Spatio-Temporal Analysis of Passenger Travel Patterns in Massive Smart Card Data

Juanjuan Zhao, Qiang Qu, Fan Zhang, Chengzhong Xu, *Fellow, IEEE*, and Siyuan Liu

***Abstract*— Metro systems have become one of the most important public transit services in cities. It is important to understand individual metro passengers' spatio-temporal travel patterns. More specifically, for a specific passenger: what are the temporal patterns? what are the spatial patterns? is there any relationship between the temporal and spatial patterns? are the passenger's travel patterns normal or special? Answering all these questions can help to improve metro services, such as evacuation policy making and marketing. Given a set of massive smart card data over a long period, how to effectively and systematically identify and understand the travel patterns of individual passengers in terms of space and time is a very challenging task. This paper proposes an effective data-mining procedure to better understand the travel patterns of individual metro passengers in Shenzhen, a modern and big city in China. First, we investigate the travel patterns in individual level and devise the method to retrieve them based on raw smart card transaction data, then use statistical-based and unsupervised clustering-based methods, to understand the hidden regularities and anomalies of the travel patterns. From a statistical-based point of view, we look into the passenger travel distribution patterns and find out the abnormal passengers based on the empirical knowledge. From unsupervised clustering point of view, we classify passengers in terms of the similarity of their travel patterns. To interpret the group behaviors, we also employ the bus transaction data. Moreover, the abnormal passengers are detected based on the clustering results. At last, we provide case studies and findings to demonstrate the effectiveness of the proposed scheme.**

***Index Terms*— Passenger behavior analysis, spatio-temporal analysis, smart card data, metro system.**

I. Introduction

WE ARE witnessing the rapid development of modern cities, in which transportation is one of the most important services. The increasing demands of efficient inner-city transportation call for the availability of metro systems in numerous cities. These metro systems have been becoming important and preferred public tools. To improve the metro services, such as schedule planning, evacuation policy making, and marketing, it is necessary to understand the spatio-temporal travel patterns of individual passengers. Here are several real-world application examples by our study, as follows.

- *Market Evaluation:* Better understanding of the travel patterns of individual passengers can enable transit authorities to evaluate their current services or help to adjust their marketing strategies to encourage higher usage [1]. For example, passengers as customers often quite concern new marketing policies, which can be reflected by their travel behaviors. Considering a company is to sell new monthly passes, to mine passenger travel patterns, showing their behaviors, can help to estimate both traffic and income variation before and after the release of the new passes.
- *Anomaly Detection:* The abuse of metro systems is an intractable problem. For example, after passing the toll-gates, home-less people can travel back and forth between lines for begging by simply changing their clothes to avoid suspicion. Another example, thefts often happen in public transit systems, because passengers are inclined to pay less attention to their belongings when they are in a crowded environment [2]. Based on passenger travel pattern analysis, we can spot these abnormal behaviors in terms of space and time, which can help metro companies detect those special passengers.
- *Social Networking:* The "familiar stranger" is an interesting phenomena in modern cities [3]. Social demands thus grow exponentially. Many developers based on the application needs are developing social apps connecting passengers with similar public mobility patterns, which would also benefit from this study.

Manuscript received August 5, 2016; revised January 9, 2017; accepted February 27, 2017. Date of publication April 4, 2017; date of current version October 30, 2017. This work was supported in part by the China National Basic Research Program (973 Program) under Grant 2015CB352400, in part by the National Natural Science Foundation of China under Grant U1401258 and Grant 61572488, in part by the Science and Technology Planning Project of Guangdong Province under Grant 2015B010129011, in part by the Basic Research Program of Shenzhen under Grant JCYJ20150630114942313 and Grant JCYJ20140610152828686, in part by the Natural Science Foundation of Guangdong Province under Grant 2015A030310326 and Grant 2016A030313183, in part by NSF under Grant CMMI 1436786 and Grant CNS 1526638, in part by the Research Program of Shenzhen under Grant JSGG20150512145714248, Grant KQCX2015040111035011, and Grant CYZZ20150403111012661, and in part by the Fundamental Research Funds for the Central Universities. This paper was presented at the IEEE 17th International Conference on Intelligent Transportation Systems, 2014, Qingdao, China, entitled Understanding Temporal and Spatial Travel Patterns of Individual Passengers by Mining Smart Card Data. The current manuscript extends the work from method design, literature review, to more extensive experimental study and more comprehensive discussion of findings. The Associate Editor for this paper was F. Chu. *(Corresponding author: Qiang Qu.)*

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

Q. Qu and F. Zhang are with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: qiang@siat.ac.cn; fan.zhang@siat.ac.cn).

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

S. Liu is with the Smeal College of Business, Pennsylvania State University, State College, PA 16801 USA (e-mail: siyuan@psu.edu).

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.2017.2679179

1524-9050 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY ROORKEE. Downloaded on December 13,2024 at 11:37:43 UTC from IEEE Xplore. Restrictions apply.

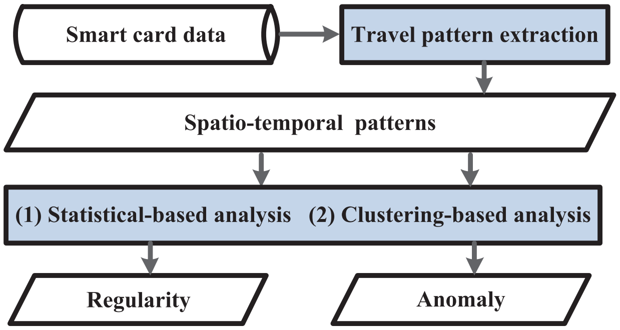

Smart card data
Travel pattern extraction
Spatio-temporal patterns
(1) Statistical-based analysis
(2) Clustering-based analysis
Regularity
Anomaly

Fig. 1. Overview.

The increasing availability of individual travel history data collected by automated fare collection (AFC) systems, where passengers tap-in or tap-out through swapping smart card, brings us challenges as well as opportunities to understand individual's travel behavior from the raw data. On this topic, some techniques have been proposed base on AFC data, including predicting the movement destination for an individual passenger [4], extracting the transit usage and access distance of a passenger [5], recognizing individual passengers' travel regularity based on the similarity of their own trips [6], etc. However, we study individual¡¯s general travel style and regularity travel patterns in terms of space and time. Based on three aspects, temporal, spatial and spatio-temporal, we define individual's travel patterns and we propose methods to retrieve the patterns. Then, for anomaly detection and regularity discovery, we analyze individual passengers' travel patterns by statistical-based and unsupervised cluster-based methods. Some interesting findings are shown in the paper. For example, we cluster metro passengers into four groups in terms of the similarity of their travel patterns. We find a group of passengers who usually take metro in one trip and return by bus. It is of importance to understand passengers¡¯ choice on transportation tools. The results may be used to improve the metro services, such as to help metro companies attract more passengers by adjusting marketing policies based on how people choose between metros and buses. Next we find outlier passengers with travel patterns different from the majority, which are useful for metro companies, e.g., to design tickets. The overview of our methods is shown in Fig. 1. First we extract each passenger's travel patterns based on their history trips. Then by statistical-based and unsupervised clustering-based methods, we able to find the hidden regularity and anomaly of the travel patterns. The contribution of this paper lies in following aspects:

- We study individual travel patterns in terms of space and time and we propose methods to retrieve the patterns from the raw smart card transactions.
- We investigate passengers travel patterns for anomaly detection and regularity discovery. We analyze individual passengers' travel patterns by statistical-based and unsupervised cluster-based methods. We also show interesting findings of passenger behaviors buried in the datasets.
- We further consider bus transaction records for metro passenger behavior analysis, which is very useful for better understanding passengers' choices on transportation tools.

The rest of this paper is organized as follows. Section II discusses the related work and dataset used in this study. Section III presents our methods to extract spatio-temporal travel patterns of individual passengers. The statistical-based and unsupervised clustering-based analyses are discussed in Section IV and Section V. Section VI concludes the paper with potential outline for future studies.

II. Background

A. Related Work

In this section, we briefly review the literature from 3 aspects: 1) studies on smart card data, 2) behavior and pattern analysis, and 3) anomaly detection, respectively, in transportation research.

1) Smart Cards: Nowadays, the data generated in cities can help to tackle some issues in big cities, such as traffic congestion, pollution and so on [7]. Smart cards or city passes have been widely used in urban public transportation systems. The smart card data are studied in an increasing number of applications, including demand analysis, scheduling and so on [8]. We survey the representative studies as follows. Munizaga and Palma [9] presented a method to estimate a public transport OD matrix. Ma *et al.* [10] proposed a Markov chain based Bayesian decision tree algorithm to infer passengers boarding stops. Trépanier *et al.* [4] predicted get-off stations of bus passengers with smart card data. Morency *et al.* [11] analyzed transit riders' travel variability, they pointed out that better understanding of travel variability can help to reduce operational costs. Utsunomiya *et al.* [5] analyzed the frequency and the consistency of daily travel patterns of passengers using smart card transactions. Bagchi and White [12] used smart card data to infer turnover rates, trip rates, and the proportion of linked trips. Chu *et al.* [13] reconstructed bus passengers' itinerary, and spatial-temporal portraits on road networks. Agard *et al.* [14] defined user types based on their temporal regularity and daily patterns, and found that travel patterns are varied by card types. Ma *et al.* [6] analyzed individual passengers' travel behaviors and regularity based on the similarity between trips. Ma and Wang [15] developed a data-driven platform for online transit performance monitoring based on smart card and GPS data. Sun *et al.* [3] used public transit transaction records to find physical encounters who are with reproducible temporal patterns. Sun *et al.* estimated the spatio-temporal density of passengers inside metro systems [16]. Kusakabe *et al.* estimated which train was boarded by a passenger using smart card data [17]. Asakura *et al.* analyzed how the passengers changed their travel behaviors to accommodate the changes of train schedules [18]. Zhang *et al.* [19] proposed a method to extract spatio-temporal segmentation information of trips in a metro system, which is importance to estimate the spatio-temporal density inside a complex metro system. Zhao *et al.* [20] proposed an approach to estimate each passenger's route choice pattern in a complex metro network. Sun *et al.* formulated three optimization models to design

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY ROORKEE. Downloaded on December 13,2024 at 11:37:43 UTC from IEEE Xplore. Restrictions apply.

demand-driven timetables for a single-track metro service [21]. Sun *et al*. also proposed an integrated bayesian statistical inference framework to characterize passenger flow assignment model in a complex metro system [22]. In this paper, we leverage smart card data to investigate individual's travel patterns in terms of space and time.

2) Behavior and Pattern Analysis: Recently, the increasing scales of urban infrastructures enable the collection of massive data, which provides the possibility to investigate user behaviors. A number of efforts have been made based on different data. One category of research is on communication data for user behavior and pattern analysis. For instance, Isaacman *et al*. [23] proposed an approach to modeling how people move within cities. Isaacman *et al*. [24] validated daily travel patterns for the people in two cities. Dufková *et al*. [25] proposed a method to predict the locations of cellphone users. Isaacman *et al*. [26] proposed to identify important locations of cellphone users. Next, GPS data are also widely used to analyze user behavioral patterns. For example, the studies [27], [28] identified user locations based on the data from taxis and buses, respectively. Ge *et al*. [29] predicted taxi passenger demands. Ye *et al*. [30] presented a framework to effectively retrieve life patterns from GPS data. Liu *et al*. [31] study rationality from the perspective of user decision on visiting a point of interest (POI).

Then, for smart card data, most of the aforementioned research [4]–[6], [9], [11]–[14] mainly focused on mainly mine first-level knowledge about travel behavior, like significant places, possible destinations, regular trips. Compared with these studies, we study individual¡¯s general travel style and regularity travel patterns in terms of space and time. Based on three aspects, temporal, spatial and spatio-temporal, we define individual's travel pattern and we propose methods to retrieve the patterns. Then, in terms of anomaly detection and regularity discovery, we analyze individual passengers' travel patterns by statistical-based and unsupervised cluster-based methods. Some interesting findings are made. Moreover, to the best of our knowledge, this is the first study to analyze metro individual passengers¡¯ travel patterns, which exhibits significance to many potential real-life applications.

3) Anomaly Detection: Anomaly detection is a problem of finding outliers in the data. Till now, many anomaly detection techniques have been specifically developed for various applications [32], [33]. The techniques can be categorized as classification-based, distanced-based, clustered-based, and statistical-based, etc . Many existing anomaly detection studies focused on travel patterns extracted by GPS data. For instance, Pang *et al*. [34] detected flawed urban planning using GPS trajectories of taxi. Zheng *et al.* [35] detected collective anomalies (a collection of nearby locations that are anomalous during a few consecutive time intervals) from multiple spatio-temporal datasets. Lee *et al*. [36] proposed a partition-and-detect framework to detect outlying sub-trajectories from massive trajectory data. Ge *et al*. [37] provided an evolving trajectory outlier detection method by computing the outlying score for each trajectory in an accumulating way. Bu *et al*. [38] proposed an outlier detection framework for continuous trajectory streams. In this paper, based on individual passengers' travel patterns extracted by smart card data, we use statistical-based and clustering-based methods to detect outlier metro passengers.

TABLE I

SMART CARD DATA FORMAT

Field	Value
CardID	identifier of a smart card
TrmnlID	metro station ID or bus ID
TrnsctTime	Transaction time
TrnsctType	Transaction type: 31 indicates bus boarding, 21 represents tap-in a metro station, and 22 is tap-out a metro station.

B. Dataset

The smart card transaction dataset used in this study was collected from Shenzhen of China, a modern and big city in China. According to the government statistics, in Nov 2014, the average number of daily passengers in Shenzhen metro system is about 2.8 millions. As the smart cards could be used both by bus and metro systems, the data are mixed with both metro and bus transactions. In this paper, although we focus on metro systems, the transactions from bus can provide additional support as discussed in latter sections. Thus, instead of discarding the bus transactions, we also leverage their properties to improve our methods. The data are collected from about 4 million smart cards. In total, there exist approximately 210 million transactions from Nov 1 to Nov 30, 2014. Among the data, 41.9% are metro transactions and bus for the rest. The data format is shown in Table I. For metro systems, passengers need to swipe their smart cards when tapping-in and tapping-out, their boarding and alighting time and stations can be obtained directly. For bus passengers, they only need to tap smart card for boarding (i.e., check-in), the AFC system on buses does not save boarding location information, only bus ids, transaction time and so on. That means both of the boarding and alighting stops can not be derived directly. To derive boarding stop, we consider GPS coordinates of every bus, which is reported every 20-40 seconds, and public bus routes . Thus by matching the time stamps in AFC records and bus GPS records and then linking bus routes data, we can locate passengers boarding stations. However, when and where did the passengers get off a bus are still unknown, which is solved in Section III.

III. INDIVIDUAL'S TRAVEL PATTERN EXTRACTION

A. Individual's Travel Pattern Definition

In order to describe the process of extracting individual's travel patterns clearly, here we clarify the concept of a trip. A trip in this paper is meaning a one way journey from one activity to another for a particular purpose, such as travelling from home to workspace. A trip may consist of some journey segments defined as transfer between same modes (e.g. from one bus to another) or different modes (e.g. from bus to metro).

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY ROORKEE. Downloaded on December 13,2024 at 11:37:43 UTC from IEEE Xplore. Restrictions apply.

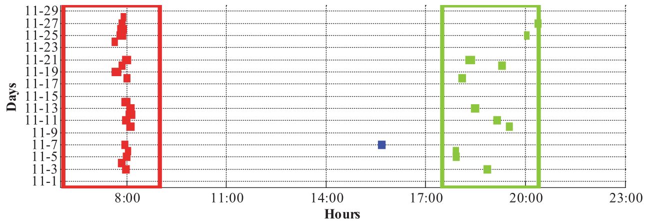


Fig. 2. The travel patterns of a metro passenger.

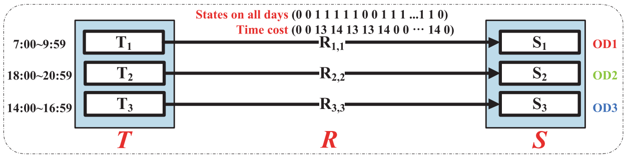

States on all days (0 0 1 1 1 1 1 0 0 1 1 1 ...1 1 0)
Time cost (0 0 13 14 13 13 14 0 0 ··· 14 0)
7:00~9:59
18:00~20:59
14:00~16:59
T1
T2
T3
R1,1
R2,2
R3,3
S1
S2
S3
OD1
OD2
OD3
T
R
S

Fig. 3. Three aspects of a passenger's travel patterns.

To illustrate the travel patterns of an individual passenger, we show an example in Fig. 2 based on a dataset of one-month smart card transactions. We plot all the trips of a passenger in each day over a month. In the figure, there are some trips represented by colored lines. Different colors represent different O-D pairs. The start and end points of a line represent begin and end time of the trip. The results show 3 observations. First, the number of most frequently visited OD pairs (approx. 95%) is two (ref., spatial patterns). Second, trips are repeated and *regular* in certain time periods (ref., temporal patterns): most days have trips within two periods, i.e, *7:00∼9:59* and *18:00∼20:59* marked by two different-colored rectangles. Third, the passenger is spatio-temporally regular, which means the trips are often from one fixed station to another during relatively fixed time periods. Accordingly, we deliberate three aspects of travel patterns of a passenger : spatial patterns $\mathcal{S}$, temporal patterns $\mathcal{T}$, and spatio-temporal patterns $\mathcal{R}$, respectively, as shown in Fig. 3 from the example in Fig. 2.

(Temporal travel pattern $\mathcal{T}$), which means the passenger has trips during the same time period over multiply days. An example of this type of patterns is: "In 80% of the days, Tom takes metro between 8:00 a.m. to 9:00 a.m.".

(Spatial travel pattern $\mathcal{S}$), which means that the passenger repeatedly visits the same O-D pairs over multiply days, like: "In 85% of the days, Tom departs from station A to station B";

(Spatio-temporal patterns $\mathcal{R}$), which represents the relationship between T and S and means that the passenger repeatedly visits the same O-D pairs during same time period over multiply days, like: "In 85% of trips during 8:00 a.m. to 9:00 a.m of all days, Tom departs from station A to station B".

B. Individual' Travel Pattern Recognition

For recognizing individual passengers' travel patterns, we first filter out the passengers who rarely take metro because such passengers have insufficient information to reveal the travel patterns. In this paper, we study those passengers who have more than seven days taking metro over one month. It is worth mentioning that the threshold (seven days) can be replaced by other value, which does not influence our method. In this following of this subsection, we first introduce how to construct the trips that belong to a specific passenger, then extract the travel patterns for the passenger.

1) Individual's Trip Construction: We construct two types of trips for a passenger. The trips of the first type are constructed using metro transactions for analyzing the passenger's travel patterns of using metro. The trips of the second type are constructed by combining bus and metro transactions, which are used to help interpret the passenger's travel patterns of the first type. Under the condition of network operation and seamless transfer between different metro lines, the first type of trips can be obtained by joining passenger's tap-in/tap-out records. While to construct the second type of trips, we need to (1) infer the alighting time and bus station for a bus passenger; (2) link journey segments(transfer between buses or bus and metro) for the passenger to derive trips.

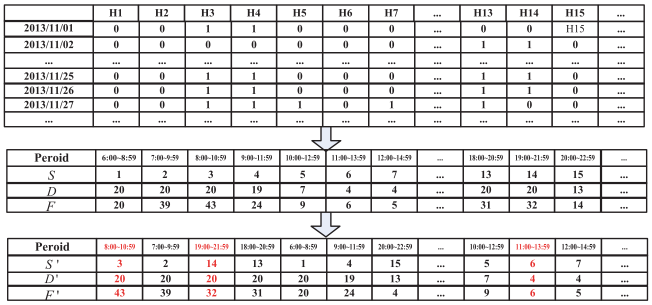
We infer the alighting stop for a bus passenger from the origin of the passenger's next boarding. We use $< R_{ID}, O_i, TO_i, Tag, V_i >$ to represent a boarding tuple (bus or metro), where O_i and TO_i represent the boarding station and time of the ith boarding of a passenger R_{ID}. Tag denotes transportation mode (1-bus, 0-metro). If $Tag = 1$, V_i is bus id. We need to infer alighting station D_i and time TD_i if $Tag = 1$. We make estimation D_i to be O_{i+1} and TD_i to be the time when the bus V_i arrived at station s, if there exists a station s the bus V_i passed through later, and $s \approx O_{i+1}$ (where $x \approx y$ are two adjacent stations, e.g., the stations of opposite directions of the same bus line at the same location, or the transit stations shared by two lines). For the other cases, D_i is set to *null* and TD_i is set to the time derived by adding an average in-vehicle time of the passengers with the same bus line and boarding station to TO_i.

For linking journey segments to derive trips, we need to consider the transfer time between two segments of a trip. Our survey reveals that more than 99% of the transfer activities in Shenzhen are less than 40 mins. So if the time difference between TD_i and TO_{i+1} is greater than 40 mins, the two segments are divided into different trips, else the two represent a transfer activity.

Through the above steps, we can obtain the trips for each individual. A trip is associated with begin time, origin station, end time and destination station (we ignore the routes or mode changes between origin and destination. In fact the routes can also be viewed as one type of travel pattern, which is not considered in this paper and left for future work). Note that the destinations of a minor part of trips are hardly to be inferred. However, this does not have a big impact on the accuracy of the individual's travel pattern recognition and regularity or anomaly analysis.

2) Temporal Pattern $\mathcal{T}$ Extraction: An ordered set $\mathcal{T} = \{T_1, T_2, ..., T_n\}$ is used to represent the temporal patterns of a passenger. An item T_i is associate with $\{T_i.t, T_i.r\}$, $T_i.t$ is a time slot, $T_i.r$ represents the proportion of the number of active days during time slot $T_i.t$. The time slots between any two items of $\mathcal{T}$ are not overlapped. e.g, $T_1.t$ and $T_2.t$ of the passenger in Fig. 2 are *7:00∼9:59* and *18:00∼20:59*, which are disjoint. T_i indicates the ith most frequent active time slot of the passenger. To find the temporal patterns, we need

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY ROORKEE. Downloaded on December 13,2024 at 11:37:43 UTC from IEEE Xplore. Restrictions apply.



	H1	H2	H3	H4	H5	H6	H7	...	H13	H14	H15	...
2013/11/01	0	0	1	1	0	0	0	...	0	0	H15	...
2013/11/02	0	0	0	0	0	0	0	...	1	1	0	...
...	...	...	...	...	...	...	...	...	...	...	...	...
2013/11/25	0	0	1	1	0	0	0	...	1	1	0	...
2013/11/26	0	0	1	1	0	0	0	...	1	1	0	...
2013/11/27	0	0	1	1	1	0	1	...	1	0	0	...
...	...	...	...	...	...	...	...	...	...	...	...	...

Peroid	6:00~8:59	7:00~9:59	8:00~10:59	9:00~11:59	10:00~12:59	11:00~13:59	12:00~14:59	...	18:00~20:59	19:00~21:59	20:00~22:59	...
S	1	2	3	4	5	6	7	...	13	14	15	...
D	20	20	20	19	7	4	4	...	20	20	13	...
F	20	39	43	24	9	6	5	...	31	32	14	...

Peroid	8:00~10:59	7:00~9:59	19:00~21:59	18:00~20:59	6:00~8:59	9:00~11:59	20:00~22:59	...	10:00~12:59	11:00~13:59	12:00~14:59	...
S'	3	2	14	13	1	4	15	...	5	6	7	...
D'	20	20	20	20	20	19	13	...	7	4	4	...
F'	43	39	32	31	20	24	4	...	9	6	5	...

Fig. 4. Temporal pattern $\mathcal{T}$ extracting procedure.

to tag time slots and counts the trips during each time slot. A time slot T_i should not only be able to reflect an individual's temporal patterns, but it also shall be convenient for aggregated analysis in the following sections. Thus we use sequential, overlapped and fixed time length slots to automatically compute the patterns over the transaction data. Note that, if we split a day into fixed intervals as tagged time slots: *00:00∼02:59*, *03:00∼05:59*, *06:00∼08:59*, *09:00∼11:59*, ..., we are hardly to count trips in the period of *08:30∼10:00*. By overlapped slots, *00:00∼02:59*, *01:00∼03:59*, ..., *07:00∼09:59*, *08:00∼10:59*, ..., all the trips in *8:30∼10:00* can be counted in the slot *08:00∼10:59*. In this paper, the length of each slot is set to 3 hours since the length of 98% trips do not exceed 2 hours in our test data in Shenzhen, China. In addition, for other reasons (such as weather), the passengers' schedule might be changed, we add 1 hour floating-time in order to cover the whole trip in the data. In the following, we explain the method for temporal pattern extraction by the three steps shown in Fig. 4.

Step 1: We use a table H to represent the temporal travel attributes of each passenger as shown in Fig. 4: each row represents one day and each column represents one hour (*H1* denotes *06:00∼6:59*, *H2* denotes *7:00∼9:59*, ..., *H19* denotes *00:00∼00:59*); each value in the table is 1 if the card of the passenger is active in the corresponding one-hour period, elsewise 0.

Step 2: This step aggregates the temporal travel attributes for each passenger. As shown in the middle part of Fig. 4, S is the time slot sequence number, D is the active days in that slot, and F is the active hours in that slot. D and F are calculated by Equation (1) (i) and (ii), respectively. Note that D_{num} is the total number of days in the collected dataset.

$$(i) D_{1,i} = \sum_{j=1}^{D_{num}} (H_{j,i} || H_{j,i+1} || H_{j,i+2}).$$

$$(ii) F_{1,i} = \sum_{j=1}^{D_{num}} (H_{j,i} + H_{j,i+1} + H_{j,i+2}). \tag{1}$$

Step 3: By D and F, we sort time slots S to obtain the ordered time slots S' as shown in the bottom table of Fig. 4. Then, we compute the nonoverlap time slots S'' by Algorithm 1. The algorithm iterates S' and add a item S'_i into S'', where S'_i has no overlap with any item in S''. Given S'', we can obtain all ordered time slots of a passenger. The top three time slots are marked in red in the bottom table of Fig. 4.

Algorithm 1 Extract the Temporal Patterns of a Passenger

```
Require: The ordered time slots: $S'$, the size of $S'$: $N$
Ensure: A set of nonoverlap time slots: $S''$
$S'' \leftarrow Null$
for $i = 0 \rightarrow N$ do
    $k \leftarrow size(S'')$
    $tag \leftarrow false$
    for $j = 0 \rightarrow k$ do
        //Are $S[i]$ and $S''[j]$ overlapping
        if $!isOverlap(S''[j], S[i])$ then
            $tag \leftarrow true$
            break;
        end if
    end for
    if $!tag$ then
        //Add $S'[i]$ to the set $S''$
        $Add(S'', S'[i])$
    end if
end for
return $S''$
```

3) Spatial Pattern S Extraction: We obtain all the trips for a passenger, and each trip has origin and destination represented by physical bus or metro stations. However, there may be more than one physical stations around a passenger' real origin or destination. These physical stations may have different spatial coordinates, although they are very close to each other. So before extracting a passenger's spatial patterns, we firstly group different stations by semantic meanings. We use the method proposed in [30] that applied OPTICS clustering method to cluster stations into several geographical regions. Then we use an ordered set $\mathcal{S} = \{S_1, S_2, ...\}$ to represent the spatial patterns of a passenger. An item S_i indicates an origin $S_i.o$, a destination $S_i.d$, the number of trips $S_i.n$ between $S_i.o$ and $S_i.d$, the proportion of the number of active days for the OD pair(from $S_i.o$ to $S_i.d$): $S_i.r$. All items in S are sorted by $S_i.r$. In other words, S_i denotes the passenger's ith most frequently accessed OD pair. To obtain the set, for a passenger, we first group all the trips by OD pairs. Then, for each OD pair, we calculate $S_i.n$ and $S_i.r$. Finally, we sort all OD pairs by $S_i.r$ in descending order.

4) Spatio-Temporal Patterns: R: We use a set $\mathcal{R} = \{R_{i,j}\}$ to represent the spatio-temporal patterns of a passenger. An item $R_{i,j}$ indicates the relationship between the ith time slots T_i in T and the jth OD pair in S, it is associated with a V-dimension boolean vector $R.S$ and a V-dimension integer vector $R.C$. $R.S$ keeps the information about whether the passenger has the trip S_j during the ith time slots T_i over the days in the data. $R.C$ indicates the time cost of the corresponding trip, which is calculated by subtracting the tap-out time of the trip from the tap-in time.

IV. Statistical-Based Travel Pattern Analysis

In this section, we proceed to analyze the regularity and anomaly of metro individual passengers' travel patterns using statistical-based method based on $\mathcal{S}$, $\mathcal{T}$, and $\mathcal{R}$.

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY ROORKEE. Downloaded on December 13,2024 at 11:37:43 UTC from IEEE Xplore. Restrictions apply.

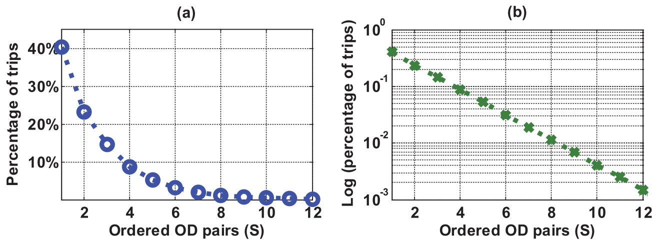


Fig. 5. Trips distribution of ordered OD pairs.

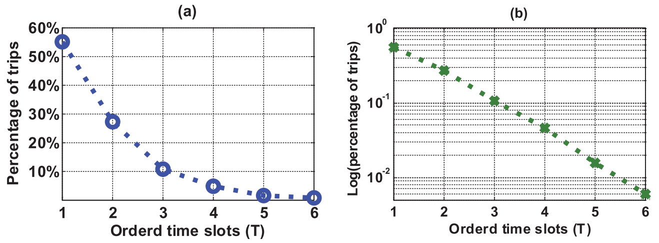


Fig. 6. Trips distribution of ordered time slots.

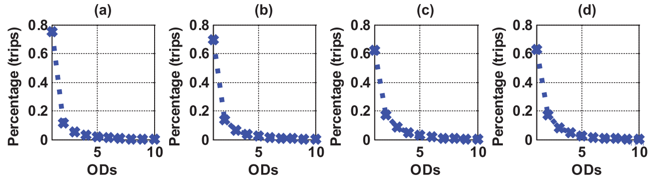


Fig. 7. The proportion of trips of different OD pairs during the first four time slots.

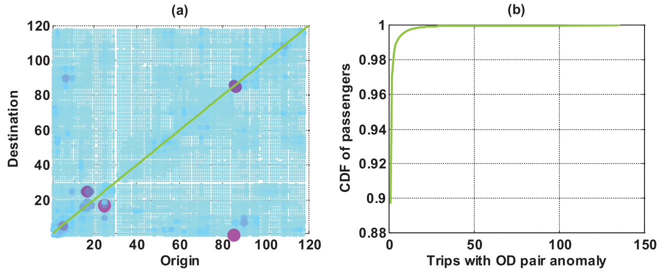


Fig. 8. (a) The passengers of different OD pairs; (b) The cumulative distribution of passengers with OD pair anomaly.

A. Travel Pattern Regularity Analysis

1) Trip Distribution Analysis Based on Spatial Pattern S: Base on all individual passenger's spatial patterns S, where all OD pairs are sorted by trip frequencies. We calculate the average proportion of trips of different OD pairs to total trips. The results are shown in Fig. 5 (a). The findings show that most of the trips are focused on the top two OD pairs (ref., about 65%). The reason is that most passengers have limited active position, e.g. between home and office. If we redraw the figure in terms of logarithmic coordinates, we can get Fig. 5 (b). The results show almost a linear line, $y(x) = e^{-0.5075*x-0.4158}$, which means that the number of trips of different OD pairs belongs to an exponential distribution. That's means most passengers will have a high proportion activity in a small area under the physical constrain.

2) Trips Distribution Analysis Based on Temporal Pattern T: Given all individual passenger's temporal patterns T where all time slots of a passenger are sorted by trip frequency. We calculate the average proportion of trips during different time slots of T. The results are shown in Fig. 6 (a), where X-axis represents the ordered time slots, Y-axis represents the proportions of trips during the time slots. The findings show that most of trips concentrated in the top two time slots, during which the sum of trips reaches 80%. That's because most passengers have regular life and schedule, e.g. everyday they leave home for work during a fixed time slot and go back home during another fixed time slot. If using logarithmic coordinates, we can get Fig. 6 (b). The time–trips relation is almost linear. The equation is $y(x) = e^{-0.5100*x-0.3987}$. According to the maximum entropy principle, it belongs to an exponential distribution. It also means that due to life regularity constrains, most people have trips in regular time.

3) Relationship Between T and S: Based on the relationship R between the time slots and the visited OD pairs of each passenger, we calculate the average proportion of trips of different OD pairs during the first four time slots in T, as shown in Fig. 7. The findings show that passengers mostly travel with one OD pair (ref., more than 60%). Particularly, in the first time slot, the percentage is almost 80%, which indicates that passengers are spatio-temporally regular. In other words, they frequently travel the same OD pairs in the relatively fixed time slots.

B. Anomaly Detection From Empirical Knowledge

1) OD Pair Based Anomaly Analysis: Fig. 8 (a) shows the number of passengers of each OD pair. In the figure, a station is indicated by a unique number (x axis represents origin, and y axis represents destination), and the size of a point represents the number of passengers of the corresponding OD pair (the bigger the more). From Fig. 8 (a), we have the following two findings.

First, without considering the special point marked by a red rectangular, the points in Fig. 8 (a) are almost symmetric along the diagonal. It means that the total number of passengers from a station to another is almost the same with that of opposite direction. The special point represents the passengers from station *FuTianKouAn* to *Luohu*, both of which are large public transport junctions in Shenzhen. The passengers between them are mostly transferring passengers, which yields an anomaly behavior of passengers.

Second, we can find that there are some points falling on the diagonal, which means that some passengers tap-in and tap-out from the same stations (we have filtered out the smart card transactions of metro staffs). If a passenger often performs such a behavior, the passenger is considered to be abnormal. Note that our hypothesis is that if the number of such trips of a passenger is greater than a threshold, the passenger is abnormal. We show the cumulative distribution of these passengers versus the number of such trips, as shown in Fig. 8 (b). Almost 95% passengers have less than five such trips. In this paper, we set the confident level to 95% (often used in statistics) or set the φ to 5, that means we set the

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY ROORKEE. Downloaded on December 13,2024 at 11:37:43 UTC from IEEE Xplore. Restrictions apply.

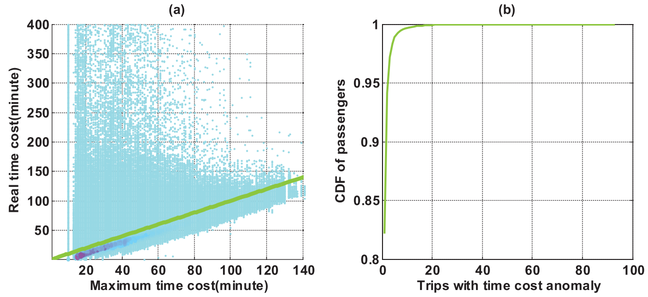


Fig. 9. (a) The time cost of trips (b) The cumulative distribution of passengers with time-cost anomaly.

top 5% of these passengers who are with the topmost such trips as spatial anomaly passengers. For the spatial anomaly passengers, we are to explain in the Section V.

2) Time Cost Based Anomaly Analysis: The time costs of various trips of same OD pair may be different. The expected maximum time L from one station to another can be estimated. Assume the minimum time from A to B is $L1$, which can be obtained by calculating the minimum time cost of all the trips of the OD pair. The longest waiting time is $L2$ and an assigned error $L3$ is set to half hour in this paper. Then L can be calculated by summing up the three: $L = L1+L2+L3$. The expected maximum time cost versus actual time cost is shown in Fig. 9 (a). We can find that there are some trips falling above the blue line $x = y$. That means there are some trips with time cost longer than the maximum value. Similar to the method for OD pair-based anomaly detection, if the number of such trips is more than a threshold ϕ, the passenger is abnormal, otherwise isn't. We show the cumulative distribution of passengers versus the total number of such trips, as shown in Fig. 9 (b). Almost 95% passengers have less than five such trips. In this paper, we set the confidence level to 95% or $\phi = 5$. That means if a passenger who has more than five such trips, the passenger shall be a time cost anomaly, otherwise normal. We will give explanation for these abnormal passengers in the following Section V .

V. Clustering-Based Travel Pattern Analysis

A. Passenger Clustering Based on Temporal Patterns

Fig. 10 shows the distribution of trips in respect of time of six special passengers over one month. We can find that the three passengers shown in Fig. 10 (a–c) are similar: they all have two peak time slots. Meanwhile, the three passengers in Fig. 10 (d–f) are similar: they all have one peak time slot. To find the similar passengers is important for metro companies to provide better services or market. Thus in this subsection, we present our method to automatically cluster passengers into groups by temporal patterns.

1) Temporal Patterns From Metro Data: For the temporal patterns of a passenger T, if the length of each slot T_i is set to 3 hours, we can get at most 8 time slots in a day. However, most of passengers mainly travel in two or three slots. In order to increase effectiveness of our cluster algorithm and avoid dimension disaster, we only select the top n slots for each passenger. The selection is based on two aspects. First, the chosen n features should convey temporal information as much as possible. Second, the n value should be as small as possible to improve the scalability of the following analysis. Similarly, we select the top n OD pairs for spatial parameters for spatial analysis in the following subsection V-B. In Fig.6 (a) and Fig.5 (a), we get that when $n = 4$, both the trip proportions of the first 4 OD pairs and the first four time slots exceed 85%, which is sufficient for analyzing passenger's travel regularity. As analyzed in section IV-A, due to physical and life regularity constraints, most passengers have most of activities in a small area during few time slots, so we set n to 4 in this paper.

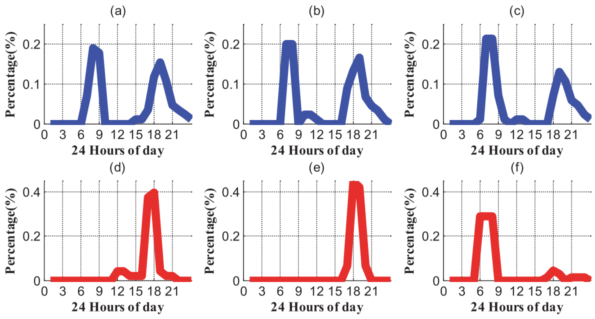


Fig. 10. The distribution of travel time of six special passengers.

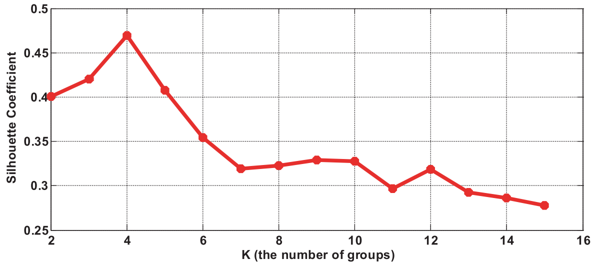


Fig. 11. Silhouette coefficient.

Next, we use K-means algorithm and city-block distance measure to cluster passengers based on the 4-dimensional temporal features: Ft_1, Ft_2, Ft_3, Ft_4. For a passenger, Ft_i is the proportion of active days during the ith time slot T_i to the total. We use average silhouette coefficient [39] to determine the number of groups. The silhouette coefficient value is used to measure of how similar a passenger is to the cluster (cohesion) compared to others (separation). For the ith passenger, the silhouette coefficient is calculated by $s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$, where a_i is the average dissimilarity of the ith passenger with all other passengers within the same cluster. b_i is the lowest average dissimilarity of ith passenger to any other cluster. The silhouette coefficient is in $[-1, 1]$, where a large value indicates that the passenger is well matched to the cluster and poorly matched to others. From Fig. 11, we find that when $k = 4$, silhouette coefficient is optimal. Thus, we obtain 4 groups: *TGrp 1*, *TGrp 2*, *TGrp 3*, and *TGrp 4*.

We plot the cluster centers of the 4 groups in Fig. 12. The results show that passengers in *TGrp 1* have one dominant time slot; passengers in *TGrp 2* have two dominant time slots; the time slots in *TGrp 3* from the biggest to the smallest have similar gap; and the time slots have no significant difference in *TGrp 4* and all are relatively small. The four groups have

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY ROORKEE. Downloaded on December 13,2024 at 11:37:43 UTC from IEEE Xplore. Restrictions apply.

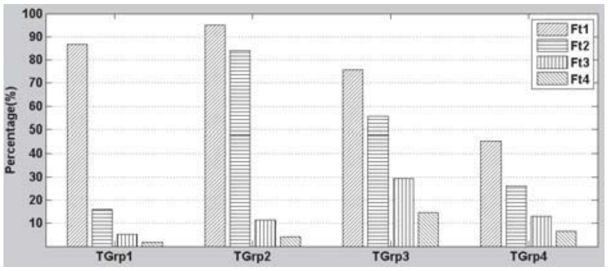


Fig. 12. The cluster centers of *TGrp 1*, *TGrp 2*, *TGrp 3* and *TGrp 4*.

TABLE II
THE JOINT DISTRIBUTION OF *TG*rp AND *BTG*rp

Group	*BTGrp-1*	*BTGrp-2*	*BTGrp-3*	*BTGrp-4*	*Total*
TGrp-1	5.48%	7.78%	2.54%	0.04%	15.83%
TGrp-2	0.00%	31.70%	0.94%	0.00%	32.64%
TGrp-3	0.03%	3.55%	23.52%	0.002%	27.11%
TGrp-4	0.58%	5.73%	15.17%	2.94%	24.41%
Total	6.08%	48.75%	42.17%	3.00%	100%

15.83%, 32.64%, 27.11%, 24.41% of the total passengers, respectively.

Specifically, consider *TGrp 2*, the passengers regularly travel at two certain periods by metro. For each period, the number of active days is more than 80% of all their own active days. Our investigation shows this group contains passengers who regularly commute back and forth to work or other activities during the morning and evening peaks.

The *TGrp 1* passengers regularly travel in one certain period by metro. During the period the number of active days is more than 85%. A question may arise, if they go to work by metro, how do they go back home? or vice versa, how do they go to work? There are few possibilities: by bus or taxi. Furthermore, there is no distinct temporal regularity for *TGrp 3* and *TGrp 4*. In order to understand the passengers' temporal pattern, we incorporate bus data to analyze in the following section.

2) Temporal Patterns From Metro and Bus Data: By incorporating bus data, we can better understand the above question. With bus data, we re-cluster all passengers using the same cluster centers and the 4 new groups are denoted as *BTGrp 1*, *BTGrp 2*, *BTGrp 3* and *BTGrp 4*. They have 6.08%, 48.75%, 42.17%, 3.00% of the total passengers, respectively.

We match the passengers to both *TGrp* and *BTGrp* in Table II. For the 15.83% passengers in *TGrp 1*, 7.78% move to *BTGrp 2*: though metro is faster and punctual, but it is more expensive than bus; to save money, some passengers take metro in one trip and return by bus. There are still 5.48% passengers remain in *BTGrp 1*: they could be divided into two categories, those who finished a round trip in three hours and those who chose other transportation tools instead of bus or metro, such as shuttles or taxi.

Fig. 13 shows the distribution of the first time slot T_1 between *TGrp 1* and *BTGrp 1*. As we can see, both curves have remarkable AM and PM peak hours; there is a remarkable reduce in the number of passengers in *BTGrp 1* compared with *TGrp 1*, especially in AM and PM peak hours. It suggests that most of the passengers move from *TGrp 1* to *BTGrp 2* are traveling during peak hours.

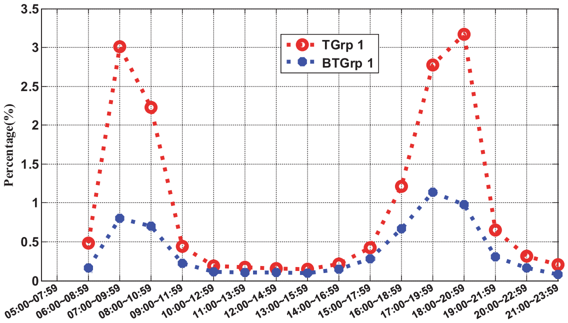


Fig. 13. The distribution of the T_1 of the passengers in *TGrp 1* and *BTGrp 1*.

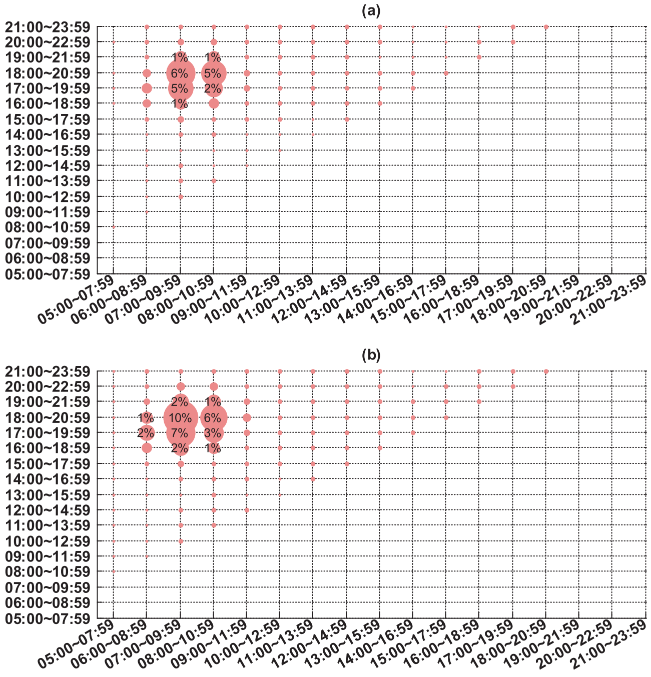


Fig. 14. Joint probability distribution of the first and second time slots of the passengers in (a)*TGrp2* (b)*BTGrp2*.

Not surprisingly, nearly all the *TGrp 2* passengers (i.e., 31.7%) fall into *BTGrp 2*; Fig. 14 shows the joint probability distribution of the first and second time slots T_1 and T_2 of the passengers in *TGrp 2* and *BTGrp 2*. As we can see, the number of passengers traveling during AM and PM peak hours increases from 22% to 37% after being reclassified. The results verify that most passengers travel in two certain periods, e.g., forth and back to work or other activities. We also notice that there are some passengers who also regularly travel during two certain periods, but not during peak hours. The possible reason is that these passengers normally work shifts, such as as regular evening, rotating shift, or some other schedule.

In Table II, for the 27.11% *TGrp 3* passengers, most of them (i.e., 23.52%) fall into *TGrp 3*. The passengers mainly rely on metro for round trips, but the travel time of them is irregular. Our investigation shows that some of them often work overtime or have a lot of leisure activities. A small portion of passengers (i.e., 3.55%) fall into *TGrp 2*.

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY ROORKEE. Downloaded on December 13,2024 at 11:37:43 UTC from IEEE Xplore. Restrictions apply.

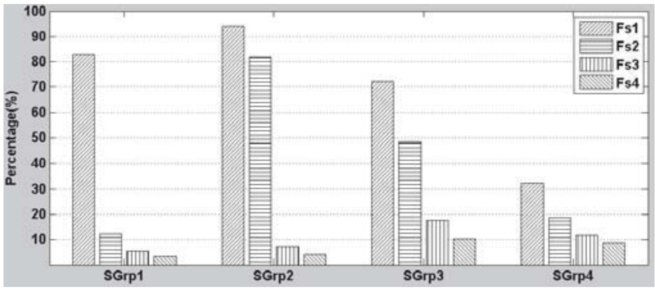


Fig. 15. The cluster centers of *SGrp 1*, *SGrp 2*, *SGrp 3* and *SGrp 4*.

The passengers regularly travel during two certain periods by metro or bus.

In Table II, for the 24.41% passengers in *TGrp 4*, most passengers (i.e., 20.9%) fall into *BTGrp 2* and *BTGrp 3*. The passengers mainly rely on bus for round trips. The rest of the passengers either rely on other tools, or have no regular temporal patterns.

B. Passenger Clustering Based on Spatial Patterns

1) Spatial Patterns From Metro Data: A passenger has 4-dimensional spatial features, Fs_1, Fs_2, Fs_3, Fs_4. The ith feature is the proportion of the passenger's active days to access the ith OD pair. Metro passengers can be spatially clustered into 4 groups: *SGrp1*, *SGrp2*, *SGrp3*, and *SGrp4*. They have 15.08%, 27.43%, 26.09%, 31.40% of the total passengers, respectively. The cluster centers of these groups are shown in Fig. 15. *SGrp1* passengers have only one frequently accessed *OD*-pair, of which the portion of days visiting the OD pair (i.e., access days) is 80%. The *SGrp2* passengers have two frequently accessed *OD*-pairs, of which the portions of access days are both greater than 80%. The *SGrp3* passengers have one relatively frequently accessed *OD*-pair and another with less accesses. There is no remarkable frequently accessed *OD*-pair in *SGrp4*.

2) Spatial Patterns From Metro and Bus Data: By incorporating bus data, we can better understand passengers' spatial patterns. With bus data, we re-cluster all passengers using the same cluster centers and the 4 new groups are denoted as *BSGrp 1*, *BSGrp 2*, *BSGrp 3* and *BSGrp 4*. They have 3.95%, 58.68%, 35.86%, 1.51% of the total passengers, respectively. We match the passengers to both $SGrp$ and $BSGrp$ in Table III. For the 15.08% passengers in *SGrp 1*, 7.54% move to *BSGrp 2*. Over half of passengers belong to *BSGrp-2*. We calculate the ratio that the origin of the first OD-pair is same with the destination of the second OD-pair, and the destination of the first OD-pair is same with the origin of the second OD-pair. The result is 87.9% in *SGrp 2* and 80.3% in *BSGrp 2*, which may suggest that most of these passengers are commuters who take a home-to-work trip.

TABLE III

THE JOINT DISTRIBUTION OF SGrp AND BSGrp

Group	*BSGrp-1*	*BSGrp-2*	*BSGrp-3*	*BSGrp-4*	*Total*
SGrp-1	3.02%	7.54%	3.77%	0.75%	15.08%
SGrp-2	0.30%	24.69%	2.19%	0.25%	27.43%
SGrp-3	0.29%	10.98%	15.91%	0.23%	26.10%
SGrp-4	0.35%	16.02%	14.76%	0.28%	31.41%
Total	3.95%	58.68%	35.86%	1.51%	100%

TABLE IV

THE CONDITIONAL PROBABILITY OF *SG*rp GIVEN *TG*rp

Group	*SGrp-1*	*SGrp-2*	*SGrp-3*	*SGrp-4*
TGrp-1	72.8%	2.2%	10.7%	14.3%
TGrp-2	1.0%	70.1%	27.6%	1.3%
TGrp-3	5.8%	15.3%	48.2%	30.7%
TGrp-4	6.8%	0.2%	9.5%	83.5%

TABLE V

SELECTIONS OF TWO CLASSES OF PASSENGERS

class	*TGrp*	*BTGrp*	*SGrp*	*BSGrp*
class-1	1	2	1	2
class-2	2	2	2	2

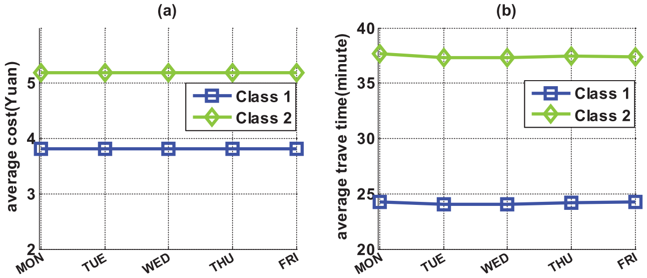


Fig. 16. Average cost and travel time of the two classes of passengers.

C. Relationship Between SGrps and TGrps

The conditional probability of *SGrp* given *TGrp* is shown in Table IV. We can observe that *SGrp* has strong correlation with *TGrp*, as 72.8%, 70.1%, 83.6% of temporal groups *TGrp 1*, *TGrp 2* and *TGrp 4* belong to the corresponding spatial group *SGrp 1*, *SGrp 2*, and *SGrp 4*. *TGrp 3* is correlated with both *SGrp 3* and *SGrp 4*. The implication is that most passengers in group 1 and 2 are spatio-temporally regular; most passengers in group 3 are relatively spatio-temporally regular; as a comparison, most passengers in group 4 are irregular in terms of both time and space. To find the reason for the passengers take metro in one trip and return by bus, we do the following analysis. First, we select two classes of passengers as shown in Table V. Class-1 represents those who are spatio-temporal regularly and take metro in one trip and take bus in the other, and class-2 represents those who take metro in round trips. Note that we have temporal label *TGrp*, extended temporal label *BTGrp* (combining with bus data), and spatial label *SGrp* and extended spatial label *BSGrp*. The comparison of average costs and travel times in five weekdays are shown in Fig. 16. The results show that the first class is less than the second class by cost and travel time. Taking metro is higher than taking bus in Shenzhen, and nearer distance from source to destination is in general more easily to find a direct bus for easy transportation. Thus, by economic reasons, some passengers may choose bus instead of metro.

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY ROORKEE. Downloaded on December 13,2024 at 11:37:43 UTC from IEEE Xplore. Restrictions apply.

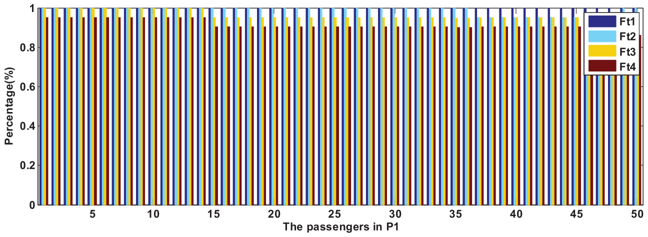


Fig. 17. The features for the top 50 passengers with most anomaly score.

D. Clustering-Based Anomaly Detection

In this section, we use a clustering-based anomaly detection method to find outlier passengers. Clustering based anomaly detection relies on the assumption that normal data instances lie close to their closest cluster centroids, while anomalies are far away from their closest cluster centroids. This technique consist of two steps: first, the data is clustered using a clustering algorithm; second, for each data instance, its distance to its closest cluster centroid is calculated as its anomaly score. Based on the result in section V-A, which cluster metro passengers based on temporal patterns, we can obtain the anomaly score for each passenger. According to anomaly scores, we choose the top 50 passengers, and the temporal features Ft_1, Ft_2, Ft_3, Ft_4 of them are shown in Fig. 17. We observe that almost all of them regularly travel by metro during four time slots. For each slot, the proportion of active days is more than 85%. We further calculate the average travel time a day for these passengers and the result is about 8 hours, which is far more than that of normal passengers with 2 hours travel time a day on average.

Till now, we have found three categories of anomaly passengers. The first category contains the passengers, whose origin and destination stations of the trips are usually the same. The second category is those, the time cost of whose trips are usually longer than the expected maximum value. The third category is the passengers whose average travel time a day is far longer than the normal passengers. For understanding these anomaly passengers, we conduct the following surveys to verify the findings.

1) Field Investigation: We consulted metro staffs, who observed that some logistics company staffs delivering packages by metro, usually bring packages to the nearest metro station to destination, and then they pass the packages to others who wait outside the fare gates of the station. Hereby, they do not need to tap-out at the station and tap-in again. By the method, within the maximum trip length (i.e., 3 hours), such a passenger only needs to pay the minimum fare by taping out from the tap-in station. We further consulted logistics company staffs who verify the observation of the metro staffs, and some of them need to work inside metro system longer than 8 hours. Therefore, their temporal patterns should be different with normal passengers. Additionally, there also exist homeless people and accompanying passengers.

2) Support From Actual Data: We have 55 thefts passengers' smart card IDs from public security department. We find that there are 15 of them belonging to at least one category of abnormal passengers in this study, 9 of them belonging to the first category, 7 of them belonging to the second category; 4 of them belonging to the third category.

Thus, these anomaly passengers may be logistics company staff, homeless passengers, thefts, and accompanying passengers. We leave further analysis for these passengers in the future work.

VI. Conclusion

In this paper, we study individuals general travel style and regularity travel patterns in metro system. Concretely, we first investigate individual¡¯ travel patterns in terms of space and time and we propose the method to retrieve the patterns based on the history smart card transaction data. Then in terms of anomaly detection and regularity discovery, we use statistical-based and clustering-based methods to understand the passengers' travel patterns. We have some important discoveries. For instance, if a passenger is temporally regular, it is very possible that the passenger is also spatially regular. We find that there are some passengers taking metro in one trip and return by bus, it is of importance to understand passenger¡¯s transportation mode choice behavior. We also find some outlier passengers with travel patterns different with general passengers. All of these discoveries is important for transportation researchers to improve metro services.

In near future, we will consider more factors such as passenger types (regular, student, staff), route choice (there may be several routes connecting two stops) to perform further analysis on individual passengers¡¯ travel patterns, and build a complete system to distinguish a special type of anomaly passengers from normal passengers. In addition, we plan to apply our research results in real applications, such as personalized scheduling and route recommending.

Acknowledgment

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

References

[1] E. C. Taylor and C. K. Jones, *Fair Fare Policies: Pricing Policies that Benefit Transit-Dependent Riders*. New York, NY, USA: Springer, 2012, pp. 251–272.

[2] B. Du, C. Liu, W. Zhou, Z. Hou, and H. Xiong, "Catch me if you can: Detecting pickpocket suspects from large-scale transit records," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 87–96.

[3] L. Sun, K. W. Axhausen, D.-H. Lee, and X. Huang, "Understanding metropolitan patterns of daily encounters," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 34, pp. 13774–13779, 2013.

[4] M. Trépanier, N. Tranchant, and R. Chapleau, "Individual trip destination estimation in a transit smart card automated fare collection system," *J. Intell. Transp. Syst.*, vol. 11, no. 1, pp. 1–14, 2007.

[5] M. Utsunomiya, J. Attanucci, and N. H. M. Wilson, "Potential uses of transit smart card registration and transaction data to improve transit planning," *Transp. Res. Record*, vol. 1971, pp. 119–126, Jan. 2006.

[6] X. Ma, Y.-J. Wu, Y. Wang, F. Chen, and J. Liu, "Mining smart card data for transit riders' travel patterns," *Transp. Res. C, Emerg. Technol.*, vol. 36, pp. 1–12, Nov. 2013.

[7] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: Concepts, methodologies, and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 3, p. 38, 2014.

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

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY ROORKEE. Downloaded on December 13,2024 at 11:37:43 UTC from IEEE Xplore. Restrictions apply.

[9] M. A. Munizaga and C. Palma, "Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile," *Transp. Res. C, Emerg. Technol.*, vol. 24, pp. 9–18, Oct. 2012.

[10] X.-L. Ma, Y.-H. Wang, F. Chen, and J.-F. Liu, "Transit smart card data mining for passenger origin information extraction," *J. Zhejiang Univ. Sci. C*, vol. 13, no. 10, pp. 750–760, 2012.

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

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

[13] K. K. A. Chu, R. Chapleau, and M. Trepanier, "Driver-assisted bus interview: Passive transit travel survey with smart card automatic fare collection system and applications," *Transp. Res. Record*, vol. 2105, pp. 1–10, 2009.

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

[15] X. Ma and Y. Wang, "Development of a data-driven platform for transit performance measures using smart card and GPS data," *J. Transp. Eng.*, vol. 140, no. 12, p. 4014063, 2014.

[16] 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.

[17] 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.

[18] Y. Asakura, T. Iryo, Y. Nakajima, and T. Kusakabe, "Estimation of behavioural change of railway passengers using smart card data," *Public Transp.*, vol. 4, no. 1, pp. 1–16, 2012.

[19] F. Zhang, J. Zhao, C. Tian, C. Xu, X. Liu, and L. Rao, "Spatiotemporal segmentation of metro trips using smart card data," *IEEE Trans. Veh. Technol.*, vol. 65, no. 3, pp. 1137–1149, Mar. 2016.

[20] J. Zhao *et al.*, "Estimation of passenger route choice pattern using smart card data for complex metro systems," *IEEE Trans. Intell. Transp. Syst.*, pp. 1–12, Aug. 2016.

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

[22] 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, Emerg. Technol.*, vol. 52, pp. 116–131, Mar. 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0968090X15000030

[23] S. Isaacman *et al.*, "Human mobility modeling at metropolitan scales," in *Proc. 10th Int. Conf. Mobile Syst., Appl., Ser. (MobiSys)*, New York, NY, USA, 2012, pp. 239–252. [Online]. Available: http://doi.acm.org/10.1145/2307636.2307659

[24] S. Isaacman *et al.*, "Ranges of human mobility in Los Angeles and New York," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops (PERCOM Workshops)*, Mar. 2011, pp. 88–93.

[25] K. Dufková, J.-Y. Le Boudec, L. Kencl, and M. Bjelica, "Predicting user-cell association in cellular networks from tracked data," in *Mobile Entity Localization and Tracking in GPS-Less Environnments* (Lecture Notes in Computer Science). Berlin, Germany: Springer, 2009, pp. 19–33.

[26] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, J. Rowland, and A. Varshavsky, "A tale of two cities," in *Proc. 11th Workshop Mobile Comput. Syst. Appl. (HotMobile)*, New York, NY, USA, 2010, pp. 19–24. [Online]. Available: http://doi.acm.org/10.1145/1734583.1734589

[27] R. Ganti, M. Srivatsa, A. Ranganathan, and J. Han, "Inferring human mobility patterns from taxicab location traces," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2013, pp. 459–468.

[28] S. Bhattacharya, S. Phithakkitnukoon, P. Nurmi, A. Klami, M. Veloso, and C. Bento, "Gaussian process-based predictive modeling for bus ridership," in *Proc. ACM Conf. Pervasive Ubiquitous Comput. Adjunct Publication*, 2013, pp. 1189–1198.

[29] Y. Ge, H. Xiong, A. Tuzhilin, K. Xiao, M. Gruteser, and M. Pazzani, "An energy-efficient mobile recommender system," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 899–908.

[30] Y. Ye, Y. Zheng, Y. Chen, J. Feng, and X. Xie, "Mining individual life pattern based on location history," in *Proc. 10th Int. Conf. Mobile Data Manage., Syst., Ser. Middleware (MDM)*, May 2009, pp. 1–10.

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

[32] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv. (CSUR)*, vol. 41, no. 3, p. 15, 2009.

[33] V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artif. Intell. Rev.*, vol. 22, no. 2, pp. 85–126, 2004.

[34] L. X. Pang, S. Chawla, W. Liu, and Y. Zheng, "On detection of emerging anomalous traffic patterns using GPS data," *Data Knowl. Eng.*, vol. 87, pp. 357–373, Sep. 2013.

[35] Y. Zheng, H. Zhang, and Y. Yu, "Detecting collective anomalies from multiple spatio-temporal datasets across different domains," in *Proc. 23rd SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst.*, 2015, p. 2.

[36] J.-G. Lee, J. Han, and X. Li, "Trajectory outlier detection: A partition-and-detect framework," in *Proc. IEEE 24th Int. Conf. Data Eng. (ICDE)*, Washington, DC, USA, Apr. 2008, pp. 140–149. [Online]. Available: http://dx.doi.org/10.1109/ICDE.2008.4497422

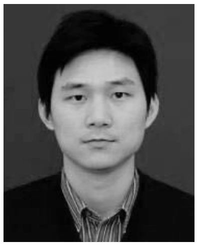
[37] Y. Ge, H. Xiong, Z.-H. Zhou, H. Ozdemir, J. Yu, and K. C. Lee, "Top-eye: Top-k evolving trajectory outlier detection," in *Proc. 19th ACM Int. Conf. Inf. Knowl. Manage.*, 2010, pp. 1733–1736.

[38] Y. Bu, L. Chen, A. W.-C. Fu, and D. Liu, "Efficient anomaly monitoring over moving object trajectory streams," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2009, pp. 159–168.

[39] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. 1, pp. 53–65, 1987.

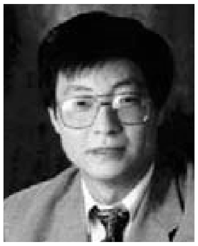
Juanjuan Zhao received the M.S. degree from the Department of Computer Science, Wuhan University of Technology, China, in 2009. She is currently working toward the Ph.D. degree with Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. She was a Research Assistant with Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, from 2009 to 2012. Her research interests include cloud computing, big data processing, streaming-data processing, data fusion technique, big-data-driven systems, spatio-temporal data mining.

Qiang Qu received the M.Sc. degree in computer science from Peking University and the Ph.D. degree from Aarhus University. He is currently an Associate Professor and the Executive Director of the Global Center for Big Mobile Intelligence, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. His research interests include large-scale data management and mining, mobile computing, and artificial intelligence.



Fan Zhang received the Ph.D. degree in communication and information system from Huazhong University of Science and Technology in 2007. He was a Post-Doctoral Fellow with The University of New Mexico and University of Nebraska-Lincoln from 2009 to 2011. He is an Associate Professor with Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. His research topics include big data processing, data privacy and urban computing.

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY ROORKEE. Downloaded on December 13,2024 at 11:37:43 UTC from IEEE Xplore. Restrictions apply.



Chengzhong Xu (F'16) received the Ph.D. degree from The University of Hong Kong in 1993. He is a Professor with the Department of Electrical and Computer Engineering, Wayne State University, USA. He also holds an adjunct appointment with Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, as the Director of the Institute of Advanced Computing and Data Engineering. He has published over 200 papers in journals and conferences. His research interests include parallel and distributed systems and cloud computing. He was the Best Paper Nominee of 2013 IEEE High Performance Computer Architecture (HPCA) and the Best Paper Nominee of 2013 ACM High Performance Distributed Computing (HPDC). He was a recipient of the Faculty Research Award, the Career Development Chair Award, and the President's Award for Excellence in Teaching of WSU. He was also a recipient of the Outstanding Oversea Scholar Award of NSFC. He 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 *China Science Information Sciences*.

Siyuan Liu received the Ph.D. degree from the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, and the second Ph.D. degree from the University of Chinese Academy of Sciences. He is an Assistant Professor with the Smeal College of Business, Pennsylvania State University. His research interests include spatial and temporal data mining, social networks analytics, and mobile marketing.

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY ROORKEE. Downloaded on December 13,2024 at 11:37:43 UTC from IEEE Xplore. Restrictions apply.