Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 1733579, 14 pages
https://doi.org/10.1155/2021/1733579

Research Article

Exploring Passengers' Dependency Variety on Stations' Functions in Urban Subway

Xia Zhao [1], Pengpeng Jiao [1], Yong Zhang [2], and Chenjing Zhou [1]

[1] *Beijing Key Laboratory of General Aviation Technology, Beijing University of Civil Engineering and Architecture, Beijing 102616, China*
[2] *Beijing Key Laboratory of Multimedia and Intelligent Software Technology, Beijing Artificial Intelligence Institute, the Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China*

Correspondence should be addressed to Pengpeng Jiao; jiaopengpeng@bucea.edu.cn

Received 5 April 2021; Accepted 29 September 2021; Published 13 October 2021

Academic Editor: Giulio E. CANTARELLA

Urban subway is taken by people in different frequencies, thus leading them to present different dependency varieties on this mode. Yet, how those passengers who possess low dependency on urban subway travel is less investigated. Under this background, we propose a framework to uncover passengers' dependency variety on stations' functions in urban subway. To begin with, nine features regarding travel frequency and time are extracted from 100 million transaction records generated by 11.45 million passengers in a month. Thus, their travel dependency on urban subway is quantified. These features are clustered into 5 distinct levels via the k-means algorithm, before an inference of subway stations' functions from 236,040 POI data sources via the LDA approach. In this way, passengers' travel purposes can be identified. How passengers with different dependency levels behave in subway stations in space and time is further explored in a visualization way. The intuitive experimental results, validated by priori user experiences and land-use plan of Beijing, show that among the 5 levels of dependency varieties, passengers in the first two groups present a relatively strong dependency on urban subway. Meanwhile, passengers in the rest three groups possess a low dependency on urban subway and display extreme travel patterns in time and frequency, greatly increasing management difficulties for transit operators. Findings in this research help distinguish passengers with low levels of subway dependency and grasp how those passengers without striking dependency travel by subway and what for so that practitioners can conduct an accurate risk assessment on them.

1. Introduction

Urban subway is taken by individuals in megacities in different frequencies, which correlates to passengers' dependency levels on this transportation mode to some extent [1]. It is reported in one typical megacity Beijing that nearly 30% percent of subway riders in the year 2019 present high dependency on the urban subway by taking it in high frequencies to conduct routine activities like commuting. Still, there is another type of passengers who are less explored by academic fields. They take account for 10% of the whole population in Beijing each year and travel in a relatively low frequency by subway, which is an alternative transportation mode for them to reach dozens of points of interest (POI), e.g., transportation hubs, tourist sites, hospitals, or universities. Their identities are versatile. They may be the elderly, babysitters, or housekeepers, who seldom go out. They may be busy travellers who utilize other transportation modes, e.g., private vehicles, bicycles, or customized buses, a lot, but still spare their affections on urban subway. They also may be visitors from somewhere else and rush into the famous center of multiple functions for a short temporal period, to achieve their personalized motivations, e.g., holiday tours, business affairs, medical treatments, friend visiting, or further education [2]. No matter what, the common floating and random travel habits presented by those passengers may disturb existing travel orders formed by commuters in a sudden. For example, they may increase the population density of a station so that deadly human stampedes may be triggered. Therefore, it is necessary to grasp the knowledge that how do passengers without striking dependency on urban subway travel and what for so that practitioners can conduct an accurate risk assessment on this type. Their low dependency towards urban subway may increase the difficulty of solving the above questions; however, a prevalence of big smart card (SC) data recording passengers' digital footprints can relieve it.

Currently, previous studies have focused on categorizing passengers' mobility patterns from the aspects of spatial regularity [3] or spatiotemporal regularity [4,5]. The corresponding results may help identify whether passengers possess travel routines or not and provide theoretical supports for customized passenger management. However, they failed to throw more light on revealing passengers with low dependency on urban subway, thus causing a blurry understanding of how and why they visit a station. Yet, passengers' hidden travel motivations can be uncovered by exploring the function of a station by using POI data [6], passengers' mobility data [7], or both [8]. But until now, no connection has been built to connect this parameter with passengers' travel dependency [9,10], with an aim to figure out passengers' potential aims of travel. Apart from the above studies, some more research gave specific attention to characterizing mobility patterns for certain special groups, such as tourists [2], people who go shopping [11], the elderly [12], pickpockets [13,14], or passengers with extreme travel patterns [15]. However, they did not quantify their travel dependency on urban subway. Above all, more endeavours are demanded to distinguish passengers with less dependency on the urban subway and analyze how they behave in public transit in space and time and what they visit a station for. Solving the above problems may guide transit operators to design tailored management of specific passengers, e.g., the elderly, tourists, or pickpockets, and thus enhance a benign interaction between human mobility styles and mature land-use plan.

Under this background, we propose a comprehensive framework to uncover passengers' dependency varieties on subway stations with different functions. To begin with, we extract several features from over 100 million transaction records generated by 11.45 million passengers in a month, to quantify their travel dependency on urban subway. Then, we cluster them into 5 distinct levels, before an inference of subway stations' functions from 236,040 POI data sources via the LDA approach. How passengers in different dependency levels behave in subway stations in space and time is further explored and validated in a visualization way. Our main contributions are summarized as follows:

(1) Throw a deep light on passengers who exhibit less dependency on urban subway and distinguish them from dominant commuters via a clustering-based approach

(2) Build a connection between stations' functions and passengers' travel interest to uncover how passengers without striking dependency travel by subway and what for

For the remaining part, we summarize the related work on mobility pattern exploration in Section 2. The methodology to infer mobility patterns is interpreted in Section 3. The data sets are provided in Section 4. Experimental results and discussions are concluded in Section 5 before a brief conclusion drawn in Section 6.

2. Related Work

Recent research focused on characterizing normal passengers, for example, commuters, by their mobility patterns in the last decade. For example, Ma et al. [16] employed the first and last trips generated by passengers daily and characterized spatiotemporal travel regularity by quantifying the repeatability pattern of certain behaviours, such as frequently departing at a similar time interval or frequently visiting adjacent stops. Zhao et al. [17] compared the distribution differences of activity space generated by workers and nonworkers using big SC data lasting for 4 months. The results showed that passengers with travel routines are more likely to travel longer but cover less area than those without routines on weekdays, while opposite patterns occur on weekends. Huang et al. [18], Long et al. [19], Zhou et al. [20], and Zou [21] shared the point that commuters should have home and workplaces and proposed a series of rules to identify these two key spots before extraction of commuters from SC data. Obviously, commuting, a dominant regular pattern generated by a majority of commuters who travel in public transit, has been explored widely by recent studies. However, patterns derived from these studies cannot be applied directly to those passengers who display little dependency towards urban subway.

Some other studies threw light on passengers who present a wide variety of heterogeneous mobility patterns based on travel features regarding temporal, spatial, or attributive factors. For instance, Lee and Hickman used SC data to examine the spatiotemporal symmetry of passenger flows along a transit route on a single day. The amounts of boarding and alighting passengers in one station in both directions were statistically compared. The research showed that significant variation existed between specific time periods and spatial environments [22]. Zhao explored passengers' hidden relation proximity by measuring their similarities on travel profiles, interaction activities, and neighbourhood-driven connections by leveraging SC data [23]. Jiang extracted 400 million trips from SC data in Seoul city and identified trip chains and transit modes via spatial distribution of travel time [12]. Considering the inability to capture dynamic human mobility from solely a single view of data, Zhang et al. [24] built a multiview learning framework to sketch passengers' digital footprints. The multisource data adopted were cellphone, SC, and taxi data. The results showed that the integrated model with multiple views better described human mobility patterns. Although the above researches explored human mobility patterns based on either spatial or temporal factors, they have ignored the fact that different groups of people may share different transit behaviours.

Under the above background, some studies separated passengers into various groups to study heterogeneous mobility patterns. For instance, Lee and Hickman analyzed historical trip records on subway card data and generated heuristic rules in terms of spatiotemporal consistency to derive hidden transit behaviours following regular forms. The research found that passengers had regular activity patterns, such as commuting from work/school to home. Furthermore, this research partitioned passengers into work-related patterns, school-related patterns, and others [25]. Ma et al. [4] proposed a comprehensive framework to mine transit mobility patterns. The research firstly identified trip chains from the SC database and used a clustering-based approach to extract the most frequent transit patterns of every passenger in terms of spatial and temporal regularity. Ultimately, the research classified passengers into five types based on regularity levels, such as passengers with extremely high or low transit regularity. Like Ma's research, Kieu et al. [5] also reconstructed multiday travel itineraries from SC transactions and analyzed the transit regularity of every passenger in terms of spatial and temporal aspects. But different from Ma, the authors partitioned passengers into four types based on spatiotemporal regularity, such as passengers with or without spatiotemporal regularity. Louail et al. [3] analyzed mobile phone data in 31 cities during workdays in Spain and extracted an origin-destination matrix of passenger flows to describe the commuting mobility of each individual from residential hotspots to work hotspots. The research characterized four categories of commuting flows in terms of spatial regularity, then classified the flows into several categories. Cui et al. [15] proposed an extreme index (EI) to evaluate the extreme level of transit behaviours on travel time and travel frequency. The research calculated the values of EI of every bus rider and characterized four types of extreme travel patterns based on time or frequency irregularity. But the research did not take a space-relevant index into account. Above all, the above research relied on employing SC to segment passengers into various subgroups. However, they merely concentrated on those travel patterns with dominant regularities. With no knowledge of the specific travel demands of passengers, research still cannot figure out why passengers visit a station and how would they behave if they do not present a strong dependency on public transit.

To solve the above issue, a clear identification of a station's function may play a critical role in solving the above issue. Recognizing the function of a station is a fundamental factor to evaluate travel demand. Previous studies have shown that this index can be mined from multiday transit records or geo-tagged POIs. For instance, Yuan et al. [8] extracted semantic data on geo-tagged POIs to infer the function of a region, before clustering them into distinct subgroups to partition the functions of regions. The research indicated that the function of a region function could be represented by the distribution of mobility patterns and ultimately extracted nine types of functions for all regions, such as areas related to commerce, entertainment, education, science, residence, and so on. Like Yuan's research, Kim et al. [26] extracted both the diurnal transit data and the geospatial subway stations in Seoul city and clustered these stations in terms of statistical travel patterns. The research derived five clusters of stations related to five different types of functions, that is, residence, commerce, tour, nature scenes, and leisure. Moreover, the research showed that the distribution of functions revealed passengers' boarding or alighting travel patterns in subway stations. But until now, few research have combined this factor with others to investigate passengers' mobility patterns.

To sum up, previous studies have tried to partition mobility patterns of passengers based on their regularities in space, time, or both, to assist the current transportation management of a majority of passengers. However, these studies did not examine those passengers who express a low dependency on urban subway, thus failing to explain why they visit a station. Uncovering the function of a subway station is of great importance to demonstrate their potential travel orientations, yet few research have combined this specific index into mobility discovery. This results in a further blurry examination of passengers' mobility patterns without dominant travel patterns in public transit. Thus, it is urgent to distinguish those passengers without striking dependency from commuters and uncover how they travel by subway and what for.

3. Methodology

A three-step framework is developed to identify passengers' dependency variety on the urban subway. To begin with, nine features regarding travel frequency and time are extracted to quantify passengers' travel dependency on the urban subway, before clustered to derive different levels of travel dependency. Next, subway stations' functions are inferred to help identify passengers' general purposes of visiting stations. Finally, the derived results are validated based on priori user experiences and the land-use plan of Beijing.

3.1. Clustering Passengers by Their Dependency Variety on Urban Subway

3.1.1. Quantifying Passengers' Dependency Variety. Considering commuters in public transit have been found to present stronger dependency on urban subway than those do not [4], frequency- and time-related features are adopted here to distinguish passengers with a variety of dependency levels.

Frequency-Related Features. Travel frequency can be quantified by two fined granularities, that is, travel times and travel days. With travel times, we can easily grasp how frequently passengers take subway; while with travel days, we can infer how the above frequency distributes over days. Thus, six specific features regarding this are extracted, with details given below.

Travel Times. Travel times are widely adopted by researchers, to name a few, [4,14,23,27], to capture passengers' travel dependency varieties. Normally, the higher passengers' travel times are, the more frequently they are used to take subway, accordingly indicating a higher dependency on this transportation mode. With this regard, three statistical values, that is, the max, the mean, and the min daily travel times, are particularly specified to capture any extreme changes behind dependency varieties every day. Moreover, the monthly travel times are also calculated in order to gain an overall picture of passengers' travel frequency within the research period, that is, a month.

Fixed Travelling Days. The number of fixed travelling days when passenger generates regular trips are employed by investigators like Ma et al. [16], Huang et al. [18], and Zhao et al. [17]; to quantify passengers' travel regularities. The criteria to examine whether a passenger has regular trips or not lie in the following conditions: (i) sharing similar first-boarding time and stations in multiple days, (ii) sharing similar last-alighting time and stations in multiple days, and (iii) sharing similar route sequences in multiple days. The higher fixed days passengers travel by subway, the higher regularities and the stronger dependency they possess. As validated by previous studies, this index performs well in differentiating commuters, a type of passengers who own strong stickness to urban subway, from other types. With this regard, we extract its two statistical values, that is, the max and the min monthly fixed days, to capture any extreme changes behind dependency varieties within the research period.

Time-Related Features. Travel time is used by scholars like Jang [10], Xue et al. [2], Yildirimoglu et al. [28], and Dharmowijoyo et al. [29] to measure how long passengers last in a transit trip. Normally, the longer time a passenger takes in a trip, the stronger dependency he/she sticks to urban subway. Particularly, three statistical values, that is, the max, the mean, and the min unit travel time, are particularly specified to capture the unit duration time of a single trip of a passenger generated within the research period.

3.1.2. Aggregating Passengers' Dependency Variety. With the extraction of the above nine features, passengers' dependency varieties on urban subway are characterized. Passengers with various levels of dependency on urban subway are then aggregated into K cluster via the k-means algorithm, which is an unsupervised clustering algorithm that partitions similar data into one class efficiently and automatically [30].

Set N_p as the total number of passengers in this paper and $\varnothing$ as the total squared distance of each point to its closest center. The input of the algorithm is the nine features of $\mathbf{N_p}$ passengers, which is quantized by a $\mathbf{N_p} \times 9$ matrix labelled as $\{f_i\}_{i=1}^{N_p}$. The entry f_i is a normalized 9-dimensional feature vector, $f_i = \{f_{i,k}\}_{k=1}^{9}$ $(i \in 1, 2, \ldots, \mathbf{N_p})$. The initial centers of the K clusters are randomly chosen and denoted as $C = \{c^1, c^2, \ldots, c^i, \ldots, c^K\}$. The cluster number K representing the levels of subway dependency is determined by a so-called index sum of the squared error (SSE), which measures the total squared distance distortion in a cluster. The mathematical representation of SSE is given in the following formula, where $\mathrm{dist}(c^j, f_i)$ indicates the distance of a feature vector f_i to a cluster center c^j. The smaller the SSE is, the better clustering performance occurs in one cluster. An optimum cluster number K is finalized when SSE reaches its minimum value.

$$\mathrm{SSE} = \sum_{j=1}^{k_2} \sum_{f_i \varepsilon c^j} \mathrm{dist}^2\left(f_i, c^j\right). \tag{1}$$

3.2. Identifying the Representative Function of a Station. After aggregation of passengers' dependency variety on the urban subway, the functions of all subway stations are further inferred, with an aim to understand what passengers in different dependency levels on urban subway visit a station for.

Previous studies showed that the function of a station strongly corresponds to the distribution of POIs around the location [8], thus attracting individuals to visit for relevant social activities. In this way, a passenger's general purpose of visiting a station can be deduced. For example, a station surrounded by one single category of POIs, for example, plenty of office buildings, probably attracts passengers to come forward for business-oriented activities. A station can be also surrounded by a quantity of POIs in different categories, for example, residential buildings, restaurants, gyms, and exhibits multiple types of functions. In this circumstance, we follow Yuan et al. [8] to merely focus on the representative function of the station, which addresses a high correlation to a typical category of POIs owning the highest frequency density. Above all, the first step is to identify the function of a subway station, which is deduced from POIs via a classical topic-inference-based approach named the latent Dirichlet allocation (LDA) proposed by Blei et al. [31].

The logical network of the LDA model, as shown in Figure 1, is indeed a multilayer Bayesian model. In this model, a mixture of observable words, denoted as $\mathbf{W}$ and colored black in Figure 2, support a latent topic with certain probabilities, while a mixture of topics, labelled as $\mathbf{Z}$, comprise a document. With all the words in each document known in advance, the LDA model is trained to infer a hidden topic of the document. With an analogy of this principle, an identification of the hidden function of a station proceeds by training the LDA model based on antecedent extraction of all observable POIs located in a 1 km radius around the station. This circular region is also known as a walkable region for individuals to reach [32]. Define $\mathbf{S}$ and $\mathbf{K}$ as a set of stations and their functions beforehand, then a station $s \in \mathbf{S}$ is regarded as a document, with its function $\mathbf{k} \in \mathbf{K}$ regarded as a topic. Its POI category $\mathbf{t}$ is then treated as a term, with one specific POI $\mathbf{n}$ as a word of a document. Moreover, set $\mathbf{w_{s,n}}$ as the $\mathbf{n}_{th}$ observable POI of the station s, and $\mathbf{z_{s,n}}$ as its $\mathbf{n}_{th}$ latent function. Based on these annotations, the POIs and the function of a station are then generated iteratively for $\mathbf{S} \times \mathbf{K}$ times with one loop introduced as follows:

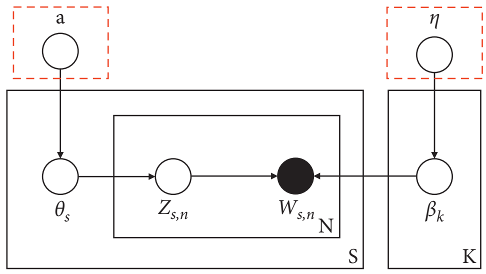


FIGURE 1: The Bayesian networks of the LDA.

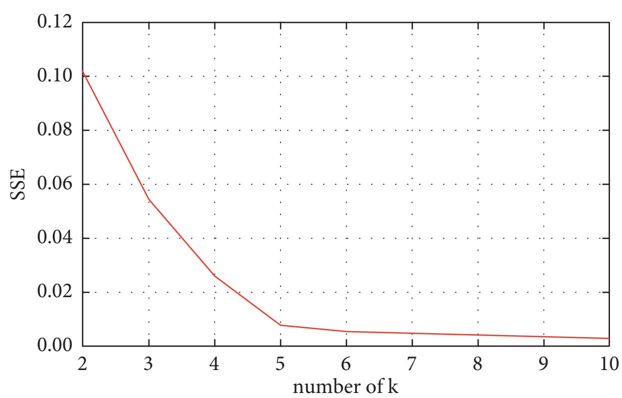


FIGURE 2: Determining an optimum cluster number of dependencies based on the SSE curve.

Each latent function $z_{s,n}$, together with each POI word $w_{s,n}$ generated by $z_{s,n}$, is regarded as following a multinomial distribution with parameters $\beta_{z_{s,n}}$ and $\theta_{\mathbf{z_{s,n}}}$ known beforehand. They are mathematically represented in formulae (2) and (3), respectively, where Mult $(\cdot)$ refers to the Multinomial distribution. Moreover, both β_k and $\theta_{\mathbf{s}}$ follow a Dirichlet distribution with parameters η and α known ahead, respectively. They are modeled by formulae (4) to (5) each, where Dir $(\cdot)$ indicates the Dirichlet distribution. The core issue of topic inference is to predict the posterior distribution $P(\theta, z, \beta_k | w, \alpha, \eta)$, which is specifically interpreted by a variety of methods, such as Gibbs sampling, with details illustrated by Blei et al. [31]. Experimental results conducted in this paper testified that the LDA model yields good performances when $\alpha = 50/K$ and $\beta = 0.01$.

$$w_{s,n} \sim \text{Mult}\left(\beta_{z_{s,n}}\right), \tag{2}$$

$$z_{s,n} \sim \text{Mult}\left(\theta_{z_{s,n}}\right), \tag{3}$$

$$\beta_k \sim \text{Dir}\,(\eta), \tag{4}$$

$$\theta_s \sim \text{Dir}\,(\alpha). \tag{5}$$

After the aforementioned function inference, stations with similar functions are also aggregated into one cluster via the k-means algorithm. The input of the algorithm is the $\mathbf{K}$ functions under $\mathbf{S}$ station zones, which is quantized by an $\mathbf{S} \times \mathbf{K}$ matrix labelled as $\{z_s\}_{s=1}^{S}$. The entry z_s is a normalized K-dimensional function vector, and $z_s = \{z_{s,k}\}_{k=1}^{K}$ $(s \in 1, 2, \ldots, S)$. Other details are similar to the clustering procedure given in Section 3.1. The difference is that we choose an index perplexity to better evaluate the optimum function categories of stations. This indicator is represented by the average likelihood of every station zone shown in formula (6), where N_s is the number of POIs in the $\mathbf{s_{th}}$ station. Normally, lower perplexity suggests better clustering results.

$$\text{Perplexity} = \exp\left\{-\frac{\sum_{s=1}^{S}\sum_{n=1}^{N_s}\ln P\left(w_{s,n}\right)}{\sum_{s=1}^{S} N_s}\right\}. \tag{6}$$

Based on the aggregated function categories of subway stations, the most representative function of a station is further annotated according to the distribution of frequency densities each POI category exhibits. The higher the frequency density of a POI category is, the higher possibility it becomes representative. In this paper, the frequency density (FD) of the i_{th} POI category in the station j with area R_j is mathematically represented in formula (7), where N_i refers to the number of POIs of the i_{th} category in a station. The representative POI category with the highest density is considered as the dominant function of a station.

$$Fd_i = \frac{N_i}{R_j}. \tag{7}$$

3.3. Validating the Derived Results. We firstly validate that whether the derived levels of passengers' travel dependency match ground truth data. In particular, we calculate the respective ratio distributions of passengers in each estimated dependency level and compare them with the associated figures summarized in the official report in Beijing, 2015 [33]. The implemented broad match can reflect the differences of ratio distributions generated theoretically and empirically in a macroscopic view.

Secondly, we invite 30 local people (permanent residents in Beijing) to manually annotate the representative function of each subway station based on their existing knowledge or life experiences, with an aim to validate whether the inferred function categories match ground truth data. Differences are particularly examined whether the stations labelled with the same functions by the LDA model still have the same (or different) functions in the opinions of the subjects.

4. Data Sets

Two kinds of digital footprints, namely, POI data and SC data, are adopted in this paper to identify passengers' dependency variety on the urban subway. An overview is given in the following.

4.1. Subway Card Data on Human Mobility. Subway card data, one kind of ubiquitous spatiotemporal data collected from the public transportation system, shared almost 40% of the overall passenger flows in Beijing, 2015. It can be extensively used to explore that how often people come to and leave a station [33,34]. The data used in this paper contain over 100 million transactions of 11,450,928 passengers from June 1 to 30, 2015. Each data record contains attributes such as smart card data IDs, entry/exit stations, and time stamps at entry/exit stations. Specific data preprocessing work has been interpreted in one of our previous work [35].

4.2. POI Data Set on Location Semantics. POI data set presents location-based semantics, such as residential buildings, dining places, tourist sites, or entertainment areas. It plays an essential role in discovering latent functions of a region [8] and can reflect passengers' rough travel demands concerning potential functions around a station. The data set generated in Beijing, 2015, contains a digital subway network owning 18 subway lines and 334 subway stations, as well as 236,040 POIs, with each POI having its latitude, longitude, and the category. The raw data are input into the LDA for function inference.

5. Results and Discussion

The derived dependency variety possessed by passengers on urban subway is firstly analyzed, followed by an exploration of the derived function categories of subway station, whose reliabilities are validated as well. Results derived in the above two sections are further integrated to dig out passengers' general purpose of visiting a subway station based on an intuitive characterization of mutual connections in space and time.

All experiments are conducted on a 64 bit server with an Intel i7 Core 3.5 G CPU and 24 GB RAM. Programs are implemented by Python 3.7.5 and Jupyter notebook. Visualization tools are adopted such as HTML5 or CSS. Gaode Map of China is used to present the visualization results.

5.1. Analyzing Passengers' Dependency Variety on Urban Subway. The clustering results of passengers' travel dependency are depicted in Figure 2. An optimum cluster number occurs when SSE reaches its minimum value. It indicates that passengers can be clustered into five levels of dependency with the strength ranging from "extremely strong" to "not at all." The derived five clusters are labelled from *c0* to *c4* for simplicity.

The respective proportion distribution of passengers in the above five clusters is displayed in Figure 3, with values ranging from 2.40% to 48.17%. The flow population in clusters *c0* and *c1* are significantly higher than the rest three kinds of clusters. In this case, the first two clusters can be regarded as having a relatively strong dependency on the urban while the rest three clusters do not.

5.2. Analyzing the Representative Function of a Station. By relying on the descriptions given above, we still cannot make an estimation on what passengers visit a station for. Thus, a further analysis of the inferred functions of subway stations is elaborated.

To begin with, the optimum function categories is finalized based on the perplexity distribution drawn in Figure 4. The LDA model converges at the 150th iteration, under which circumstance, the variation trend of perplexity is exhibited with function categories ranging from 2 to 10. The smallest perplexity emerges when the topic number is 8, suggesting that the optimum function category can be set as 8, with each category labelled from *s0* to *s7*, respectively.

The derived function categories of subway stations are further annotated based on their distributions in frequency density (FD) and space. As shown in Figure 5(a), all the function categories are highlighted by their frequency densities, with a darker part indicating a higher density. A ranking is also derived inside each category based on its internal frequency density values, which is therefore short for interval ranking (IR) for simplicity. Besides, Figure 5(b) depicts the spatial distribution of the eight function categories, distinguished by colors, for all subway stations in Beijing, with illustrations given as follows:

(1) Stations with airport-relevant function (*s*0): 2 subway stations are included that are near airport terminals in Beijing. Representative POIs are founded, such as international travel agencies, shuttle bus terminals, and express lines, together with some other auxiliary POIs, such as convenience stores, gas stations, or ATM banking.

(2) Stations with government-relevant function (*s*1): 26 subway stations located within the 2nd Ring Road of Beijing are included. They are surrounded by a quantity of representative POIs such as government agencies and embassies.

(3) Stations with residence-relevant function (*s*2): 55 subway stations, which are distributed outside the 4th Ring Road of Beijing, appear in this category and are featured with residential buildings around. Auxiliary POIs are founded in high densities as well, such as restaurants, sports centers, convenience stores, and supermarkets, to support residents' living.

(4) Stations with bus terminal function (*s*3): 9 subway stations exist in this cluster and are surrounded by 31 bus terminal stations in various scales.

(5) Stations with scenic-relevant function (*s*4): 16 subway stations fall into this category, featured with famous historical sites, to name one, the Forbidden City. Tourism-relevant services, e.g., restaurants, hotels, and souvenir shops, are provided as well.

(6) Stations with commerce-relevant function (*s*5): 24 subway stations occur in the category, where emerge a certain number of business districts, e.g., CBD and Chinese Silicon Valley. Auxiliary facilities are built as well to meet the potential demands of workers, such as restaurants, banking services, shopping malls, and gyms.

(7) Stations with education/health-relevant function (*s*6): 86 subway stations are contained in this cluster and are located within the 2nd to 4th Ring Roads of Beijing. Educational or medical POIs are dominant

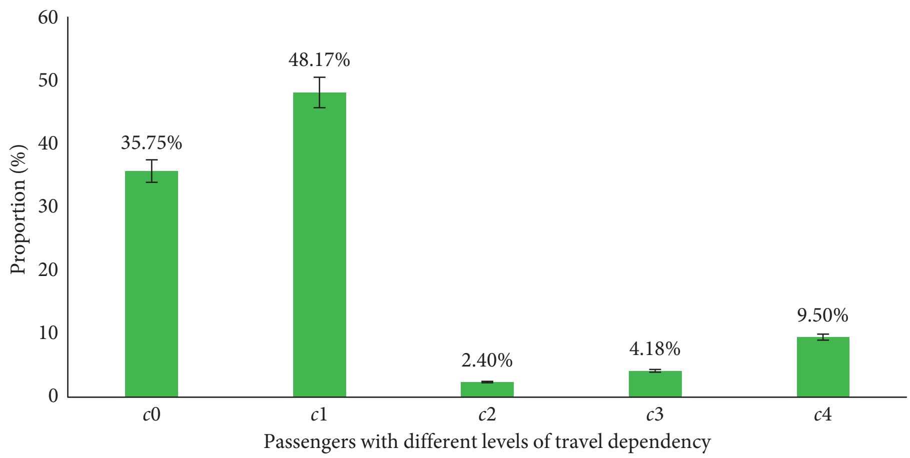


FIGURE 3: The overall proportion distribution of passengers in different dependency levels.

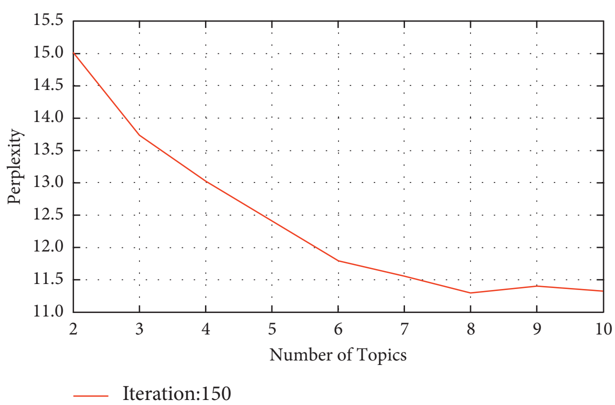


FIGURE 4: Determining an optimum function category based on the perplexity curve.

and present a high yet uneven density, supplemented with several secondary facilities, such as banking service, parking lots, restaurants, or convenience stores.

(8) Stations with railway terminal function ($s7$): 16 subway stations exist in this cluster, with most of them regarded as key transportation hubs, such as Beijing West Railway Station.

5.3. Validating the Derived Results

5.3.1. Validating the Derived Dependency Variety Possessed by Passengers. The respective ratio distributions of passengers in the five dependency levels, together with the priori figures given in the official report, are both shown in Figure 6 for comparison.

The government has already concluded three kinds of mobility patterns, i.e., life-related travel, fixed travel, and others, according to passengers' travel behaviours in the Beijing subway. The first type given in the official report accounts for 37.12% among all and links to elastic activities such as shopping, dining, friend visiting, or exercise. This type has a medium dependency on urban subway and presents random regularity in travel frequency, duration time, and frequently visited stations. This type is in line with the cluster $c0$ derived in this paper concerning ratio distribution, where passengers account for 35.75% among all and present a similar elastic travel regularity among their records.

The second type summarized in the official report accounts for 44.25% of all the trips, which includes regular activities such as commuting between work places (or school) and home. This type has a high level of dependency on urban subway by presenting regularity in travel frequency, duration time, and frequently visited stations. These patterns are almost in line with the patterns shown in cluster $c1$ derived in this paper, which accounts for 48.17% among all passengers.

The official third type, named "others," accounts for 18.63% among all in the government report without specifically figuring out who generates these travels. But this proportion is approximate to the accumulated proportions, that is, 16.08% in total, of the rest clusters labelled from $c2$ to $c4$ in this paper. The respective proportion of clusters $c2$, $c3$, and $c4$ is 2.4%, 4.18%, and 9.5%, respectively, as shown in Figure 3. This indicates that we not only discern the existing first two types of mobility patterns clarified in the official report by their ratio distributions but also explain what is the specific meaning of the ambiguous word "others" given in the official report. The specific travel dependency variety of these ambiguous patterns are further interpreted in Section 5.4 so as to fill the gap of discovering nondominant mobility patterns of passengers who present less dependency on urban subway with persuasive evidence.

5.3.2. Validating the Inferred Function Categories of Subway Stations. Among the 30 participants mentioned in Section 3.3, 26 (88.67%) of them agree that a good match is obtained in annotating the function categories of subway stations, either based on the results derived in this paper or based on their priori knowledge. We further compare the spatial distribution of the derived functional zones against the overall land-use plan of Beijing (2004–2020) [33], to get a

	s0		s1		s2		s3		s4		s5		s6		s7	
POIs	FD	IR	FD	IR	FD	IR	FD	IR	FD	IR	FD	IR	FD	IR	FD	IR
Dinning	0.000010	6	0.050969	6	0.034287	11	0.052501	5	0.062110	3	0.096774	4	0.085135	6	0.039683	10
Lottery	0.000010	6	0.076208	4	0.058823	7	0.032715	8	0.031016	5	0.000384	13	0.099671	3	0.066313	4
Plaza	0.000010	6	0.000005	14	0.000006	14	0.037387	6	0.000009	15	0.376772	1	0.000003	13	0.000009	16
Shopping	0.000010	6	0.056656	5	0.057310	9	0.057997	3	0.031365	4	0.036125	6	0.093957	4	0.047654	8
Transport	0.000010	6	0.023781	10	0.089463	4	0.022822	12	0.023679	9	0.012225	11	0.036341	12	0.052636	7
Gas station	0.001604	3	0.000005	14	0.123577	3	0.004960	16	0.000009	15	0.000003	15	0.000003	13	0.010788	14
Finance	0.001803	2	0.033747	9	0.013113	12	0.035600	7	0.004900	13	0.144221	3	0.080790	8	0.053632	6
Education	0.000209	4	0.020369	12	0.051989	10	0.015265	14	0.018438	10	0.019248	9	0.126556	1	0.033976	13
Vehicles	0.000110	5	0.000005	14	0.241603	1	0.017876	13	0.000009		0.000003	15	0.000003	13	0.000190	15
Others	0.000010	6	0.040950	8	0.061847	6	0.023097	11	0.028133	6	0.017778	10	0.065254	10	0.035154	12
Parking Lot	0.000010	6	0.021019	11	0.080725	5	0.053737	4	0.008830	11	0.185107	2	0.522088	11	0.038234	11
Healthcare	0.000010	6	0.048857	7	0.058374	8	0.027769	10	0.008656	12	0.006754	12	0.092456	5	0.063868	5
Leisure	0.000010	6	0.019070	13	0.128787	2	0.032028	9	0.024814	8	0.033757	7	0.082765	7	0.046205	9
Administrator	0.000010	6	0.100417	3	0.000006	14	0.014029	15	0.025251	7	0.030491	8	0.079000	9	0.070027	3
Residence	0.000010	6	0.115744	2	0.000062	13	0.084240	2	0.085868	2	0.040317	5	0.105964	2	0.073922	2
Tourism	0.000010	6	0.000005	14	0.000006	14	0.000014	17	0.646790	1	0.000003	15	0.000003	13	0.000009	16
Police	0.000010	6	0.392174	1	0.000006	14	0.000014	17	0.000009	15	0.000030	14	0.000003	13	0.000009	16
Railway	0.000010	6	0.000005	14	0.000006	14	0.000014	17	0.000096	14	0.000003	15	0.000003	13	0.367672	1
Bus stop	0.000010	6	0.000005	14	0.000006	14	0.487923	1	0.000009	15	0.000003	15	0.000003	13	0.000009	16
Airport	0.996125	1	0.000005	14	0.000006	14	0.000014	17	0.000009	15	0.000003	15	0.000003	13	0.000009	16

(a)

(b)

Figure 5: Function distribution of subway station zones: (a) distribution of parameters FD and IR of POI categories and (b) function distribution in the digital subway map of Beijing.

broader match. For instance, many governmental or municipal departments, configured as located in the downtown city in the derived category $s1$, are actually located within the 2nd Ring Road of Beijing, while hundreds of residence buildings, featured as appearing outside the 4th Ring Road of Beijing in the derived category $s2$, are actually located in suburbs. Similar matches are also found for the rest function zones.

Additionally, the amount of passenger flows located in each function category is further quantified to validate whether the estimated distribution of passenger flows matches the priori distribution summarized by the office report in Beijing, 2015 [33]. To be specific, we calculate the flow proportion P_c^k in the c-th pattern concerning the k-th function category based on formula (8), with results demonstrated in Figure 6.

$$P_c^k = \frac{\sum_{s=1}^{S_k} f_c^k}{\sum_{c=0}^{C-1} \sum_{k=0}^{K-1} \sum_{s=1}^{S_k} f_c^k}, \tag{8}$$

where we have the following:

(1) f_c^k: the exit flow of the c-th dependency level concerning the k-th function category

(2) S_k: the number of stations corresponding to the k-th function category

(3) C: the number of passengers' dependency levels on urban subway

(4) K: the number of function categories of subway stations

As is shown in Figure 7, passengers with flows ranking top three have frequently visited stations with categories labelled as s6, $s5$, or $s2$, which provide functions relevant to education, health, residence, and commerce. This matches the actual status of labour mobility in Beijing that hotspot POIs, for example, residence buildings, office buildings, hospitals, and schools, always attract a huge amount of people, especially labour force in cluster $c1$, to visit. Moreover, no significant difference exists in the functionality distribution in each dependency level. It is probably because the land-use plan of Beijing (2004–2020) expects to promote a balanced, intensive, and polycentric development among multiple regions [19]. Thus, a resident who has a travel-out plan can easily find a nearby region to visit with conducting his/her desired activities.

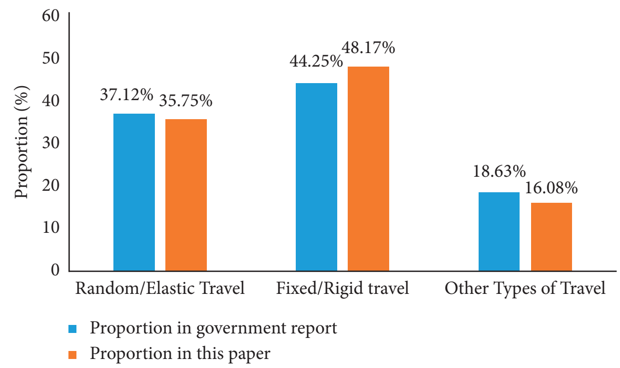


Figure 6: A comparison of ratio distributions of mobility patterns derived from the official report and from this paper.

5.4. Characterizing Passengers' General Purpose of Visiting a Subway Station. After validation of the above results, how passengers in different dependency levels behave in subway stations in space and time is further explored in a visualization way. The aim is to characterize passengers' general purposes of visiting a subway station, thus facilitating transit operators to build a mutual connection between stations' functions and passengers' travel interests.

Figure 8 demonstrates how frequent and how long passengers with different dependency varieties travel on urban subway. Moreover, Figure 9 presents the spatial distribution of the top 15 subway stations visited by passengers with different dependency varieties. The size of a circle represents the amount of passenger flows, while its color represents flow types of passengers. Circles in pink, yellow, and green colors reflect entry flows, exit flows, and flow differences between them, respectively. Finally, Figure 10 draws an example of how passengers in different groups distribute spatiotemporally on Tuesday, June 4, 2013. The x-axis is the time series ranging from 4 am to 11 pm with 1 hour's interval. The y-axis is the proposed station function ranging from $s0$ to $s7$. It is easy to find that for every pattern, there are striking morning rush hours (7~9 am) and evening rush hours (5~7 pm) in all station functions except the airport-relevant station zone $s0$.

Based on the results shown in the aforementioned three figures, we can deduce the following:

(1) Passengers of type $c0$ have a rather sparse and random distribution on travel frequency and duration time. They travel 1 to 5 times monthly at most, with an average duration time between 30 and 90 minutes. They often appear in large-scale transportation hubs located within the downtown city. The representative station zones are $s1$ (e.g., XiDan), $s7$ (e.g., DongZhiMen), $s5$ (e.g., CBD), or $s6$ (e.g., NanLuoGuXiang). Besides, they arrive at the transportation hubs continuously in different densities at any time period of the day. Apparently, these passengers have both optional departure or arrival time and flexible origins or destinations. They may arrive for multiple purposes, such as shopping, friend visiting, dining out, or leisure entertainment. Thus we term type $c0$ as the elastic travel.

(2) Passengers of type $c1$ present the highest level of dependency on urban subway. They follow a regular travel pattern in travel frequency and duration by generating 23 to 65 travels in 20 to 30 fixed days monthly. The average duration time is between 30 and 90 minutes. Besides, they appear within the 6th Ring Road of Beijing. Several representative station zones are $s5$ (e.g., CBD), $s6$ (e.g., Tsinghua University), or $s2$ (e.g. HuiLongGuan). They arrive at station zone $s5$ or $s6$ intensively during morning rush hours and arrive at station zone $s2$ or $s6$ intensively during evening rush hours. No such sign exists in other time periods of a day. Apparently, these passengers have a rigid travel schedule. They commute every weekday in fixed rush hours between home and fixed hotspots such as workplaces or schools. And their travel purposes may be to work or study on time and leave by time. Thus, we term type $c1$ as the rigid travel. Passengers in this cluster have the highest probability to be commuters.

(3) Passengers of type $c2$ own a rare regularity on travel frequency, duration time, and visited stations. They

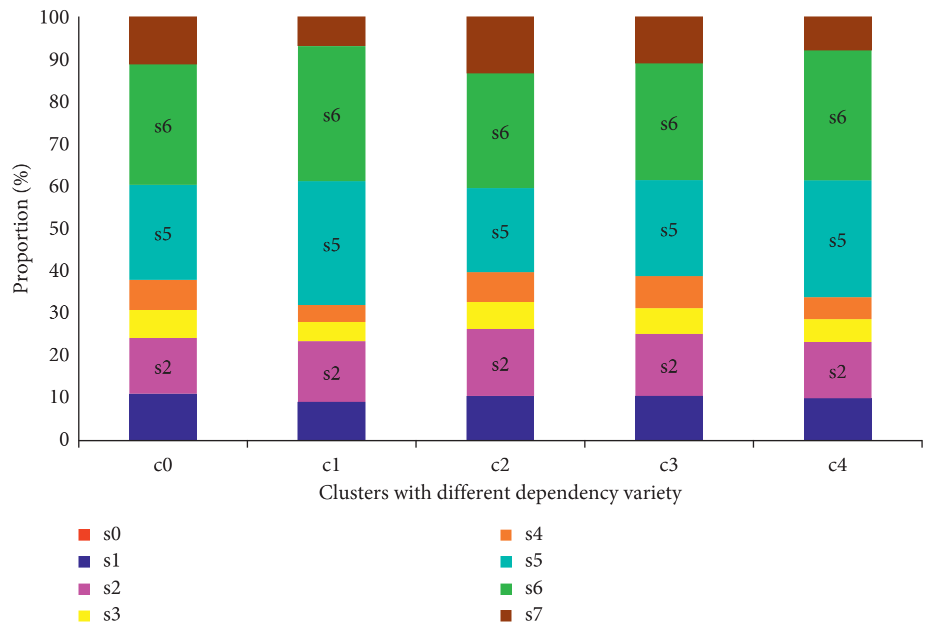


Figure 7: Specific proportion distribution over station functionality.

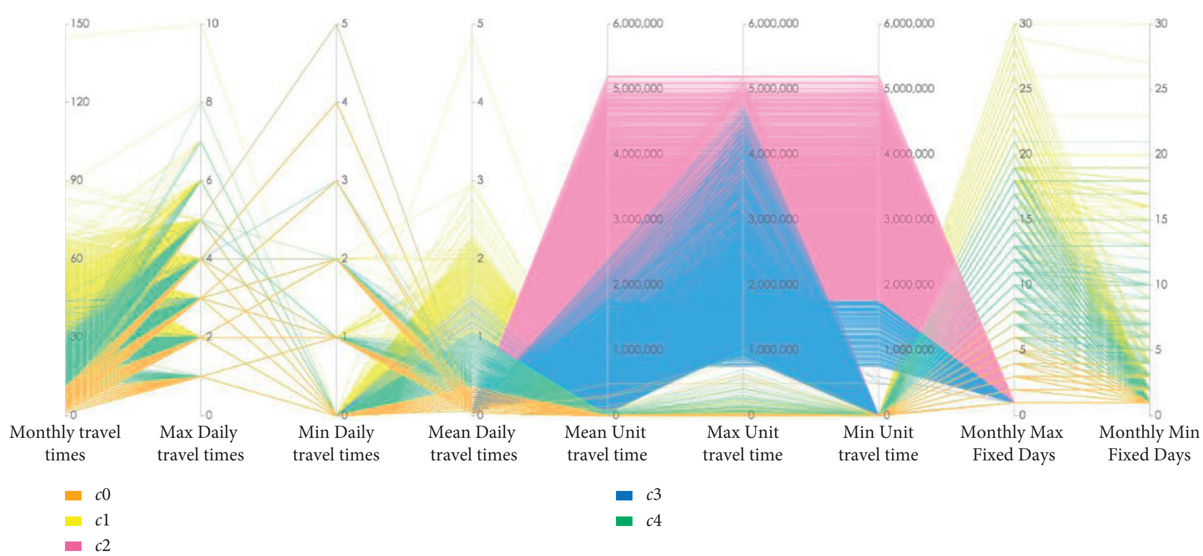
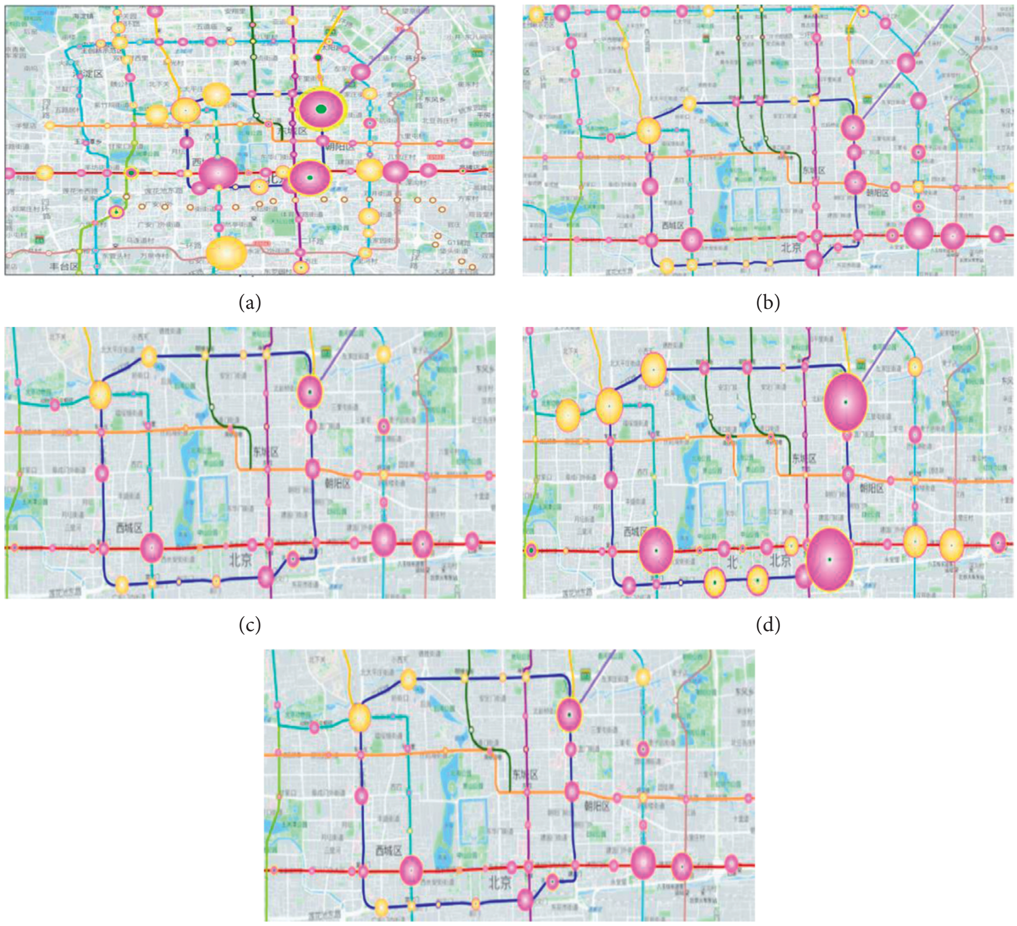


Figure 8: Statistical distribution of travel features describing travel dependency variety.

only travel 1 or 2 times monthly. But their average duration time is above 90 minutes, significantly longer than other clusters in every single trip. They mainly appear within the 4th Ring Road of Beijing. Several representative station zones are *s7* (e.g., Beijing Railway Station or Dongzhimen), *s6* (e.g., South Luogu Lane), *s4* (e.g., Beijing zoo or Qianmen), *s1* (e.g., the Forbidden City), or *s6* (e.g., Peking University or Third Hospital of Peking University). Obviously, these station zones have hotspots for scenic viewing or medical treatment. They arrive at these station zones continuously in different densities at any time period of the day. They are probably the floating population from other provinces of China. They generate floating tides on tourism or medical treatment. Thus, we term type *c2* as the infrequent floating travel with a long duration time once.

(4) Passengers of type *c3* own an unstable frequency by having travels 1 to 12 times in a month. They usually travel for 30 to 50 minutes in every single trip, but sometimes, they have aperiodic long travels with duration time reaching to or even exceeding 90 minutes. This extreme long travel is caused by the long distance between origin and destination zones. They have frequently been to several representative zones such as *s2* (e.g., Longze), *s6* (e.g., Tiantongyuan), or *s7* with large-scale transit hubs (e.g.,

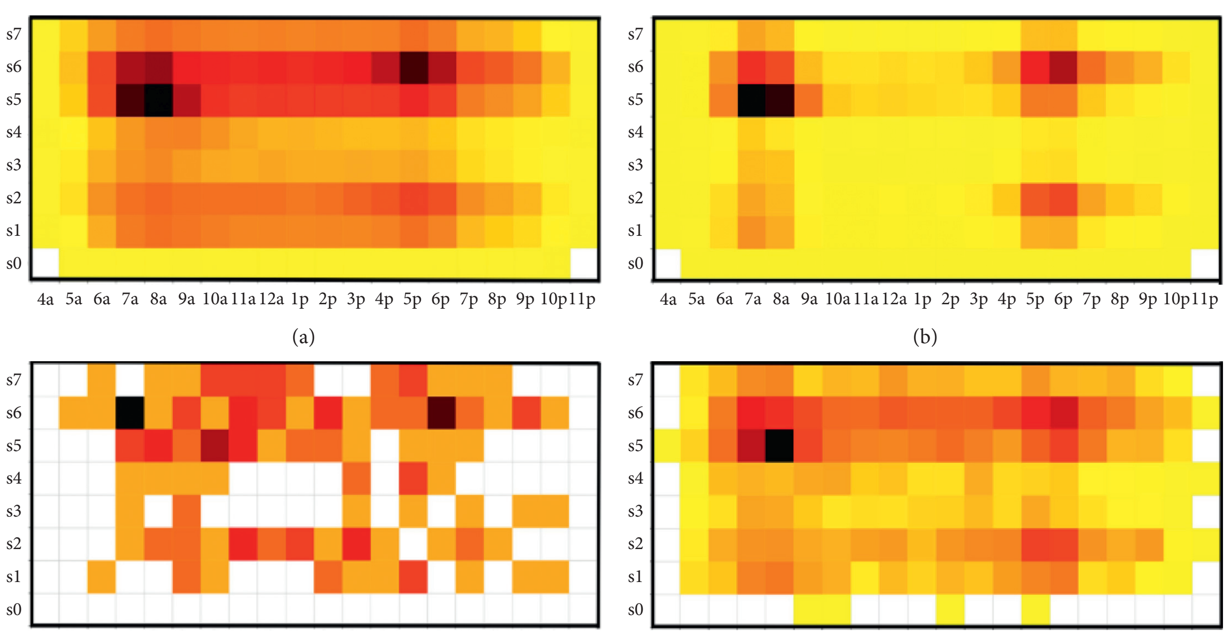


(a) (b)

(c) (d)

(e)

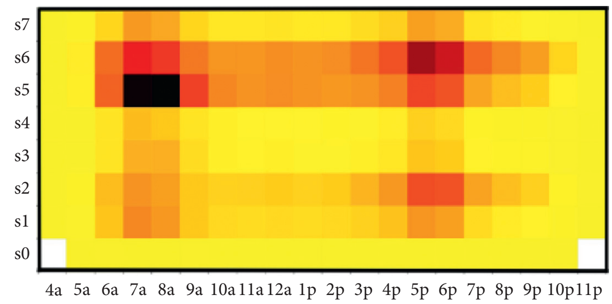
FIGURE 9: Spatial distribution of passengers by their dependency variety: (a) $c0$, (b) $c1$, (c) $c2$, (d) $c3$, and (e) $c4$.



(a) (b)

(c) (d)

FIGURE 10: Continued.



(e)

Figure 10: Spatiotemporal distribution of passengers' dependency levels regarding station functions: (a) $c0$, (b) $c1$, (c) $c2$, (d) $c3$, and (e) $c4$.

Dongzhimen or Xizhimen), which are with large-scale transit hubs. They arrive at the above station zones continuously in different densities at any time period of the day. They have aperiodic long-distance travels within Beijing, such as business travel (e.g., meetings) or leisure travel (e.g., visiting friends/family or outdoor recreation). Thus, we term type $c3$ as the aperiodic outdoor travel with a long duration time once.

(5) Passengers of type $c4$ own a rather high frequency by having travelled 12 to 34 times in a month. But their average duration time is lower than 30 minutes once in a trip, relatively shorter than other types of mobility patterns. They have frequently been to the station zone $s5$ or $s6$ within the 4th Ring Road or north part of Beijing. They can reach stations with functions $s5$ or $s6$ at any time of the day. Apparently, their purposes are not for work or education because neither time distribution nor frequency distribution uniforms to that of rigid travel. Their trip purposes are suspected to be abnormal activities, such as stealing, begging, or express delivery. But more evidences are needed to testify this hypothesis. Thus, we term type $c4$ as the frequent abnormity-suspected travel with a short duration time once.

6. Conclusions

This paper proposes a comprehensive framework to infer passengers' dependency levels on urban subway. Key features regarding travel frequency and time are extracted to quantify passengers' travel dependency on urban subway, which are clustered into five groups to distinguish their dependency variety. The function categories are further estimated for subway stations to help identify passengers' general purposes of visiting stations. Finally, experimental results on the deducted functions of stations are validated based on priori user experiences. The main findings are summarized as follows:

(1) Eight hidden functions of station zones are identified automatically. The top three functional station zones with the highest passenger flows are related to commerce, education/health, and residence. The functionality results well match the overall land-use plan in Beijing, 2015.

(2) Five mobility patterns are inferred with specific travel purposes. The first two patterns, with dominant regularity, are identified as the elastic and rigid mobility patterns. They present a medium level, together with a high level of dependency on the urban subway. Both of them are found to match the priori travel patterns proposed by the official report of Beijing in proportion distribution and spatio-temporal regularity.

(3) The rest three mobility patterns present nondominant travel regularity on the urban subway. Although they show a relatively low dependency and are in a small population, they share extreme, irregular, or versatile patterns in travel frequency, duration time, and frequently visited stations, which may increase the management difficulties to urban subway. All of them well match the statistical mobility patterns of Beijing residents proposed in the official report mentioned above. Most importantly, the successful exploration of travel patterns without dominant regularity may fill the research gap on uncovering the minority/extreme travel patterns that are less explored in previous research.

Data Availability

Data resources are provided by the Transportation Information Center and TOCC in Beijing. The introduction of our used data is given in the following link: https://www.researchgate.net/publication/335652865_OpenVisCardChains?_sg=HE3p3Jk6SUANbkm9NojZCrys7xoWtQvDNEF8PldRI62ObBE9InQMRGSoWKB_DmIi7rajRsU_u1qBh1oZ4c6FXEC3Oa2jPGlANluyeBuU.Ai4_SpHe6a_YZV3PVh8VVO409AKSCGng-POnDgccBYEJ6I_wiSCoIBy1p2LqSJL4L7GYuncjmzM8gh-ruJYGIQ.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the China Postdoctoral Science Foundation under Grant 2021M690332; in part by the Basic Scientific Research Foundations for Municipal Universities under Grant X21061; in part by the National Natural Science Foundation of China under Grants U1811463, 61632006, 52172301, 62072015, and 5170080357; and in part by the Beijing Talents Project under Grant 2017A24.

References

[1] L. Hansson and J. Holmgren, "Reducing dependency on special transport services through public transport," *Transportation Research Procedia*, vol. 25, pp. 2454–2464, 2017.

[2] M. Xue, H. Wu, C. Wei, W. S. Ng, and G. H. Goh, "Identifying tourists from public transport commuters," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1779–1788, New York, NY, USA, August 2014.

[3] T. Louail, M. Lenormand, O. García Cantú et al., "Uncovering the spatial structure of mobility networks," *Nature Communications*, vol. 6, no. 6007, pp. 6007-6008, 2015.

[4] X. Ma, Y.-J. Wu, Y. Wang, F. Chen, F. Chen, and J. Liu, "Mining smart card data for transit riders' travel patterns," *Transportation Research Part C: Emerging Technologies*, vol. 36, pp. 1–12, 2013.

[5] L. M. Kieu, A. Bhaskar, 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.

[6] X. Cao, G. Cong, C. S. Jensen, and M. L. Yiu, "Retrieving regions of interest for user exploration," *Proceedings of the VLDB Endowment*, vol. 7, no. 9, pp. 733–744, 2014.

[7] J. Gan, J. Zhang, and S. Zheng, "Where you really are: user trip based city functional zone ascertainment," in *Proceedings of the IEEE 37th International Performance Computing and Communications Conference*, pp. 1–8, Orlando, FL, USA, November 2018.

[8] N. J. Yuan, Y. Zheng, X. Xie et al., "Discovering urban functional zones using latent activity trajectories," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 3, pp. 712–725, 2015.

[9] J.-A. Carrasco and E. J. Miller, "The social dimension in action: a multilevel, personal networks model of social activity frequency between individuals," *Transportation Research Part A: Policy and Practice*, vol. 43, no. 1, pp. 90–104, 2009.

[10] W. Jang, "Travel time and transfer analysis using transit smart card data," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2144, no. 1, pp. 142–149, 2010.

[11] T. Hu, R. Song, Y. Wang, X. Xie, and J. Luo, "Mining shopping patterns for divergent urban regions by incorporating mobility data," in *Proceedings of the Conference on Information and Knowledge Management*, pp. 569–578, Indianapolis,IN USA, October 2016.

[12] S. Kim and G. F. Ulfarsson, "Activity space of older and working-age adults in the puget sound region, Washington," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2494, no. 1, pp. 37–44, 2015.

[13] B. Du, C. Liu, W. Zhou, Z. Hou, and X. Hui, "Detecting pickpocket suspects from large-scale public transit records," *IEEE Transactions on Knowledge and Data Engineering*, vol. 99, pp. 465–478, 2018.

[14] X. Zhao, Y. Zhang, H. Liu et al., "Detecting pickpocketing gangs on buses with smart card data," *IEEE Intelligent Transportation Systems Magazine*, vol. 11, no. 3, pp. 181–199, 2019.

[15] Z. Cui, Y. Long, R. Ke, and Y. Wang, "Characterizing evolution of extreme public transit behaviour using smart card data," in *Proceedings of the IEEE International Conference on Smart Cities*, pp. 142–148, Guadalajara, Mexico, October 2015.

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

[17] X. Zhao, M. Cui, and D. Levinson, "Temporal variations in daily activity networks using smartcard data," in *Proceedings of the 99th Transportation Research Board Annual Meeting*, pp. 1–39, Washington DC, USA, January 2020.

[18] J. Huang, D. Levinson, J. Wang, D. Levinson, J. Wang, and H. Jin, "Job-worker spatial dynamics in Beijing: Insights from smart card data," *Cities*, vol. 86, pp. 83–93, 2019.

[19] Y. Long and J.-C. Thill, "Combining smart card data and household travel survey to analyze jobs-housing relationships in Beijing," *Computers, Environment and Urban Systems*, vol. 53, pp. 19–35, 2015.

[20] J. Zhou and Y. Long, "Jobs-housing balance of bus commuters in Beijing: exploration with large-scale synthesized smart card data," *Journal of Transportation Research Record*, vol. 2418, no. 3, pp. 1–10, 2014.

[21] Q. Zou, X. Yao, P. Zhao et al., "Detecting home location and trip purposes for cardholders by mining smart card transaction data in Beijing subway," *Transportation*, vol. 45, no. 3, pp. 919–944, 2018.

[22] S. G. Lee and M. Hickman, "Are transit trips symmetrical in time and space?" *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2382, no. 1, pp. 173–180, 2013.

[23] X. Zhao, Y. Zhang, Y. Hu, H. Liu, and B. Yin, "Modeling relation proximity of passengers using public transit smart card data," *IEEE Intelligent Transportation Systems Magazine*, 2020.

[24] D. Zhang, J. Zhao, F. Zhang, and T. He, "coMobile: real-time human mobility modeling at urban scale using multi-view learning," in *Proceedings of the ACM International Conference on Advances in Geographic Information Systems*, pp. 1–10, Bellevue, WA, USA, November 2015.

[25] S. Lee and M. Hickman, "Trip purpose inference using automated fare collection data," *Public Transportation*, vol. 6, no. 1-2, pp. 1–20, 2014.

[26] M.-K. Kim, S.-P. Kim, J. Heo, and H.-G. Heo, "Ridership patterns at subway stations of Seoul capital area and characteristics of station influence area," *KSCE Journal of Civil Engineering*, vol. 21, no. 3, pp. 964–975, 2017.

[27] S. Hu, J. Weng, W. Zhou, P. Lin, and Z. Liu, "A data-fusion based method for identifying individual passenger travel dependence on public transportation," in *Proceedings of the IEEE Intelligent Transportation Systems Conference*, pp. 142–148, Auckland, New Zealand, October 2019.

[28] M. Yildirimoglu, Y. Limniati, and N. Geroliminis, "Investigating empirical implications of hysteresis in day-to-day travel time variability," *Transportation Research Part C: Emerging Technologies*, vol. 55, pp. 340–350, 2015.

[29] D. B. E. Dharmowijoyo, Y. O. Susilo, and A. Karlström, "Relationships among discretionary activity duration, its travel time spent and activity space indices in the Jakarta

Metropolitan Area, Indonesia," *Journal of Transport Geography*, vol. 54, pp. 148–160, 2016.

[30] D. T. Pham, S. S. Dimov, S. S. Dimov, and C. D. Nguyen, "Selection of k in k-means clustering," *Proceedings of the Institution of Mechanical Engineers - Part C: Journal of Mechanical Engineering Science*, vol. 219, no. 1, pp. 103–119, 2005.

[31] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[32] Y. Long and Z. Shen, *Geospatial Analysis to Support Urban Planning in Beijing*, Spring GeoJournal Library, New York, NY, USA, 2015.

[33] H. Wen, M. Sun, and S. Wang, *Beijing Transport Annual Report 2015*, Publication Beijing transportation research center, Beijing, China, 2015.

[34] M. K. Mahrsi, E. Côme, J. Baro, and L. Oukhellou, "Understanding passenger patterns in public transit through smart card and socioeconomic data: a case study in rennes, France," in *Proceedings of the international Workshop on Urban Computing of ACM*, pp. 1–9, New York, NY, USA, August 2014.

[35] X. Zhao, Y. Zhang, Y. Hu, Z. Qian, H. Liu, and B. Yin, "Interactive Visual exploration of human mobility correlation based on smart card data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 8, pp. 4825–4837, 2020.