TIME) AS PRE_TRANS_CODE,*

*LAG (t. STATION_ID, 1, NULL) over (partition by t. TICKET_ID*

*ORDER by t. TICKET_ID, t. TXN_DATE, t. TXN_TIME) AS PRE_STATION_ID*

*FROM metro. metro201905 t*;

The following SQL statement fetches and stores the data that meets the four conditions of the OD pair to table "metro201905_final," with the structure shown in Table 2:

*CREATE TABLE metro201905_final AS*

*SELECT t. TICKET_ID,*

*t. TXN_DATE,*

*t. TXN_TIME AS EXITTIME,*

*t. STATION_ID AS STATION_ID_out,*

*t. PRE_TXN_TIME AS ENTERTIME,*

*t. PRE_STATION_ID AS STATION_ID_IN*

*FROM metro201905_t1 t*

*WHERE t. TICKET_ID = t. PRE_TICKET_ID*

*AND t. TXN_DATE = t. PRE_TXN_DATE*

*AND t. TRANS_CODE = 8*

*AND t. PRE_TRANS_CODE = 7*

*AND t. TXN_TIME > t. PRE_TXN_TIME.*

*3.3. Theoretical Analysis and Comparison of Two Methods.* Compared with the table join mode of setting time threshold, the analysis function method proposed in this paper avoids the way of setting connection conditions and thresholds for table join and can accurately distinguish the three situations of metro IC card swiping data and will not cause the records exceeding the threshold to be discarded due to the influence of threshold size. Table 5 shows theoretical comparison of different methods for orbit OD matching.

## 4. Experiment

*4.1. Experimental Dataset Construction*

*4.1.1. Reasons for Dataset Construction.* In the OD matching and calculation of the metro, only the records of entering and leaving the station can be matched, but the missing records in and out of the station cannot be matched. However, the situation of missing entering and exiting

Table 5: Theoretical comparison of different methods for metro OD matching.

| Method | Is accuracy affected by the selected threshold? | Does the method need to set corresponding thresholds for different datasets? | Can missing records classification be realized |
|---|---|---|---|
| Based on table join | Yes | Yes | No |
| Based on analysis function | No | No | Yes |

station records in the original metro data sheet is recorded, and the specific missing data is not known:

(1) It is impossible to know the data that can match OD completely

(2) The OD results obtained by different matching methods cannot prove the correctness of the matching

Therefore, we hope to have a real dataset that knows the specific missing data situation (including the amount of missing data and the number of missing pieces) and then compare the OD results obtained by different methods with the OD results of the real dataset to prove its correctness.

*4.1.2. Dataset Construction Process.* The flow chart for building the dataset is shown in Figure 4 and dataset construction process is mainly divided into the eight following steps:

Step 1: select the card number, transaction date, arrival time, card type, and inbound station ID fields from the OD dataset of complete matching metro. The inbound time field is named as the transaction date, and the inbound station ID field is named as the station ID, and these data are stored in the inbound information table;

Step 2: add the inbound and outbound flag field in the inbound information table, and update the field of all data in this table to inbound station;

Step 3: select the card number, transaction date, outbound time, card type, and outbound station ID fields from the complete matching OD dataset. The outbound time field is named as the transaction date, and the outbound station ID field is named as the station ID, and these data are stored in the outbound information table;

Step 4: add the inbound and outbound flag field in the outbound information table, and update all data in this table to outbound;

Step 5: merge the inbound and outbound information tables into a single inbound/outbound table, and arrange the data in the table randomly once;

Step 6: randomly extract a part of inbound data according to proportion;

Step 7: in the remaining data after deducting the extracted inbound data, a part of outbound data with different card numbers from the incoming data just extracted is randomly extracted;

Step 8: after deducting the extracted outbound data, the remaining data is the constructed dataset.

*4.2. Introduction to Instance Objects and Datasets.*

Research object: the IC card data of rail transit in April and May 2019 in Xiamen City, Fujian Province, are selected. Data details are shown in Table 6.

Dataset: select the metro IC card data, static station information data, and static card type data in April and May 2019.

*4.3. Evaluation Methods and Indicators*

*4.3.1. Evaluation Method.*

(1) Using the original IC card data, the OD results obtained by the analysis function method are compared with the OD results based on the table join method.

(2) First of all, a correct metro IC card dataset is constructed; that is, each passenger's inbound card has a corresponding outbound card swipe, and there is no error or omission in the data. Then the OD matching of the dataset is carried out by using the method based on table join and the method based on analysis function. Finally, the results are verified with the real OD results.

(3) Firstly, a missing metro IC card swiping dataset is constructed; that is, the passengers have a lack of entry or exit records in a subway trip, and then the OD matching of the dataset is carried out by using the method based on table join and the method based on analysis function. Finally, the results are verified with the real OD results.

*4.3.2. Evaluating Indicator.*

(1) The accuracy of matching;

(2) The rate of wrong connection caused by the same passenger taking the subway for several times within 5,400 seconds and the ratio of not connecting for more than 5,400 seconds are counted as evaluation indexes. The specific calculation formula is as follows:

$$k_1 = \frac{N_{\text{err}}}{N_{\text{od}}}, \qquad k_2 = \frac{N_{\text{et}}}{N_{\text{od}}}. \tag{1}$$

In the above equation, $k_1$ represents the ratio of the number of records of wrong connection caused by multiple

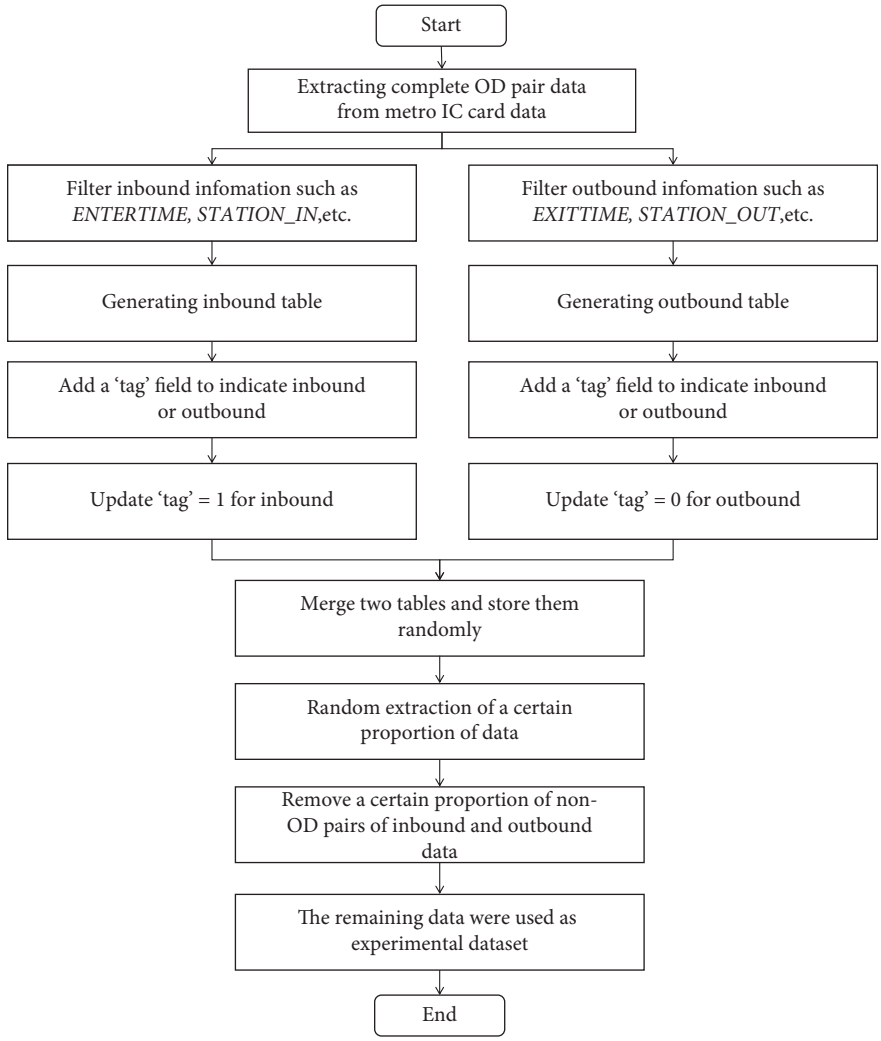


Figure 4: The dataset construction process of complete metro OD data eliminating some inbound and outbound information.

Table 6: Data volume of all datasets used.

| Dataset | Time | Total |
|---|---|---|
| IC card data of metro in April | April 2019 | 8709743 |
| IC card data of metro in May | May 2019 | 9494211 |
| Static site information data | — | 24 |
| Card main type data | — | 10 |

Note: the main types of cards include one-way ticket, stored value ticket, period ticket, commemorative ticket, employee ticket, e-card, financial IC card, transportation card, and e-ticket.

subway ride records of a certain passenger within 5,400 seconds; $N_{err}$ represents the number of records of wrong connection caused by multiple subway ride records of a passenger in 5,400 seconds; $N_{od}$ represents the amount of passenger flow after OD; $k_2$ represents the proportion of the records of passengers who have not successfully matched OD due to a subway trip time of more than 5,400 seconds; $N_{et}$ represents the number of OD pairs whose time difference exceeds the threshold of 5,400 seconds.

*4.4. Experimental Results.* The OD matching results of metro IC card data in April and May 2019 based on table join and analysis function are shown in Tables 7 and 8.

According to the OD data calculated by the analysis function method and the dataset construction process of Section 4.1.2, the datasets of April and May 2019 are constructed, respectively, and the OD matching based on table join and that based on analysis function method are compared again. Table 9 shows the OD matching results of the original IC card data in April and May 2019 through the analysis function. The results of dataset construction are shown in Table 10. Table 11 shows the experimental results of the datasets constructed with the original IC card data in

Table 7: OD matching results of two methods based on April 2019 data.

| Method | $N_{ic}$ | $N_{od}$ | $N_{err}$ | $N_{et}$ | $k_1$ | $k_2$ |
|---|---|---|---|---|---|---|
| Based on table join | 8709743 | 4409973 | 92050 | 10486 | 2.09% | 0.12% |
| Based on analysis function | 8709743 | 4409973 | 0 | 0 | 0 | 0 |

Table 8: OD matching results of two methods based on May 2019 data.

| Method | $N_{ic}$ | $N_{od}$ | $N_{err}$ | $N_{et}$ | $k_1$ | $k_2$ |
|---|---|---|---|---|---|---|
| Based on table join | 94942111 | 4808486 | 103967 | 11408 | 2.16% | 0.12% |
| Based on analysis function | 94942111 | 4715927 | 0 | 0 | 0 | 0 |

Table 9: Data analysis after OD matching and calculation of original IC card data.

| Date | $N_{ic}$ | $N_{od}$ | $N_{mi}$ | $N_{mo}$ | $k_1$ | $k_2$ |
|---|---|---|---|---|---|---|
| April 2019 | 8,709,743 | 4,328,409 | 28,367 | 24,558 | 2.16% | 0.12% |
| May 2019 | 9,494,211 | 4,715,927 | 33,542 | 28,815 | 0 | 0 |

Table 10: The OD data structure of complete metro is lack of some information of entrance and exit stations.

| Date | $N_{ic}$ of dataset | $N_{mi}$ of dataset | $N_{mo}$ of Dataset | $N_{in}$ of dataset | $N_{out}$ of dataset | $N_{od}$ of dataset |
|---|---|---|---|---|---|---|
| April 2019 | 8,656,818 | 28,567 | 24,238 | 4,299,842 | 4,304,171 | 4,275,604 |
| May 2019 | 9,431,854 | 33,011 | 28,295 | 4,682,916 | 4,687,632 | 4,654,621 |

Table 11: OD matching based on table join and analysis function using constructed dataset.

| Date | Method | $N_{od}$ | $N_{err}$ | $N_{et}$ | $k_1$ | $k_2$ | $N_{mi}$ | $N_{mo}$ |
|---|---|---|---|---|---|---|---|---|
| April 2019 | Based on table join | 4,444,086 | 178,815 | 10,333 | 4.02% | 0.12% | 0 | 0 |
| | Based on analysis function | 4,275,604 | 0 | 0 | 0 | 0 | 28,567 | 24,238 |
| | Real results in Table 10 | 4,275,604 | — | — | — | — | 28,567 | 24,238 |
| May 2019 | Based on table join | 4,753,677 | 79,714 | 11,310 | 1.68% | 0.12% | 0 | 0 |
| | Based on analysis function | 4,654,621 | 0 | 0 | 0 | 0 | 33,011 | 28,295 |
| | Real results in Table 10 | 4,654,621 | — | — | — | — | 33,011 | 28,295 |

April and May 2019 by using the methods of table join and analysis function.

*4.5. Analysis of Experimental Results.* By comparing the relationship between the OD passenger flow and the total passenger flow, we can see that the relationship between the table join method and the total passenger flow is as follows:

(1) Based on table join method, OD passenger flow volume × 2 > total passenger flow volume.

(2) Therefore, it can be considered that there are redundant matches in the metro OD results obtained by the intratable join method, which is unreasonable to some extent.

(3) By comparing the metro OD obtained by the two methods, it can be seen that the method based on the analysis function has good performance in the aspects of multiple trips in 5,400 seconds and is unconnected in more than 5,400 seconds.

(4) By matching the constructed datasets, the OD matching results obtained by the proposed analysis function method are the same as the real data, but the table join method is not the same as the real data.

(5) Using the constructed dataset, the missing inbound and outbound data stored in the missing data table of inbound and outbound stations obtained by the analysis function method are combined with the extracted inbound and outbound data to perform OD matching, and the matching result is the same as the real data.

Therefore, it can be considered that the analysis function method is better than the table join method in OD matching of metro.

## 5. Conclusion and Future Work

This paper proposes an OD matching algorithm based on analysis function for metro IC card data. Compared with the

previous table join OD matching algorithm, the method based on analysis function avoids the setting of time threshold, so as to deal with the situation where passengers stay too long in the subway station. Taking Xiamen metro in and out IC swiping card dataset in April and May 2019 as an example, this method is more accurate and powerful than table join method and can identify the correct OD and the wrong or real IC swiping card records in and out of the station.

In the future work, this method should be verified at IC card inbound and outbound dataset of passengers of complex metro network with transfer stations. Time and space complexity of this method should be analyzed and optimized so that it has excellent accuracy and also less time and space cost and could be applied to large-scale dataset.

## Data Availability

The GPS data used to support the findings of this study were supplied by Xiamen GNSS Development and Application Co., Ltd. under license and so cannot be made freely available.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

This work was supported by China National Social Science Fund (19BXW110).

## References

[1] H. Shin, "Analysis of subway passenger flow for a smarter city: knowledge extraction from Seoul metro's "untraceable" big data," *IEEE Access*, vol. 8, pp. 69296–69310, 2020.

[2] X. Chen, C. Yang, X. Xu, and Y. Gong, "Anomaly detection in metro passenger flow based on random matrix theory," in *Proceedings of the 2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pp. 625–630, Auckland, New Zealand, October 2019.

[3] W. Yu, H. Bai, J. Chen, and X. Yan, "Anomaly detection of passenger OD on nanjing metro based on smart card big data," *IEEE Access*, vol. 7, pp. 138624–138636, 2019.

[4] X. Liu, Z. Hu, H. Ling, and Y.-M. Cheung, "MTFH: a matrix tri-factorization hashing framework for efficient cross-modal retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 964–981, 2021.

[5] H. Zhiyuan, Z. Liang, X. Ruihua, and Z. Feng, "Application of big data visualization in passenger flow analysis of Shanghai metro network," in *Proceedings of the 2017 2nd IEEE International Conference on Intelligent Transportation Engineering (ICITE)*, pp. 184–188, Singapore, September 2017.

[6] D. Li, Y. Tang, and Q. Chen, "Multi-mode traffic demand analysis based on multi-source transportation data," *IEEE Access*, vol. 8, pp. 65005–65019, 2020.

[7] H. Moon, K. Oh, S. Kim, and J.-Y. Jung, "Analysis of regional transit convenience in Seoul public transportation networks using smart card big data," *Journal of Korean Institute of Industrial Engineers*, vol. 42, no. 4, pp. 296–303, 2016.

[8] J. Kim, J. Corcoran, and M. Papamanolis, "Route choice stickiness of public transport passengers: measuring habitual bus ridership behaviour using smart card data," *Transportation Research Part C: Emerging Technologies*, vol. 83, pp. 146–164, 2017.

[9] J. Li, Y. Lv, J. Ma, and Q. Ouyang, "Methodology for extracting potential customized bus routes based on bus smart card data," *Energies*, vol. 11, no. 9, pp. 2224–2228, 2018.

[10] Y. Chen, L. Zhou, L. Sheng et al., "KNN-block dbscan: fast clustering for large-scale data," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 6, pp. 3939–3953, 2021.

[11] Y. Chen, L. Zhou, N. Bouguila et al., "Block-dbscan: fast clustering for large scale data," *Pattern Recognition*, vol. 109, Article ID 107624, 2021.

[12] Y. Chen, X. Hu, and W. Fan, "Fast density peak clustering for large scale data based on kNN," *Knowledge-based System*, vol. 187, Article ID 104824, 2020.

[13] E. H. Lee, H. Lee, S.-Y. Kho, and D.-K. Kim, "Evaluation of transfer efficiency between bus and subway based on data envelopment analysis using smart card data," *KSCE Journal of Civil Engineering*, vol. 23, no. 2, pp. 788–799, 2019.

[14] J. Ha and S. Lee, "A study on the monitoring of urban activity and spatial structure changes using public transportation big data-Based on the smart card data and the prestige centrality index in Seoul, Republic of Korea," *Journal of Korea Planning Association*, vol. 52, no. 6, pp. 73–90, 2017.

[15] Y. Zhou, Z. Fang, Q. Zhan, Y. Huang, and X. Fu, "Inferring social functions available in the metro station area from passengers' staying activities in smart card data," *ISPRS International Journal of Geo-Information*, vol. 6, no. 12, pp. 394–396, 2017.

[16] Y. Zhang, E. Yao, J. Zhang, and K. Zheng, "Estimating metro passengers' path choices by combining self-reported revealed preference and smart card data," *Transportation Research Part C: Emerging Technologies*, vol. 92, pp. 76–89, 2018.

[17] Z. Wang, Y. Hu, P. Zhu, Y. Qin, and L. Jia, "Ring aggregation pattern of metro passenger trips: a study using smart card data," *Physica A: Statistical Mechanics and Its Applications*, vol. 491, pp. 471–479, 2018.

[18] J. Zhao, F. Zhang, L. Tu et al., "Estimation of passenger route choice pattern using smart card data for complex metro systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 4, pp. 790–801, 2017.

[19] X. Lin, X. Xiao, and Z. Li, "A scalable approach to inferring travel time in Singapore's metro network using smart card data," in *Proceedings of the 2018 IEEE International Smart Cities Conference (ISC2)*, pp. 1–8, Kansas City, MO, USA, September 2018.

[20] X. Yang, B. Ning, X. Li, and T. Tang, "A two-objective timetable optimization model in subway systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 5, pp. 1913–1921, 2014.

[21] P. Pan, H. Wang, L. Li, Y. Wang, and Y. Jin, "Peak-hour subway passenger flow forecasting: a tensor based approach," in *Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pp. 3730–3735, Maui, HI, USA, November 2018.

[22] M. Ni, Q. He, and J. Gao, "Forecasting the subway passenger flow under event occurrences with social media," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 6, pp. 1623–1632, 2017.

[23] Z. Wang and X. Cai, "Research on passenger flow prediction of Beijing subway based on spatiotemporal correlation

analysis," in *Proceedings of the 2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, pp. 279–283, Chengdu, China, April 2019.

[24] F. Chen, Z. Yin, Y. Ye et al., "Taxi hailing choice behavior and economic benefit analysis of emission reduction based on multi-mode travel big data," *Transport Policy*, vol. 97, 2020.

[25] D. Sun and S. Guan, "Measuring vulnerability of urban metro network from line operation perspective," *Transportation Research Part A: Policy and Practice*, vol. 94, 2016.

[26] Oracle, "Introduction to oracle *LAG ()* function," 2020, https://www.oracletutorial.com/oracle-analytic-functions/oracle-lag/.