Clustering-based Travel Pattern Recognition in Rail Transportation System Using Automated Fare Collection Data

Yupeng Chen, Yang Zhao, Kwok Leung Tsui
School of Data Science
City University of Hong Kong
Hong Kong Special Administrative Region, P.R. China
chenyupeng386@163.com, yang.zhao@my.cityu.edu.hk, kltsui@cityu.edu.hk

***Abstract*—Passenger travel pattern analysis is essential for the design and development of public transport network. Nowadays, Automated Fare Collection (AFC) systems are widely exploited in the operation and management of public transportation. The data collected from AFC systems provide valuable information to analyze passenger behavior. This research aims to investigate passenger mobility patterns from both temporal and spatial perspectives. We present a hybrid topic-clustering method for extracting travel feature and grouping passengers based on their travel patterns. Our proposed method is illustrated using a real AFC dataset of the metro transportation system in Shenzhen, China. The results showed that four temporal travel patterns were well identified. Comparison of travel behavior indicated that metro travelers with different travel time selections also have different activity areas.**

Keywords- Public transport network; AFC data; travel pattern; hybrid topic-clustering method;

I. INTRODUCTION

Nowadays, Automated Fare Collection (AFC) systems are widely exploited in the operation and management of public transit system. While AFC systems is mainly utilized for automatically collect ticket fare, they also record large volume of smart card users' travel information, such as transaction time, boarding stop, and alighting stop, etc. Mining AFC data enables us understanding passenger travel patterns, which can facilitate the traffic network operation and improve the service of public transit system.

From the perspective of the method employed in travel pattern recognition using AFC data, clustering analysis is a common approach for discovering passenger groups with different mobility patterns. Zhao et al. [1] performed passenger travel patterns clustering via K-means algorithm to detect outlier transit users, such as logistics company staffs and thieves. Agard et al. [2] presented a two-step clustering framework to group public transit users. The number of clusters in K-means was determined by the results of the hierarchical clustering method. In order to produce a flexible solution for spatial and temporal pattern analysis, Bhaskar et al. [3] developed a two-level DBSCAN method to mine regular ODs and habitual time for further disaggregating passengers of different travel behaviors. For trading off robustness with computational efficiency, Ortega-Tong et al. [4] employed the K-medoids clustering algorithm to estimate homogeneous passenger groups based on sociodemographic characteristics and travel features.

The ever-increasing volume of data presents two major challenges for the analysis of AFC travel records: travel feature extraction and the selection of data mining techniques. For these two issues, in this paper, we propose a hybrid topic-clustering method to discover typical passenger groups with various temporal travel patterns. We believe that this paper not only provides a better knowledge of inhabitants' trip behavior in the metropolis but also improves the public transit service and urban planning.

The rest of this paper is organized as follows. Section II provides a description of trip construction and temporal trip record retrieval. The hybrid topic-clustering approach to extract temporal travel habits and cluster passengers is presented in section III. The AFC data used in this study is described in section IV. Cluster result and travel pattern comparison are discussed in Section V. Section VI concludes this paper and our future research work.

II. TRAVEL RECORD PROCESSING

One of the critical tasks in AFC data preprocessing for subsequent data analysis is constructing and processing the complete trip chains of passengers. In this section, we illustrate the method of aggregating multi-stage travel segments and retrieving temporal trip records.

A. Trip Construction

One entrance record with the subsequent exit record constitute a complete journey segment $\langle O, T^{in}, D, T^{out} \rangle$ which

This work was supported by the National Natural Science Foundation of China (No. 11471275), and the Research Grants Council Theme- based Research Scheme (No. T32-101/15-R).
978-1-7281-0860-5/19/$31.00 ©2019 IEEE

contains four attributes: origin station O, check-in time T^{in}, destination station D, and check-out time T^{out}. Tuple $\left\langle O_m, T_m^{in}, D_m, T_m^{out} \right\rangle_l$ records the m th journey segment of passenger l, where $m = 1,\ldots,M$, $l = 1,\ldots,L$.

In the case of a journey that passenger cannot directly reach the final destination, one trip may contain several journey segments because of the transfer activities. To link these discrete journey segments together, we employ the "time threshold" method [5, 6] to identify transfers. This approach determines whether there is a transfer by setting an upper limit of the time interval between two journey segments. Specifically, if the time difference between T_{m+1}^{in} and T_m^{out} is less than the time threshold, the two segments are aggregated into one trip $\left\langle O_m, T_m^{in}, D_{m+1}, T_{m+1}^{out} \right\rangle_l$, otherwise they will be divided into different trips. It is worth noting that the time thresholds given by different studies are not consistent, depending on the characteristics, culture, and habits of the city studied.

B. Temporal Trip Record Retrieval

The temporal trip record for transit user refers to the trip counts over time. Suppose we separate one day into I time slots $\mathbf{T} = (T_1,\ldots,T_i,\ldots,T_I)$, for passenger l, his temporal trip trajectory during the study period can be denoted as $\mathbf{t}_l = \{t_{l1},\ldots,t_{ln},\ldots,t_{lN_l}\}$, where $t_{ln} \in \mathbf{T}$ represents the time slot of his n th trip and N_l is the total number of trips he has made. Then we sum the trips in same time slot and use an ordered set $\mathbf{TP}_l = \{TP_{l1},\ldots,TP_{li},\ldots,TP_{lI}\}$ to describe the temporal trip record of passenger l, where TP_{li} is the trip counts in time slot T_i.

III. HYBRID TOPIC-CLUSTERING METHOD

Temporal trip records imply people's travel patterns at time level. However, the relatively fixed travel time and the repeatability of daily travel make $\mathbf{TP}_l$ a high dimensional sparse vector [7, 8]. This creates obstacles for further feature extraction and passenger groups partition. To address this issue, a hybrid topic-clustering method is proposed. By fitting a Latent Dirichlet Allocation (LDA) model, we first extract people's temporal travel habits from their AFC trip records. Then, based on the output of LDA, Passenger-Temporal travel habit distribution, we use the Hierarchical Clustering (HC) method to cluster transit users.

A. Temporal Travel Feature Extraction

LDA is a generative probabilistic model used in text mining. This method represents documents as mixtures of topics, where each topic is characterized by a discrete distribution over words [9]. As an unsupervised machine learning technique, LDA can automatically find the underlying structure of topics. In this scenario, we reflect the problem of identifying the latent temporal travel habits of transit users to the problem of discovering the latent topics of a document.

Assume there are K types of temporal travel habits $\mathbf{H} = (H_1,\ldots,H_k,\ldots,H_K)$ that represent people's various mobility patterns in the time dimension. As shown in table 1, we regard each passenger as a "document" containing a collection of "words", with each word being a time slot t_{ln}, and "topics" means temporal travel habits $\mathbf{H}$. In other words, a passenger having multiple selections of temporal travel habits is just like a document containing a variety of topics.

Fig. 1 shows the analogy using an example. We use the Passenger-Time slot matrix instead of the Document-Word matrix as the input to LDA. The temporal trip record of passenger l is denoted by $\mathbf{TP}_l$, the value of TP_{li} denotes the occurrences of the i th time slot.

One key inferential problem in LDA modeling is to estimate the Topic-Word distribution (a distribution over words for each latent topic) and the Document-Topic distribution (a distribution over the topics for a given document), which can be accomplished by a variety of approximate inference approaches, such as Markov Chain Monte Carlo and Gibbs sampling [10]. In this case, the Temporal travel habit-Time slot distribution (Topic-word distribution) reveals the differences between people's various temporal move patterns, and the Passenger-Temporal travel habit distribution (Document-Topic distribution) quantifies transit user's preference for travel time selection and is used as the input of HC algorithm.

B. Identification of Typical Passenger Groups

Hierarchical clustering is a widely used technique for cluster analysis, which offers a repeatable, hierarchical, and non-parametric cluster structure [11, 12]. Compared with K-means, it is easier to determine the number of clusters by cutting the dendrogram, which offers hierarchical and structured output.

The temporal travel habits extracted by LDA model reflect the homogeneity and heterogeneity of people's travel in the time dimension. Next, based on the Passenger-Temporal travel habit distribution we use HC method to cluster passengers.

Fig. 2 depicts an overview of our study framework. First, from individual passenger's history AFC data, we construct complete trips and retrieve the temporal trip record. Then, using the proposed hybrid topic-clustering method we extract the latent temporal travel habits and cluster passengers. Finally, we analyze and compare people's travel behavior from both spatial and temporal perspectives.

TABLE I. ANALOGY FROM PASSENGER-TEMPORAL TRAVEL HABITS TO DOCUMENT-TOPICS

Components of trip record		Components of text
The set of time slots $\mathbf{T}$	→	Vocabulary
Passenger l	→	Document
Temporal travel habit H_k	→	Topic
Time slot t_{ln}	→	Word

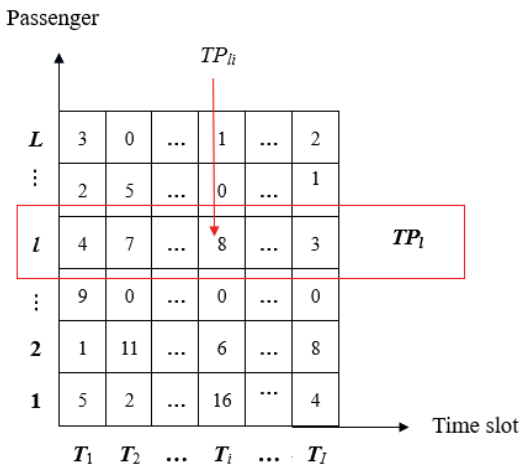


Figure 1. Passenger-Time slot matrix

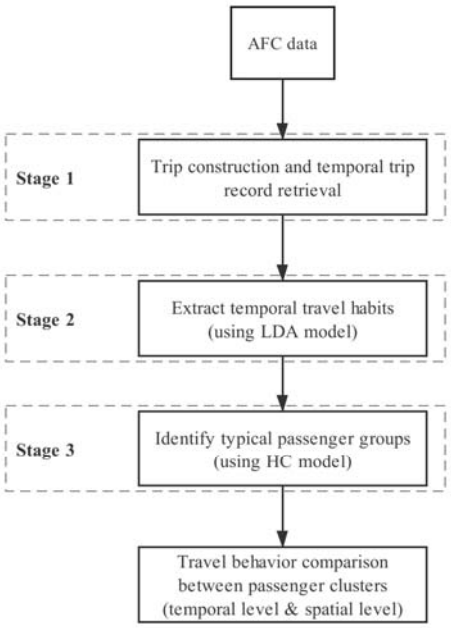


Figure 2. Analytical framework of travel pattern recognition using AFC data

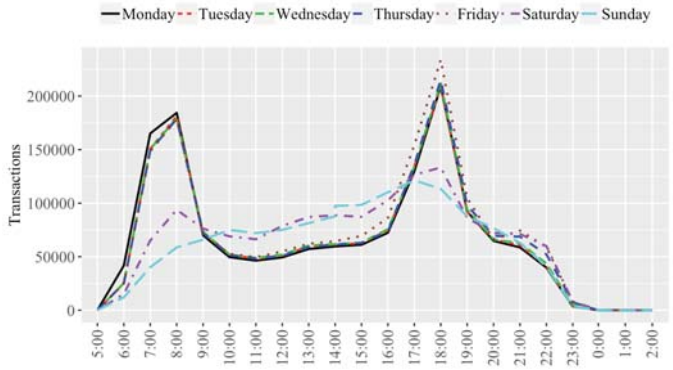
IV. DATA DESCRIPTION

Experiments of this study are conducted on real AFC dataset collected from Shenzhen Metro Cooperation in China. By 2013, the Shenzhen metro system operates 5 lines and 118 stations. The original dataset spans 20 weekdays from Oct 14, 2013 to Nov 8, 2013, with an average of more than 2 million people taking the metro every day. To improve the clustering result, in this work, we remove those passengers who take metro less than seven days, and then randomly select 50000 passengers as our study objects. The attributes and description of AFC transaction records are given in table 2.

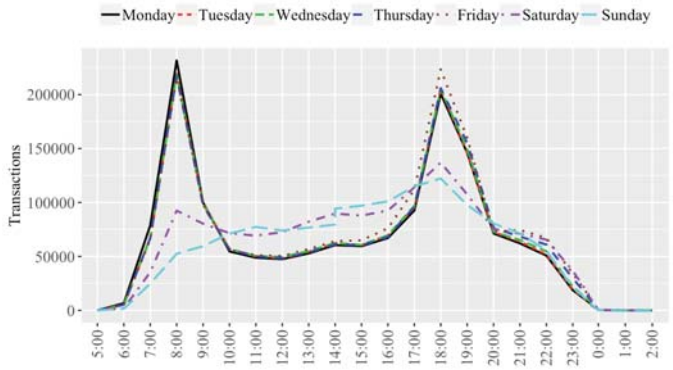
TABLE II. SMART CARD DATA FORMAT

Field	Description
Card_ID	Identifier of a smart card
Stat_ID	Metro station ID
Trans_Time	Transaction time
Trans_Type	21 represents tap-in a metro station, 22 represents tap-out a metro station.

The typical operation time of Shenzhen metro is from 06:30 to 23:00, but there are some travel records outside this period. To fully present people's movement in the city, we merge previous day with 05:00~23:59 and the next day with 00:00~02:59 (no transaction during 03:00~04:59), and thus convert the calendar day to logical day, i.e. from 5:00 am to next 02:59 am. Fig. 3 presents the total check-in counts and check-out counts per hour during weekdays. The passenger flow exhibits a double-humped shape over one day, both the counts of check-in and check-out have the morning peak between 07:00 and 09:00, and the evening peak between 17:00 and 19:00. This reflects inhabitants' commuting time and clearly shows that the general working hours in Shenzhen are from 08:00 to 19:00. Additionally, the trip survey in Shenzhen shows that more than 99% of the transfer activities are less than 40 minutes [1]. Based on the survey result, in this study, we use 40 minutes as the time threshold for identifying transfer activities.



(a) Check-in



(b) Check-out

Figure 3. Hourly transactions during five weekdays

V. RESULTS

A. Two distributions estimated by the LDA model

Fig. 4 shows the top ten most probable time slots of the two extracted temporal travel habits respectively. This visualization lets us understand the latent information extracted from AFC data. The most ten common time slots in temporal travel habit 1 contain all daytime hours from 08:00 to 17:59. This ten-hour time span suggests that it represents a travel pattern with flexible commuting time choice. Those most ten common time slots in temporal travel habit 2 are included into two blocks, 07:00 to 08:59 and 17:00 to 22:59, containing both morning peak and evening peak. The time slot distribution suggests that the second temporal travel habit represents a travel pattern with fixed and regular commuting time choice. Another important observation is that the same time slot has distinctly different probabilities in the two travel habits. For instance, the probability of time slot "08:00~08:59" in habit 1 is 0.038, while in habit 2 is 0.086.

Fig. 5 displays the distribution over two extracted temporal travel habits for ten passengers. As mentioned in section III, the Passenger-Temporal travel habit distribution quantifies passenger's preference in travel time selection. In Fig. 5, by comparing the probability of each travel habit for a certain passenger, we can get his degree of inclination towards different temporal move patterns. For example, passenger 3 and 7 are more likely to be in the second habit than the first, which means they are more inclined to take the metro during the morning and evening peaks.

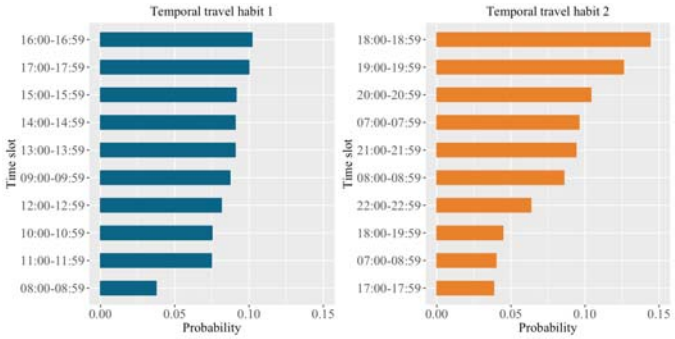


Figure 4. The Temporal travel habit-Time slot distribution

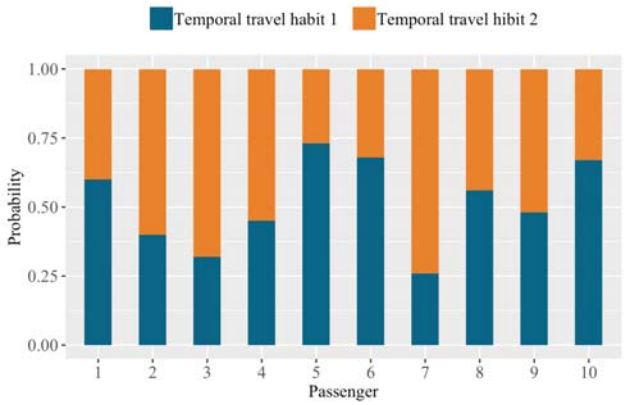


Figure 5. The Passenger-Temporal travel habit distribution

B. Passenger Groups Identified by HC Method

The cluster dendrogram is obtained by inputting the Passenger-Temporal travel habit probability matrix to the HC model, as shown in Fig. 6. According to the structure of dendrogram and within-cluster sum of square of different number of clusters, we divide passengers into four groups. Fig. 7 graphically maps the clustering result on the two temporal travel habits. The four clusters can be well separated into two chunks by the point (0.5,0.5), which is the demarcation point of the two extracted temporal travel habits. Compared with cluster 2 and 3, passengers in cluster 1 and 4 are more interested in the second temporal travel habit, but obviously, cluster 4 has the strongest preference for it. This means that after the rough partitioning by LDA, we implement further intra-group segmentation by hierarchical clustering. Fig. 8 shows that the four groups have 40.28%, 7.54%, 35.74%, 16.44% of the total passengers, respectively.

C. Travel Behavior Comparison on Temporal Level

Through statistics of different passenger groups' trip records, their temporal travel regularity could be inferred. Fig. 9 shows the travel frequency of four passenger groups over one logical day. The four passenger groups have notably different travel time selection: passengers in cluster 1 have three dominant time slots, 07:00~07:59, 08:00~08:59, and 18:00~18:59, which is the morning rush hours and evening rush hour in general. The morning peak proportion is 27.57%, the evening peak proportion is 23.55%, trips in peak hours occupy half of the daily travel demand. Passengers in cluster 2 have similar travel frequencies in each time slot between 09:00~16:59, 72.59% of trips occur during this time bucket. In cluster 3, the passenger volume is the largest in 08:00~08:59, 17:00~17:59 and 18:00~18:59, there are 43.76% of trips between 09:00 and 16:59. Although passengers in cluster 2 and 3 have flexible travel time (temporal travel habit 1), their travel patterns in time are opposite. Cluster 4 has a similar temporal travel pattern as cluster 1, but trips of cluster 4 reach the morning peak (17.83%) between 07:00~07:59, not the time slot 08:00~08:59 in cluster 1.

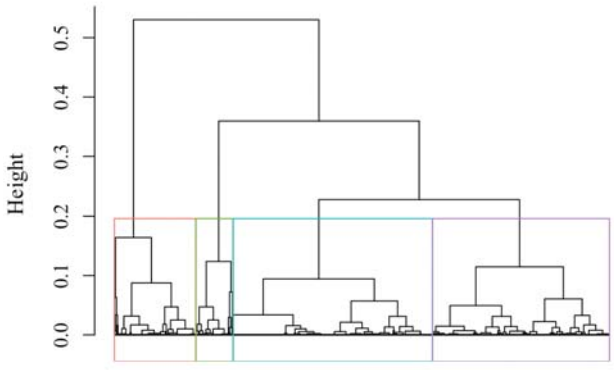


Figure 6. Cluster dendrogram

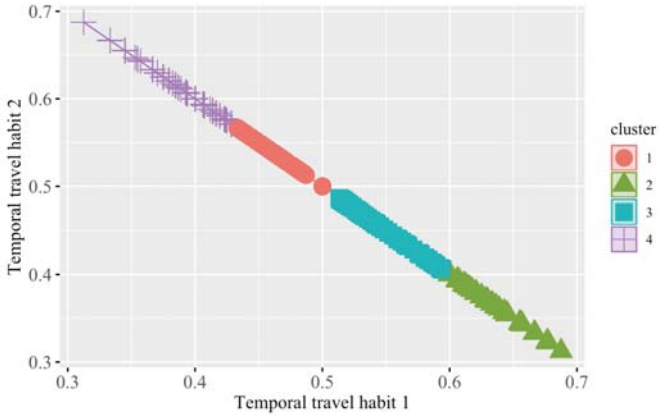


Figure 7. Passenger groups identified by LDA-HC method

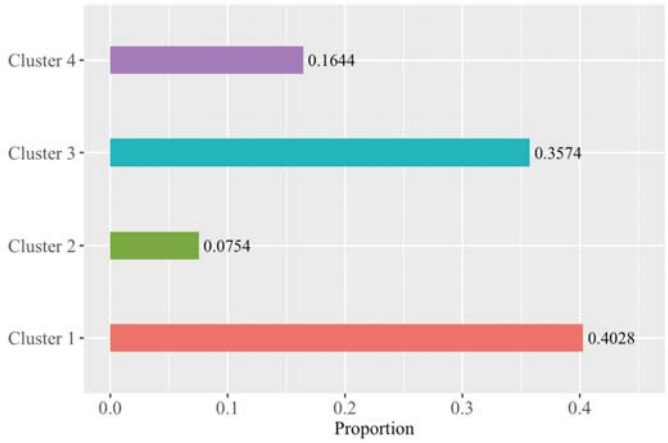


Figure 8. The proportion of each passenger group

In sum, from Fig. 9, cluster 1 and 4 contain passengers who regularly commute back and forth between home and workplace during the morning and evening peaks. Passengers in these two clusters can be viewed as "commuters" and constitute the majority (56.72%) of Shenzhen Metro users. On the contrary, cluster 2 and 3 may contain more "noncommuters", which have flexible temporal travel patterns. Fig. 3 only reflects the temporal travel patterns of commuters. However, cluster 2 and 3 have a significant proportion of trips during the day time, which do not fit the double-peak commuting pattern.

D. Travel Behavior Comparison on Spatial Level

Above comparison reveals metro users' differences in travel time selection. To investigate their mobility patterns in space, we further analyzed the spatial distribution of passengers.

Aggregating the check-in and check-out passenger flow, Fig. 10 lists the top 10 busiest metro stations for each cluster. With reference to government policy and regulatory documents, such as "Shenzhen City Master Plan" and "Shenzhen City Land Utilization Master Plan", in Fig. 10, we also give the functional attribute of the area where the metro station is located. For residential area, cluster 1 and 4 have 5 "residential" stations in different districts, including Pingzhou (Baoan district), Gangsha (Futian district), Baishizhou (Nanshan district), Huangbeiling (Luohu district), and Wuhe (Longgang district). But among the 10 stations of cluster 2 and 3, Gangsha is the only one located in residential area. This means that the passengers in cluster 2 and 3 have a higher consistency in the choice of residence place. For working area, passengers in cluster 3 focus on business zone and recreation zone. While passengers in cluster 4 have a noteworthy proportion of trips (9.4%) in the high-tech zone. The comparison suggests that metro users with various travel time also have different activity areas.

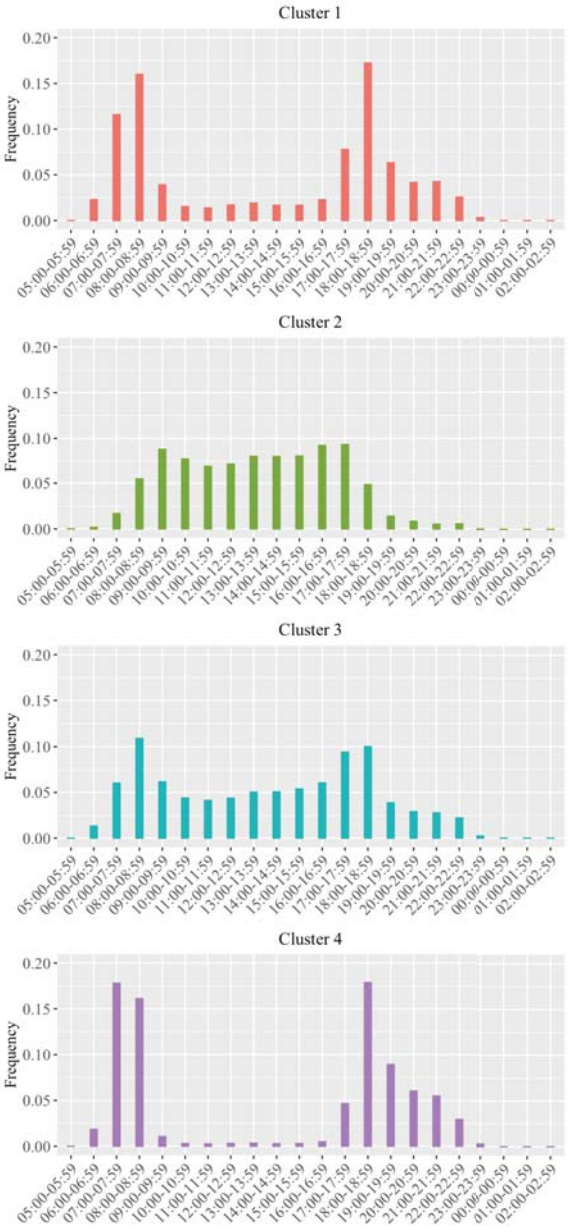


Figure 9. Travel frequency during each time slot

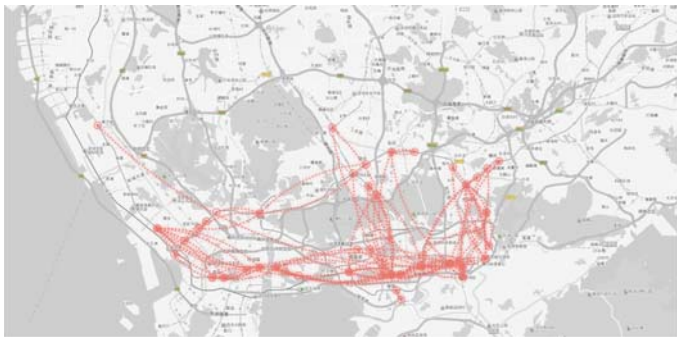
	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	Dajuyuan	Huaqianglu	Dajuyuan	Pingzhou
2	Pingzhou	Dajuyuan	Huaqianglu	Dajuyuan
3	Huizhanzhongxin	Chegongmiao	Laojie	Chegongmiao
4	Laojie	Gangsha	Gangsha	Huizhanzhongxin
5	Huaqianglu	Laojie	Huizhanzhongxin	Gaoxinyuan
6	Chegongmiao	Huizhanzhongxin	Chegongmiao	Wuhe
7	Gangsha	Huaqiangbei	Shijiezhichuang	Shenda
8	Baishizhou	Guomao	Guomao	Shenzhenbeizhan
9	Huangbeiling	Gouwugongyuan	Huaqiangbei	Laojie
10	Gaoxinyuan	Luohu	Gouwugongyuan	Baishizhou

Business area
Residential area
Recreation area
High tech zone
Border port
Railway station

Figure 10. Top 10 busiest metro stations of each cluster

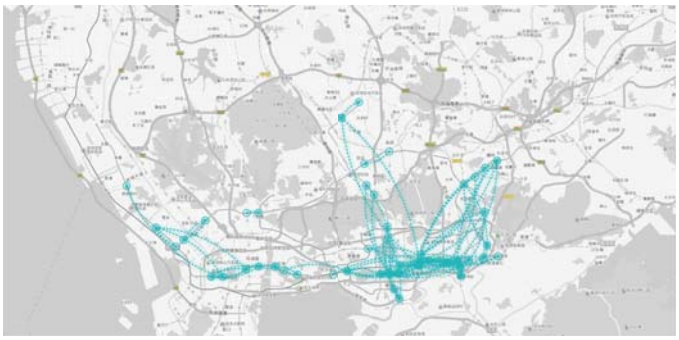
Through analysis of the trip proportion of stations, we can distinguish the residential area and working area, but we cannot tell the connections between metro stations. Thus, we derive an OD (Origin-Destination) matrix for each cluster. The top 100 frequent OD pairs are shown in Fig. 11.

On weekdays, the busiest OD pairs of cluster 1 are Pingzhou and Gaoxinyuan, Pingzhou and Taoyuan, Huangbeiling and Dajuyuan. For cluster 4, the largest connections are Pingzhou and Gaoxinyuan, Pingzhou and Shenzhen University, Baishilong and Huizhanzhongxin. These trips are mainly from residence place to high-tech zone and business zone. The center of trip generation is Pingzhou, Huangbeiling, and Baishilong, which is basically consistent with the residential center of cluster 1. The busiest OD pairs of cluster 2 are Gangsha and Huaqianglu, Gangshabei and Huaqiangbei, Luohu and Dajuyuan. For cluster 3, the largest connections are Gangsha and Huaqianglu, Guomao and Huaqainglu, Gouwugongyuan and Chegongmiao. Since more passengers in these two clusters take the recreation area as the destination, the flow direction is obviously different from cluster 1 and 4.

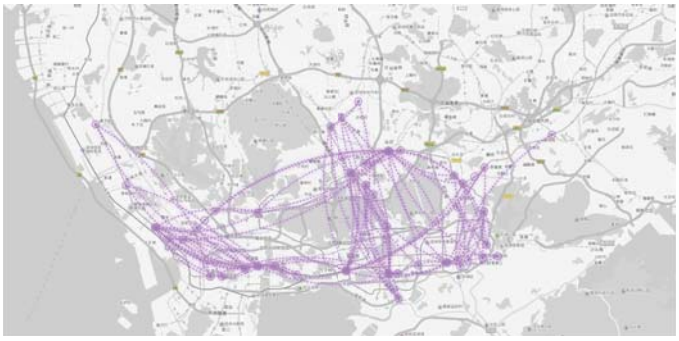


(a) The top 100 frequent OD pairs of cluster 1

(b) The top 100 frequent OD pairs of cluster 2



(c) The top 100 frequent OD pairs of cluster 3



(d) The top 100 frequent OD pairs of cluster 4

Figure 11. Connections between different stations in four clusters

VI. CONCLUSION

This study aims to identify inhabitant mobility patterns using AFC data. Concretely, in terms of the travel feature extraction and passenger groups partition, we proposed a hybrid topic-clustering method for this task.

Using twenty-weekday AFC data of Shenzhen metro system, we studied the temporal and spatial travel regularities of smart card users. Based on passengers' preference of travel time selection, four latent clusters are identified among metro riders. For temporal travel patterns, passengers in clusters 1 and 4 can be identified as "commuters" with regular commuting time. Cluster 2 and 3 can be considered as "noncommuters" having flexible travel time choices. The analysis of spatial travel patterns makes us identify the hot metro stops and the main activity areas of each passenger group. The comparison of travel behavior in both time and space indicated that metro users with different travel times also have significant differences in the choice of activity areas.

In the future, we will incorporate bus transaction records and travel routes into our study, which is helpful for comprehensive analyzing passenger traveling habits. We also plan to apply our method to spatial travel patterns and build a more elaborate system for passenger classification and anomaly detection.

ACKNOWLEDGMENT

We acknowledge all the participants in the study.

REFERENCES

[1] J. Zhao, Q. Qu, F. Zhang, C. Xu, and S. Liu, "Spatio-temporal analysis of passenger travel patterns in massive smart card data," *IEEE Transactions on Intelligent Transportation Systems,* vol. 18, no. 11, pp. 3135-3146, 2017.

[2] B. Agard, C. Morency, and M. Trépanier, "Mining public transport user behaviour from smart card data," *IFAC Proceedings Volumes,* vol. 39, no. 3, pp. 399-404, 2006.

[3] A. Bhaskar and E. Chung, "Passenger segmentation using smart card data," *IEEE Transactions on intelligent transportation systems,* vol. 16, no. 3, pp. 1537-1548, 2015.

[4] M. A. Ortega-Tong, "Classification of London's public transport users using smart card data," Massachusetts Institute of Technology, 2013.

[5] C. Seaborn, J. Attanucci, and N. H. Wilson, "Analyzing multimodal public transport journeys in London with smart card fare payment data," *Transportation research record,* vol. 2121, no. 1, pp. 55-62, 2009.

[6] N. Nassir, M. Hickman, and Z.-L. Ma, "Activity detection and transfer identification for public transit fare card data," *Transportation,* vol. 42, no. 4, pp. 683-705, 2015.

[7] N. Lathia, C. Smith, J. Froehlich, and L. Capra, "Individuals among commuters: Building personalised transport information services from fare collection systems," *Pervasive and Mobile Computing,* vol. 9, no. 5, pp. 643-664, 2013.

[8] R. N. Buliung, M. J. Roorda, and T. K. Remmel, "Exploring spatial variety in patterns of activity-travel behaviour: initial results from the Toronto Travel-Activity Panel Survey (TTAPS)," *Transportation,* vol. 35, no. 6, p. 697, 2008.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research,* vol. 3, no. Jan, pp. 993-1022, 2003.

[10] K. Gimpel, "Modeling topics," *Inform. Retrieval,* vol. 5, pp. 1-23, 2006.

[11] F. Murtagh, "A survey of recent advances in hierarchical clustering algorithms," *The computer journal,* vol. 26, no. 4, pp. 354-359, 1983.

[12] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.